UNIVERSIDADE FEDERAL DO PARANÁ

CLAUDIA MARINA SCHELLIN BECKER

MARCADORES DO CROMOSSOMO Y (Y-SNPS E Y-STRS): APLICAÇÕES NA INFERÊNCIA DE ANCESTRALIDADE PATERNA E INTERPRETAÇÃO DE MISTURAS FORENSES

CURITIBA

2023

CLAUDIA MARINA SCHELLIN BECKER

MARCADORES DO CROMOSSOMO Y (Y-SNPS E Y-STRS): APLICAÇÕES NA INFERÊNCIA DE ANCESTRALIDADE PATERNA E INTERPRETAÇÃO DE MISTURAS FORENSES

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Genética, Programa de Pós-graduação em Genética, Setor de Ciências Biológicas, Universidade Federal do Paraná.

Orientadora: Profa. Dra. Danielle Malheiros

Coorientador: Prof. Dr. Celso Teixeira Mendes Junior

CURITIBA

2023

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIAS BIOLÓGICAS

Becker, Claudia Marina Schellin.
Marcadores do cromossomo Y (Y-SNPs e Y-STRs): aplicações na inferência de ancestralidade paterna e interpretação de misturas forenses. / Claudia Marina Schellin Becker. – Curitiba, 2023.
1 recurso on-line : PDF.

Tese (Doutorado) – Universidade Federal do Paraná, Setor de Ciências Biológicas. Programa de Pós-Graduação em Genética.
Orientadora: Prof.ª Dra. Danielle Malheiros.
Coorientador: Prof. Dr. Celso Teixeira Mendes Junior.

1. Haplótipos. 2. Genética forense. 3. Miscigenação. 4. População. I. Malheiros, Danielle. II. Mendes-Junior, Celso Teixeira. III. Universidade Federal do Paraná. Setor de Ciências Biológicas. Programa de Pós-Graduação em Genética. IV. Título.

Bibliotecária: Rosilei Vilas Boas CRB-9/939

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação GENÉTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **CLAUDIA MARINA SCHELLIN BECKER** intitulada: **Marcadores do cromossomo Y (Y-SNPs e Y-STRs): aplicações na inferência de ancestralidade paterna e interpretação de misturas forenses**, sob orientação da Profa. Dra. DANIELLE MALHEIROS FERREIRA, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutora está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 29 de Setembro de 2023.

Assinatura Eletrônica
03/10/2023 13:46:24.0
DANIELLE MALHEIROS FERREIRA
Presidente da Banca Examinadora

Assinatura Eletrônica
12/10/2023 09:39:36.0
SILVIENE FABIANA DE OLIVEIRA
Avaliador Externo (UNIVERSIDADE DE BRASÍLIA)

Assinatura Eletrônica
17/11/2023 17:35:53.0
MARCIA HOLSBACH BELTRAME
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
04/10/2023 14:41:55.0
LIANA ALVES DE OLIVEIRA
Avaliador Externo (FACULDADES INTEGRADAS DO BRASIL)

Assinatura Eletrônica
05/10/2023 11:04:02.0
CELSO TEIXEIRA MENDES JUNIOR
Coorientador(a) (FACULDADE DE FILOSOFIA CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO)

Dedico este trabalho a meus pais, meu marido e meus filhos, cuja compreensão, apoio e incentivo têm sido constantes e incondicionais ao longo desta jornada.

AGRADECIMENTOS

Agradeço a Deus pela oportunidade, força e inspiração que me permitiram realizar meu doutorado.

Agradeço imensamente à minha orientadora Profa. Dra. Danielle Malheiros, pela presença e apoio constantes, por apostar em mim durante todos estes anos e, mesmo em frente a tantos desafios, oportunizar meu crescimento com profissionalismo e carinho. Ao meu coorientador, Prof. Dr. Celso Teixeira Mendes Junior, que embarcou conosco nessa jornada num momento decisivo, agradeço pela receptividade e por ter acreditado em mim, investindo tempo e compartilhando oportunidades e conhecimento. A ambos, agradeço a orientação, dedicação, disponibilidade, apoio e incentivo. Obrigada por realizarem este doutorado ao meu lado.

Agradeço à minha família, em especial minha mãe Annelie, meu marido Ermelino e meus filhos Melissa e Felipe, pelo amor, incentivo, paciência e compreensão que fizeram toda a diferença para que eu prosseguisse nessa jornada.

Agradeço ao Prof. Dr. Ulf Gregor Baranow, meu tio e padrinho, pela inspiração, conselhos, apoio e incentivo.

Agradeço à Iriel Araceli Joerin Luque, por sempre encontrar prontamente um tempinho para mim, por compartilhar seu conhecimento e experiência e pela inestimável ajuda e simpatia, em tantas ocasiões.

Agradeço ao Henrique Trigo de Castro Junior, pela parceria, por sua valiosa contribuição e constante disponibilidade em ajudar em momentos tão variados, sempre dando conta do recado com muita agilidade, simpatia, competência e profissionalismo.

Agradeço ao Vitor Matheus Soares Moraes, por habilmente conduzir a execução dos procedimentos nos programas, por preparar e apresentar os dados gerados para minha análise e por dedicar seu tempo e expertise contribuindo fundamentalmente para o sucesso desta pesquisa.

Agradeço à Profa. Dra. Maria Luíza Petzl-Erler, por disponibilizar as amostras das populações para que este estudo pudesse ser realizado, e ao Prof. Dr. Danillo Gardenal Augusto, por ter realizado a genotipagem das amostras em larga escala.

Agradeço aos meus colegas do LGMF, Anna Carolina, Camilinha, Danilo, Halina, Jaqueline, Josiane, Juliane, Leonardo, Luciellen, Marcelo, Marianna, Paula, Pedro e Ricardo, pelo apoio e cooperação sempre que se fizeram necessárias.

Agradeço ao Programa de Pós-graduação em Genética da Universidade Federal do Paraná, pelo incentivo à pesquisa e por todo o conhecimento transmitido. Um agradecimento especial à Priscila Jansen e às alunas Bruna Terezinha Magnabosco Ferreira da Cruz e Sara Santos Matsunaga, do LGMH, pela imensa ajuda em separar e preparar as amostras para minhas análises.

Agradeço à Polícia Científica do Paraná, pela disponibilização da infraestrutura e por ter apoiado este estudo.

Agradeço à CAPES e ao CNPq, pela disponibilização de recursos financeiros para a realização desta pesquisa.

Com sabedoria se constrói a casa, e com discernimento se consolida. Pelo conhecimento seus cômodos se enchem do que é precioso e agradável"
(Provérbios 24:4)

RESUMO

A genética forense objetiva a comparação de perfis genéticos encontrados em vestígios de crimes com possíveis suspeitos. Nos casos de violência sexual conjuntos de marcadores forenses STR (*Short Tandem Repeats*) autossômicos e do cromossomo Y são comumente analisados. Misturas genéticas entre vítima e agressor ou agressores podem comprometer a interpretação de perfis autossômicos, restando apenas o cromossomo Y para análise. Quando a autoria do delito é desconhecida, informações relativas à ancestralidade paterna do agressor ou agressores, a partir do conhecimento da diversidade local, podem auxiliar nas investigações. Haplogrupos do cromossomo Y determinados por polimorfismos de nucleotídeo único ou SNPs (*Single Nucleotide Polimorfisms*), refletem a origem biogeográfica de grupos de indivíduos com ancestrais comuns, mas seu uso forense ainda não é rotineiro. Entretanto, haplogrupos compartilham marcadores Y-STRs característicos e sua predição a partir de haplótipos Y-STRs tem se consolidado. Preditores de haplogrupos a partir de haplótipos Y-STR foram desenvolvidos, mas suas taxas de erros ainda limitam seu uso com fins forenses na população miscigenada brasileira. Considerando o contexto acima, os objetivos deste trabalho foram: (1) avaliar o desempenho do programa *STRUCTURE*, ainda não utilizado para este fim, para inferência de haplogrupos a partir de haplótipos; (2) Estimar os parâmetros forenses dos marcadores analisados na população em pauta; (3) Investigar, em misturas forenses de dois indivíduos, como se comportam as médias da diversidade genética dos Y-STRs nos diferentes haplogrupos. Foram genotipados para 23 marcadores Y-STRs do *kit* PowerPlex® Y23 (Promega) 180 indivíduos do sexo masculino, sendo 172 urbana do Mato Grosso do Sul (n=125), São Paulo (n=17), Paraná (n=24), Rondônia e Roraima (n=2), Paraguai (n=1) e 3 sem origem geopolítica informada, e 8 indígenas do Paraná, todos previamente classificados em haplogrupos de Y-SNPs. A inferência dos haplogrupos desta população, tomando como referência 573 indivíduos do painel do HGDP-CEPH, foi feita com o *STRUCTURE,* comparando-se com outros preditores, *HAPEST* e *NevGen*. Observou-se maior taxa de acertos do *STRUCTURE* (89,94%, 85,47% e 81,56%, respectivamente), e baixa taxa de erros (4,47%, 13,41% e 3,35%, respectivamente). *STRUCTURE* obteve resultados corretos em uma porção significativa das situações, incluindo aquelas em que outros programas divergiram ou não geraram resultados, indicando sua eficácia. A associação de dois programas poderia incrementar a segurança dos resultados obtidos. O conjunto das populações revelou valores elevados de diversidades haplotípica (DH) e gênica (DG) (DH=0,9999 e DG=0,6920). A DH, a probabilidade de coincidência (PC) e o poder de discriminação (PD) obtidos para os haplogrupos europeu (DH=0,9999; PC=6,94x10^{-3}; PD=0,9865), africano (DH=1,000; PC=9,09x10^{-2}; PD=1,000) e ameríndio (DH=0,9948; PC=6,37x10^{-2}; PD=0,8947) confirmaram a eficácia do painel de Y23 nestas populações. Foram realizadas simulações computacionais de misturas dos perfis de Y-STR de pares de indivíduos. Dentro do mesmo haplogrupo os marcadores se mostraram menos diferentes do que entre haplogrupos distintos. Seu padrão de diferenças nas misturas parece promissor para detectar tendências e permitir a inferência de ABG, porém requer investigações mais minuciosas, refinamento das técnicas e aprimoramento da metodologia. O aumento do número amostral possibilitaria

uma investigação mais robusta, tanto do uso do *STRUCTURE* para inferência de haplogrupos a partir de Y-STRs, quanto da análise de haplogrupos em misturas genéticas.

Palavras-chave: Haplótipos de Y-STRs; haplogrupos de Y-SNPs; inferência de ancestralidade biogeográfica; programa *STRUCTURE*; misturas forenses; diversidade haplotípica; populações miscigenadas brasileiras.

ABSTRACT

Forensic genetics aims to compare genetic profiles found in crime scene evidence with potential suspects. In cases of sexual violence, sets of forensic autosomal Short Tandem Repeats (STRs) and Y-chromosome markers are commonly analyzed. Genetic mixtures between victims and individuals from different haplogroups can complicate the interpretation of autosomal profiles, leaving only the Y-chromosome for analysis. When the identity of the individuals is unknown, information regarding the paternal ancestry of the individuals, derived from knowledge of local diversity, can assist in investigations. Haplogroups of the Y-chromosome, determined by Single Nucleotide Polymorphisms (SNPs), reflect the biogeographic origin of groups with common ancestors, yet their forensic use is not routine. However, haplogroups share characteristic Y-STR markers, and their prediction from Y-STR haplotypes has gained acceptance. Predictors of haplogroups from Y-STR haplotypes have been developed, but their error rates still limit their use for forensic purposes in the admixed Brazilian population. Given the context above, the objectives of this work were: (1) to evaluate the performance of the STRUCTURE program, not yet used for this purpose, for haplogroup inference from haplotypes; (2) to estimate the forensic parameters of the markers analyzed in the population under study; (3) to investigate how the genetic diversity means of Y-STRs behave in different haplogroups in forensic mixtures involving two individuals. A total of 180 males were genotyped for 23 Y-STR markers from the PowerPlex® Y23 kit (Promega), including 172 urban individuals from Mato Grosso do Sul (n=125), São Paulo (n=17), Paraná (n=24), Rondônia and Roraima (n=2), Paraguay (n=1), and 3 with no reported geopolitical origin, as well as 8 indigenous individuals from Paraná, all previously classified in Y-SNP haplogroups. Haplogroup inference for this population, using 573 individuals from the HGDP-CEPH panel as a reference, was performed with STRUCTURE, comparing with other predictors, HAPEST and NevGen. STRUCTURE showed a higher accuracy rate (89.94%, 85.47%, and 81.56%, respectively) and a low error rate (4.47%, 13.41%, and 3.35%, respectively). STRUCTURE obtained correct results in a significant portion of situations, including those where other programs diverged or did not generate results, indicating its effectiveness. The combination of two programs could enhance result reliability. The pooled populations revealed high haplotypic diversity (HD) and gene diversity (GD) values (HD=0.9999 and GD=0.6920). HD, the probability of coincidence (PC), and the power of discrimination (PD) obtained for European (HD=0.9999; PC=6.94x10-3; PD=0.9865), African (HD=1.000; PC=9.09x10-2; PD=1.000), and Amerindian haplogroups (HD=0.9948; PC=6.37x10-2; PD=0.8947) confirmed the effectiveness of the Y23 panel in these populations. Computational simulations of mixtures of Y-STR profiles from pairs of individuals were conducted. Within the same haplogroup, markers showed fewer differences than between distinct haplogroups. Their pattern of differences in mixtures seems promising for detecting trends and allowing Ancestry-by-Genetics (ABG) inference, but it requires further investigations, technique refinement, and methodology improvement. Increasing the sample size would enable a more robust investigation, both for the use of STRUCTURE for haplogroup inference from Y-STRs and for haplogroup analysis in genetic mixtures.

Keywords: Y-STR haplotypes; Y-SNP haplogroups; Biogeographic ancestry inference; Structure; Forensic mixtures; Haplotypic diversity; Brazilian admixed populations.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABG	Ancestralidade biogeográfica
AFR	Haplogrupos africanos
AIMs	Ancestry Informative Markers
AMR	Haplogrupos indígenas
CEPH	Centre D'Êtude du Polymorphisme Humain
DG	Diversidade Gênica
DH	Diversidade Haplotípica
EUR	Haplogrupos europeus
FTDNA	Family-tree DNA
FUNAI	Fundação Nacional do Índio
GWAS	Genome Wide Association Studies
HGDP	Human Genome Diversity Project
INDELs	Insertion-deletion events
ISFG	DNA Commission of the International Society of Forensic Genetics
ISOGG	International Society of Genetic Genealogy
PC	Probabilidade de Coincidência
PD	Poder de Discriminação
SMM	Stepwise mutation model
SNP	Single Nucleotide Polimorfism
SNV	Single nucleotide variation
STR	Short Tandem Repeat
SWGDAM	Scientific Working Group on DNA Analysis Methods
UME	Unique Mutation Event
YHRD	The Y Chromosome Haplotype Reference Database

SUMÁRIO

1. INTRODUÇÃO

O principal objetivo deste trabalho de doutoramento foi investigar e avaliar a utilidade de marcadores genéticos STR do cromossomo Y na inferência da ancestralidade paterna em contextos de análise forense de vestígios coletados em locais de crime. A obtenção do conhecimento sobre a ancestralidade paterna tem como propósito propiciar o direcionamento das investigações criminais em casos nos quais não há suspeitos identificados. Em essência, o estudo concentrou-se em duas linhas de investigação, a partir da quantidade de perfis genéticos presentes na amostra forense:

1) Em amostras de fonte única, isto é, constituídas por apenas um perfil genético, objetivou-se avaliar o potencial de inferência de haplogrupos e predição de ancestralidade biogeográfica paterna de indivíduos da população brasileira (urbanos e indígenas) ao se utilizar haplótipos de Y-STRs e o programa *STRUCTURE* como ferramenta de análise e classificação. A performance dessa ferramenta foi estimada tendo-se como referência a determinação de haplogrupos a partir de Y-SNPs para todos os indivíduos da amostra populacional. Ainda, o desempenho do *STRUCTURE* foi comparado com o de outras ferramentas já estabelecidas e usadas com o propósito de inferência de haplogrupos a partir de haplótipos de Y-STRs. Nesta seção também são apresentadas informações sobre a diversidade dos haplótipos Y-STR na amostra populacional estudada, bem como as estimativas de parâmetros forenses relacionados ao sistema comercial utilizado para a genotipagem dos Y-STRs.

2) Em amostras de Y-STRs que compreendem fonte múltipla de material genético, isto é, misturas envolvendo dois doadores do sexo masculino, avaliou-se a viabilidade de inferir a ancestralidade paterna dos contribuintes. Tal avaliação foi realizada com base em padrões de diferenças intra-locus, ou seja, de diferenças de diversidade, uma vez que a ferramenta *STRUCTURE* não possui aplicabilidade para perfis múltiplos. Esta seção analisa os comportamentos dos marcadores Y-STR em vários cenários de misturas forenses, considerando indivíduos de haplogrupos iguais ou distintos, com o

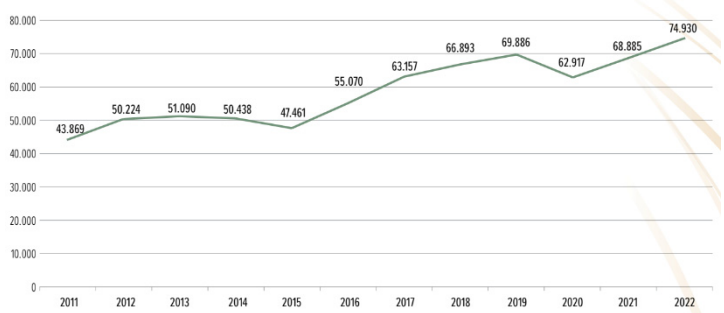
objetivo de identificar padrões que possibilitem a inferência da origem biogeográfica paterna dos contribuintes envolvidos nas misturas.

2. REVISÃO DE LITERATURA

2.1 Análise do Cromossomo Y na Genética Forense: possibilidades de aplicação nos casos de violência sexual

No Brasil, estima-se que ocorrem cerca de 822 mil casos de violência sexual a cada ano, o que equivale a dizer que são duas agressões por minuto. Existe uma clara tendência, preocupantemente crescente, desta modalidade de crime (**Figura 1**), tendo sido, em 2023, constatado crescimento de 7% em relação ao ano anterior. A proporção de subnotificação é preocupante, com apenas 8,5% dos casos chegando ao conhecimento das autoridades policiais. O estudo ainda aponta que o maior número de casos de estupro, entre os anos de 2009 e 2019, ocorreu entre jovens com pico de idade aos 13 anos, sendo que, nos casos envolvendo crianças e adolescentes com idades de 0 a 13 anos, 13,9% dos autores são desconhecidos (INSTITUTO DE PESQUISA ECONÔMICA APLICADA - IPEA, 2023). Entre as vítimas de 14 anos ou mais, 22,8% dos estupros foram praticados por desconhecidos (FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA, 2023).

Figura 1 - Evolução do número de estupros (Brasil, 2011-2022)



Fonte: Anuário Brasileiro de Segurança Pública (2023)

Conforme preconiza a legislação brasileira, quando a infração deixar vestígios, deve ser realizado o exame do corpo de delito (BRASIL, 1941). Assim, mediante solicitação da polícia judiciária, os vestígios procedentes da agressão são coletados e encaminhados aos Laboratórios de Genética Forense das

diversas unidades da Polícia Científica do país, onde representam um expressivo número de materiais.

A rotina de análise da genética forense compreende o uso de um conjunto pré-estabelecido de marcadores que são comparados entre vestígios e suspeitos (COURT, 2021). Atualmente, os STRs (do inglês, *short tandem repeats*) são os marcadores de escolha na rotina forense, sendo constituídos por um tipo de polimorfismo de comprimento, formado por nucleotídeos organizados em sequências curtas que formam blocos de tamanho variável (dois a sete pares de bases), os quais se repetem consecutivamente por um número variável de vezes (MORAN *et al.*, 2004). Os marcadores STR de interesse forense incluem tanto os localizados nos cromossomos autossômicos quanto aqueles distribuídos ao longo do cromossomo Y (Y-STR). A identificação da origem de conjunto de STRs depende da existência de um indivíduo específico a ser confrontado. Na falta de um suspeito, bancos de perfis genéticos alimentados com perfis genéticos obtidos de criminosos condenados e vestígios podem oferecer a possibilidade de uma coincidência (JOBLING, 2022).

A avaliação estatística do valor de prova obtido a partir de marcadores STR autossômicos, tipicamente herdados de maneira independente, envolve a “regra dos produtos”, na qual as probabilidades de coincidência calculadas individualmente para cada STR são multiplicadas entre si, o que reduz a chance de uma coincidência aleatória (COURT, 2021). Na herança dos marcadores Y-STR, por sua vez, a regra dos produtos não pode ser aplicada em virtude de sua característica transmissão em blocos (chamados haplótipos) entre as gerações patrilineares. Neste caso, devem ser utilizadas as frequências haplotípicas disponíveis em bancos de dados de referência, como o YHRD (do inglês The *Y Chromosome Haplotype Reference Database*), disponível em www.yhrd.org. Essas frequências são obtidas em estudos populacionais e sua utilização é baseada em modelos estatísticos que seguem diretrizes internacionais de interpretação, como as do SWGDAM (do inglês *Scientific Working Group on DNA Analysis Methods*) e do ISFG (do inglês *DNA Commission of the International Society of Forensic Genetics*) (KAYSER, 2017). Através das frequências haplotípicas do YHRD, é possível estimar a probabilidade de coincidência do haplótipo Y-STR encontrado em um vestígio de crime (KAYSER, 2017).

Embora a aplicação dos STRs autossômicos e do cromossomo Y para identificação humana encontre-se atualmente muito bem estabelecida, oferecendo a possibilidade de resolução da maior parte dos casos sob investigação (CHUNG; FUNG; HU, 2010; KAYSER, 2017), sua capacidade resolutiva requer que os genótipos de STRs autossômicos ou haplótipos de Y-STRs obtidos sejam identificados de maneira inequívoca.

Quando um vestígio de violência sexual apresenta a deposição simultânea do material biológico de mais de um indivíduo, como no caso de múltiplos agressores ou de vítimas masculinas, obtém-se uma mistura do material genético de todos os contribuintes, situação que dificulta a individualização do perfil STR de cada um e que representa substancial risco de comprometimento das análises (MEYER *et al.*, 2015). A não ser que a diferença de proporção de contribuição dos indivíduos componentes dessa mistura seja muito evidente, a individualização do perfil de cada um representa um desafio significativo ao analista (KAYSER, 2017; SHARMA; SK, 2018). Por outro lado, a desproporção entre os componentes pode também dificultar a detecção do perfil do agressor, como num cenário de vítima feminina de estupro, pois, dada a natureza íntima da coleta do vestígio, o material da vítima frequentemente sobrepuja o material do autor, mascarando-o (KAYSER, 2017; MEYER *et al.*, 2015). O fenômeno de amplificação preferencial do material genético do componente majoritário, juntamente com o possível compartilhamento de alelos entre vítima e suspeito, compromete a individualização do perfil autossômico do autor nestes casos, gerando, frequentemente, resultados inconclusivos (KAYSER, 2017).

Por estes motivos, evidencia-se a complexidade da interpretação de misturas e da necessidade do estabelecimento de metodologias e abordagens estatísticas para sua deconvolução, o que têm constituído tema de pesquisa intensa nos últimos anos (GRAVERSEN; MORTERA; LAGO, 2019; MORTERA, 2020).

Em virtude de seus locos de cópia única, o perfil de Y-STRs frequentemente auxilia a determinar o número de indivíduos presente numa mistura genética (GUSMÃO *et al.*, 2017) e, nos casos nos quais a vítima não possua o

cromossomo Y, pode ajudar na individualização do contribuinte masculino (KAYSER, 2017; MEYER *et al*., 2015).

Eventualmente o confronto direto através de marcadores STR pode não ser suficiente para oferecer a elucidação de um crime, como bem ilustrado no caso Vaatstra, um crime ocorrido na Holanda no ano de 1999. Neste caso emblemático de violência sexual seguida de assassinato de uma jovem, as investigações iniciais não conduziram a um suspeito através do perfil de STR autossômicos obtido de vestígio coletado da vítima. Tampouco foram encontradas coincidências em consulta aos bancos de perfis genéticos. As características do crime e a falta de sua resolução levaram a população local a suspeitar intensamente dos refugiados estrangeiros que viviam na região, o que resultou em conflitos entre moradores, refugiados e as autoridades policiais. Estas solicitaram então o auxílio do Laboratório Forense de Pesquisa em DNA do Departamento de Genética Humana do Centro Médico da Universidade de Leiden. Através do perfil dos marcadores STR do cromossomo Y encontrado no material coletado da vítima, foi inferida a origem biogeográfica do autor, sendo este identificado como originário do oeste europeu. Essa constatação removeu as suspeitas sobre estrangeiros e concentrou os esforços investigativos em indivíduos da própria região. Contudo, o caso permaneceu sem solução por quatorze anos. Como último recurso para resolver o caso Vaatstra, foi realizada, em 2012, uma campanha direcionada à população local visando a coleta voluntária de esfregaço bucal, da qual mais de 6.600 homens participaram. Já no primeiro lote de análises do DNA do cromossomo Y obtido durante a campanha, os investigadores identificaram indivíduos que seriam parentes próximos do doador do perfil genético do vestígio. Tendo rastreado a família paterna do autor do crime, foi possível direcionar as investigações e descobrir a identidade do assassino, o qual curiosamente também havia participado da campanha. O confronto genético final confirmou a coincidência entre os perfis de STRs autossômicos obtidos do vestígio e do suspeito, proporcionando a resolução do caso (KAYSER, 2017).

Um exemplo marcante adicional diz respeito à resolução do homicídio de outra jovem, Pamela Cahanes, 34 anos após o ocorrido. Durante todo esse período, a suposição era de que o perpetrador fosse de origem europeia. No

entanto, em 2015, a análise do perfil de Y-STRs permitiu a inferência de uma origem africana, o que por fim levou à identificação do assassino (FITZPATRICK, 2022).

Como nos casos acima descritos, na impossibilidade do confronto direto ou na eventualidade de buscas infrutíferas nos bancos de perfis genéticos, qualquer informação adicional que possa ser obtida a partir do DNA, como a fenotipagem forense (predição de características externas visíveis, como a pigmentação de pelo, cabelo ou de olhos, ou de fatores que alteram parte de tais características, como a idade cronológica) ou a inferência de ancestralidade, poderia ajudar a direcionar as investigações indicando as probabilidades de identificação de um infrator desconhecido (CANALES SERRANO, 2020; GENTILE *et al.*, 2019; HAMCZYK *et al.*, 2020; HEIDEGGER *et al.*, 2022; KAYSER, 2017; KAYSER *et al.*, 2023; MONTESANTO *et al.*, 2020; PORRAS-HURTADO *et al.*, 2013; WOŹNIAK *et al.*, 2021).

A abordagem específica dos Y-STR como instrumento para inferência do haplogrupo Y-SNP, embora de aplicabilidade a ser aprimorada para populações miscigenadas (PENA; BORTOLINI, 2004), foi fundamental para o sucesso do caso Vaatstra e do caso Cahanes, destacando sua importância na solução de casos criminais em que os STR autossômicos não tenham sido elucidativos (MEYER *et al.*, 2015).

Dados de haplótipos compostos por Y-STRs já vêm sendo usados com sucesso em abordagens antropológicas desde então, revelando informações sobre a ancestralidade biogeográfica que antes só eram obtidas por meio de marcadores binários (SNPs, do inglês *Single Nucleotide Polymorphisms*) (BELEZA; LOPES; CARRACEDO, 2003; COURT, 2021; DE KNIJFF, 2000; KAYSER, 2017; NEBEL *et al.*, 2001). Assim, casos bem-sucedidos de inferências de parentesco visando a busca familiar têm sido relatados, inclusive para situações de misturas forenses de material genético (CHUNG; FUNG; HU, 2010; DØRUM; KAUR; GYSI, 2017; GRAVERSEN; MORTERA; LAGO, 2019; GREEN; MORTERA, 2017; KAUR *et al.*, 2015).

Por similaridade, a abordagem da inferência de haplogrupos dos perfis extraídos de misturas poderia, portanto, configurar uma ferramenta auxiliar em

casos de perfis múltiplos nos quais, após a individualização, não se obtenha sucesso nos confrontos genéticos. A indicação da provável ancestralidade biogeográfica de cada um dos perfis identificados poderia servir como uma informação auxiliar para propiciar o direcionamento em investigações criminais na busca pelos agressores, embora esta aplicabilidade apresente limitações no caso de populações com ancestralidades compartilhadas.

2.2 Ancestralidade Biogeográfica no Contexto Forense

Genética de populações pode ser definida como o ramo da genética dedicado a explicar a origem das variações genéticas (como frequências alélicas e genotípicas) observadas tanto dentro quanto entre populações, bem como a dinâmica evolutiva (como mutações, seleção natural, fluxo gênico e deriva genética) a que as populações estiveram sujeitas. A partir das variações entre as populações é possível deduzir-se a estrutura populacional ali existente. Da mesma forma, torna-se possível construir análises filogenéticas, que possibilitam a obtenção e organização de informações de ancestrais mais distantes e a identificação da ordem na hierarquia de sua descendência (UNDERHILL; KIVISILD, 2007). Dentro deste contexto, torna-se possível a inferência de ancestralidade biogeográfica (ABG) de indivíduos, a partir da compreensão da diversidade genética entre diferentes grupos populacionais e sua distribuição global (COURT, 2021). Enquanto ancestralidade é um termo de conceito mais amplo, relacionado a conceitos sociais baseados em raça e etnia (como aspectos de linguagem e cultura), a ancestralidade biogeográfica se refere à região geográfica de origem dos ancestrais de um indivíduo. Logo, a inferência da ABG implica em que sejam desconsiderados os conceitos como raça, etnia, linguagem, religião ou outras manifestações de cultura ou tradição (SCHNEIDER; PRAINSACK; KAYSER, 2019).

Múltiplos estudos desta natureza vêm sendo conduzidos há décadas, buscando desvendar os processos pelos quais ocorreu o povoamento de todas as regiões do globo, além de determinar os locais de origem das populações e estimar as prováveis rotas de suas migrações. Dada a história evolutiva humana mais recente, é esperado que as análises de ancestralidade biogeográfica

indiquem uma grande prevalência da contribuição de várias origens diferentes em indivíduos, decorrentes da sucessiva remoção de barreiras culturais, sociais e geográficas que originalmente limitavam os movimentos demográficos (PHILLIPS, 2015), possibilitando as trocas de material genético entre populações. Logo, a inferência da ABG pode ser realizada a partir de dados genéticos, os quais são transmitidos por herança biológica e, por este motivo, são passíveis de individualização (SCHNEIDER; PRAINSACK; KAYSER, 2019).

A diversidade genética de grupos populacionais decorre de eventos relacionados à processos evolutivos, como mutações, seleção, isolamento genético, migrações ocorridas durante o curso da história, além da estrutura de casamentos e da deriva (COURT, 2021; SCHNEIDER; PRAINSACK; KAYSER, 2019; UNDERHILL; KIVISILD, 2007), bem como do progressivo aumento da população (COURT, 2021). Através destes eventos, marcadores genéticos podem apresentar frequências distintas em diferentes regiões geográficas, podendo ser mais comuns em algumas e raros em outras, tornando-se fontes altamente informativas sobre a provável localidade de sua origem (SCHNEIDER; PRAINSACK; KAYSER, 2019). A distribuição dos polimorfismos que não são impactados pela seleção natural reflete a história da mobilidade da espécie humana desde quando as populações se dispersaram da África, ao longo de milhares de anos (COURT, 2021).

Atualmente a análise forense através dos STRs autossômicos, que são marcadores genéticos que apresentam esta categoria de polimorfismos, é a abordagem padrão e o método de escolha nas análises genéticas investigativas (JOBLING, 2022). Adicionalmente, diversos estudos têm apontado para sua aplicabilidade na inferência de ancestralidade genética, com a recomendação de que, em tais análises, sejam associados outros marcadores de ancestralidade, em especial os AIMs (do inglês *Ancestry Informative Markers)* (JOBLING, 2022; PORRAS-HURTADO *et al.*, 2013).

Marcadores genéticos informativos de ancestralidade são, em sua maioria, representados por SNPs, os quais compõem, no universo forense, painéis já muito bem estabelecidos para possibilitar a inferência da origem biogeográfica, em populações não miscigenadas, de amostras coletadas em

locais de crime a nível continental (Europa, África Sub-Saariana, Ásia Oriental, Ásia Meridional, Oceania e América) (SCHNEIDER; PRAINSACK; KAYSER, 2019).

AIMs podem incluir marcadores autossômicos, de DNA mitocondrial e do cromossomo Y. Sua transmissão se dá através das gerações, cada qual por um mecanismo diferente, o que possibilita diferentes abordagens para compreensão da origem biogeográfica. Por exemplo, marcadores autossômicos refletem a origem de ambos os genitores simultaneamente, são tipicamente sujeitos à recombinação e os descendentes carregarão apenas metade dos respectivos marcadores parentais. No caso do cromossomo Y, os marcadores são transmitidos somente de pai para filho, o que, por sua vez, permite apenas a inferência da origem biogeográfica da linhagem paterna. Os marcadores do cromossomo Y são transmitidos essencialmente inalterados para as gerações seguintes conforme seu padrão de herança já mencionado, mantendo preservadas suas linhagens originais fortemente correlacionadas com regiões continentais (PHILLIPS, 2015; SCHNEIDER; PRAINSACK; KAYSER, 2019). Se por um lado existe essa forte correlação, marcadores do cromossomo Y podem não ser representativos da totalidade da ancestralidade dos indivíduos, justamente por sua característica uniparental relacionada a linhagem paterna, levando a possíveis erros de interpretação de origem biogeográfica. Este viés é particularmente impactante se considerarmos populações miscigenadas, que apresentam ancestralidades compartilhadas em decorrência de suas respectivas histórias demográficas (PHILLIPS, 2015). Por estes motivos, marcadores uniparentais têm sido alvo de extensos estudos com objetivo de reconstrução da ordem de ramificação genealógica. No entanto, cabe ressaltar que marcadores uniparentais não são informativos para determinar a linhagem paterna de mulheres.

Ao contrário do ideal, a ABG apresenta limitações e dificilmente pode ser definida com exatidão, limitando-se, na maior parte das vezes, a uma ancestralidade a nível continental (COURT, 2021). Além disso, marcadores uniparentais, em populações altamente miscigenadas como a brasileira, não são tão informativos sob o aspecto de uma possível correlação com fenótipos (PENA; BORTOLINI, 2004). A diversidade é estruturada geograficamente (JOBLING,

2012), portanto a qualidade da inferência da origem biogeográfica realizada depende tanto da informatividade dos marcadores genéticos utilizados quanto do conhecimento prévio dos dados genéticos da população utilizada como referência (SCHNEIDER; PRAINSACK; KAYSER, 2019). A complexidade da inferência da ancestralidade fica evidente através da análise dos dados genéticos e da avaliação do efeito que a distância geográfica ou a presença de barreiras geofísicas exercem sobre a livre circulação dos indivíduos e sobre o acasalamento aleatório. Durante os processos demográficos, tais barreiras definiram por algum tempo como se dariam os acasalamentos e influenciaram fortemente a estrutura populacional. Caso esta permanecesse intacta, os testes de ancestralidade forense realizados hoje refletiriam a exata ancestralidade verificada naquela época (PHILLIPS, 2015; UNDERHILL; KIVISILD, 2007). Com a progressiva redução do impacto destas barreiras, fica evidente que o espectro de ancestralidade deve estar bem representado e estabelecido na população referência para que possa fornecer inferências confiáveis (SCHNEIDER; PRAINSACK; KAYSER, 2019). Além disso, torna-se necessário estabelecer criteriosamente os grupos populacionais baseados em padrões de diferenciação de estrutura genética (PHILLIPS, 2015).

2.3 Componentes de Ancestralidade da População Brasileira

A trajetória de colonização do território brasileiro levou a uma composição tri-híbrida bastante heterogênea, abrangendo Indígenas, Europeus e Africanos em diferentes proporções, formando uma população significativamente miscigenada (CAROLINO *et al.*, 2019; PEREIRA *et al.*, 2020; RODRIGUES DE MOURA *et al.*, 2015). Variações nas proporções de contribuição destas ancestralidades são observadas nas diferentes regiões geopolíticas do país em virtude das diferenças na dinâmica da colonização do país (RODRIGUES DE MOURA *et al.*, 2015).

Embora não haja um consenso no tamanho da população nativo americana que habitava o país no início da colonização do Brasil, o número estimado de indígenas era de, aproximadamente, 3 milhões de indivíduos pertencentes a diversos grupos étnicos (JOERIN *et al.*, 2022). A partir da colonização, iniciou-

se o processo de miscigenação entre portugueses e indígenas (SOUZA *et al.*, 2019), além da ocorrência de conflitos e doenças. A população ameríndia é atualmente estimada em cerca de 1.693.535 de indivíduos (IBGE, 2023).

A população africana foi introduzida no Brasil sob a forma do comércio de escravizados que perdurou até o ano de 1870. No período, estima-se que cerca de quatro milhões de africanos originários de da costa Atlântica (África Central e ocidental, como Guiné, República Democrática do Congo, Angola, Moçambique e Nigéria), tenham aportado no país (PENA; BORTOLINI, 2004; SOUZA *et al.*, 2019). Sua presença impactante na região nordeste se deve à expansão do cultivo da cana de açúcar. Com o declínio desta e o início das atividades mineradoras no estado de Minas Gerais, durante o século XVII, observou-se a transferência de um expressivo número de africanos escravizados à região sudeste, o que justifica a presença atual de seus descendentes na região (SOUZA *et al.*, 2019).

A colonização europeia iniciou-se em 1500 com a vinda de portugueses, seguida de italianos, espanhóis e alemães. A chegada de imigrantes asiáticos e do oriente médio, originários principalmente do Japão, Líbano e Síria, teve início no século XX (SOUZA *et al.*, 2019).

Estima-se que no período compreendido entre 1500 e 1972, os imigrantes em terras brasileiras fossem constituídos essencialmente por 58% de europeus, 40% de africanos e 2% de asiáticos (SOUZA *et al.*, 2019). Variações nos processos de colonização e ocupação em terras brasileiras, possibilitaram que se instalassem, entre as cinco regiões geopolíticas do país, distintos graus de miscigenação (CALLEGARI-JACQUES *et al.*, 2003; RODRIGUES DE MOURA *et al.*, 2015; SOUZA *et al.*, 2019). Um estudo de revisão baseado em estimativas de ancestralidade advindas unicamente de marcadores autossômicos (INDELs, SNPs ou STR/VNTRs) evidenciou o predomínio de ancestralidade europeia no conjunto de todas as regiões brasileiras, com maior prevalência nas regiões Sul e Sudeste (RODRIGUES DE MOURA *et al.*, 2015; SOUZA *et al.*, 2019). A maior contribuição Africana é encontrada na região Nordeste (RODRIGUES DE MOURA *et al.*, 2015). A ancestralidade ameríndia apresenta a menor prevalência na população brasileira, sendo mais expressiva na região Norte (RODRIGUES

DE MOURA *et al.*, 2015; SOUZA *et al.*, 2019). Em geral, os estudos corroboram os resultados do Censo Demográfico de 2022 da distribuição do perfil de autodeclaração por cor ou raça nas regiões geopolíticas do Brasil, que demonstram no país o predomínio de indivíduos que se denominam negros e pardos, seguido de indivíduos brancos e, em menores proporções, amarelos e indígenas (RODRIGUES DE MOURA *et al.*, 2015; SOUZA *et al.*, 2019) (**Tabela 1**).

Entretanto, faz-se importante salientar que a autoclassificação étnico-racial no Brasil, utilizada por exemplo no censo demográfico, pode não refletir corretamente a ancestralidade genética (com exceção dos autodeclarados indígenas), principalmente se considerarmos as contribuições matrilineares ou patrilineares isoladamente (CAROLINO *et al.*, 2019; JOERIN *et al.*, 2022; SOUZA *et al.*, 2019).

Tabela 1 - Estimativa da Distribuição (%) de Indivíduos das Regiões Geopolíticas Brasileiras por Cor ou Raça Autodeclarada de Acordo com Censo Demográfico de 2022.

Unidades da Federação	Cor ou raça Branca (*)	Cor ou raça Parda (*)	Cor ou raça Preta (*)	Cor ou raça Indígena (*)	Cor ou raça Amarela (**)
NORTE	17,70	73,40	7,50	0,35	0,13
NORDESTE	24,70	63,10	11,40	0,25	1,57
CENTRO-OESTE	34,70	55,80	8,70	0,09	0,12
SUDESTE	50,70	38,70	9,60	0,06	1,48
SUL	75,10	19,90	4,40	0,04	0,52

Fonte: (*) IBGE, Censo Demográfico 2022. Disponível em: https://static.poder360.com.br/2022/07/populacao-ibge-2021-22jul2022.pdf (Primeiros Resultados). Acesso em Agosto/2023. (**) IBGE, Censo Demográfico 2010, Elaboração DIEESE. Disponível em: https://ecosol.dieese.org.br/ws2/tabela/economia-solidaria/estimativa-da-populacao-ocupada-por-cor-ou-raca. Acesso em Julho/2023.

A caracterização genética da ancestralidade de populações miscigenadas de três regiões geopolíticas do Brasil (Centro-Oeste, Sudeste e Sul) através do estudo de haplogrupos da porção não recombinante do cromossomo Y e do DNA mitocondrial evidenciou ainda, no Brasil, uma mistura genética assimétrica, com viés do sexo, na qual os indivíduos do sexo masculino apresentam em geral maior ascendência europeia e as mulheres, maior ascendência africana ou nativa americana, nas três regiões brasileiras, possivelmente em virtude do sistemático abuso, extensivamente documentado, de mulheres ameríndias e

negras escravizadas praticado pelos europeus durante o processo de colonização (BORTOLINI *et al.*, 2003; JOERIN *et al.*, 2022; PENA *et al.*, 2009).

2.4 Haplótipos e Haplogrupos do cromossomo Y

Neste estudo, utilizaremos o termo "haplótipos" para indicar os marcadores Y-STR, enquanto empregaremos os "Y-SNPs" quando houver referência aos haplogrupos. Estes termos encontram-se explicados a seguir.

Aproximadamente 95% da extensão do cromossomo Y (60 Mb) é denominada região não recombinante ou NRY (do inglês, *non-recombining Y*), sendo basicamente constituída por elementos repetitivos (ALI; HASNAIN, 2002; QUINTANA-MURCI; KRAUSZ; MCELREAVEY, 2001), inversões e sequências palindrômicas (COURT, 2021). A NRY é considerada o maior bloco não recombinante do genoma humano, sendo transmitida de forma intacta para os indivíduos do sexo masculino das gerações seguintes, num estado haploide (DE KNIJFF, 2000; QUINTANA-MURCI; KRAUSZ; MCELREAVEY, 2001; UNDERHILL; KIVISILD, 2007).

Centenas de marcadores polimórficos têm sido identificados ao longo de toda a NRY (CHEN *et al.*, 2021; DE KNIJFF, 2000; QUINTANA-MURCI; KRAUSZ; MCELREAVEY, 2001), tais como: (1) marcadores bialélicos com baixa taxa mutacional (na ordem de 10^{-9}), representados por variantes polimórficas bialélicas (Y-SNPs) e eventos de inserção e deleção (INDELS, do inglês *insertion-deletion events),* e que representam eventos mutacionais únicos ou UMEs (do inglês, *unique mutation events*) na evolução humana; (2) microssatélites (Y-STRs) de evolução moderadamente rápida, com frequência média de mutação (na ordem de 10^{-3}) (COURT, 2021; JANNUZZI *et al.*, 2020); e (3) minissatélites de evolução rápida.

Em virtude do mecanismo de sua transmissão sem recombinação e por linhagem paterna, uma vez que a mutação ocorre, ela permanece no pool genético. Por essa razão, a herança uniparental do cromossomo Y possui maior influência da deriva genética do que os autossomos, o que pode levar a

diferenças genéticas entre regiões geográficas simplesmente por acaso. Como consequência, marcadores do Y se tornam extremamente informativos para investigações sobre a história das populações e a origem humana (JORDAMOVIC´ *et al.*, 2021; KAYSER, 2017).

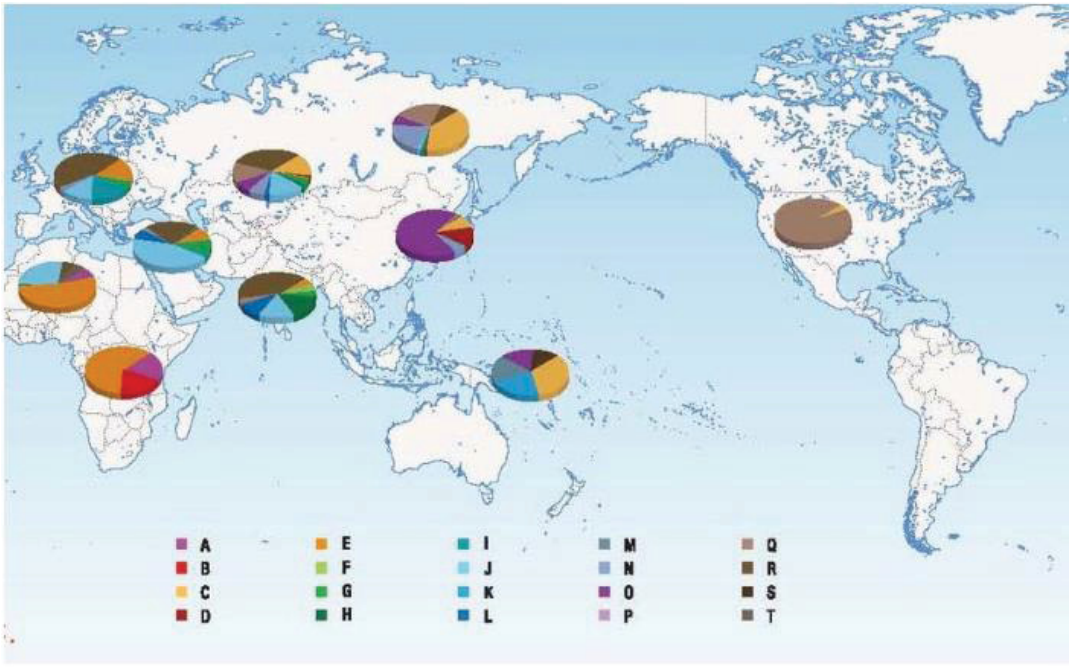
Blocos haplotípicos de Y-STRs exibem alta variabilidade e poder de discriminação, permitindo a diferenciação de cromossomos Y distintos entre si (COURT, 2021; DE KNIJFF, 2000; GUSMÃO *et al.*, 2017), sendo, por esta razão rotineiramente utilizados para investigações forenses de testes de paternidade, análises de parentesco e busca familiar. Sua aplicabilidade é particularmente importante em casos de violência sexual, quando há a contribuição de um agressor do sexo masculino e uma vítima de sexo feminino (KAYSER, 2017).

Haplogrupo do cromossomo Y é o termo empregado para se referir a uma linhagem que conecta os indivíduos, do sexo masculino, através de um evento mutacional binário em comum. Esta ligação também conecta estes indivíduos a um mesmo indivíduo ancestral no qual a marca mutacional se originou, indicando um vínculo por descendência. A cada novo evento mutacional subsequente desta natureza, novas conexões podem ser feitas, formando um novo braço desta linhagem (CHEN *et al.*, 2021; DE KNIJFF, 2000; KAYSER, 2017; QUINTANA-MURCI; KRAUSZ; MCELREAVEY, 2001). Os eventos mutacionais podem estar relacionados a indivíduos ou a pequenos grupos, porém seu grande impacto consiste no fato de que podem também definir haplogrupos preservados e profundamente enraizados filogeneticamente em distribuições geográficas mais ou menos amplas (KAYSER, 2017). De fato, muitos dos haplogrupos demonstram relação com sua distribuição geográfica, sendo esta a possibilidade que permite o rastreio de migrações e processos demográficos que moldaram as populações atuais (JORDAMOVIC´ *et al.*, 2021). Os Y-SNPs, em virtude de sua baixa taxa mutacional (na ordem de 10^{-9}) que caracteriza eventos moleculares únicos na história evolutiva humana, consistem em uma excelente abordagem para se compreender a origem de populações humanas e suas rotas migratórias (COURT, 2021; JANNUZZI *et al.*, 2020). Esses eventos mutacionais característicos possibilitam a detecção de haplogrupos que compartilham a mesma mutação, o que permite a modelagem de relacionamentos evolutivos das populações em árvores filogenéticas (COURT, 2021). Haplogrupos definidos por

Y-SNPs apresentam alta especificidade geográfica definindo origens com bastante precisão (JANNUZZI *et al.*, 2020).

Os haplogrupos são nomeados alfabeticamente, sendo que os mais antigos são os haplogrupos A e B, essencialmente restritos à África, suportando portanto a provável ancestralidade africana da população humana (COURT, 2021) (**Figura 2**).

Figura 2 - Representação geográfica da provável distribuição geográfica dos principais haplogrupos Y no mundo (Karafet et al., 2008).

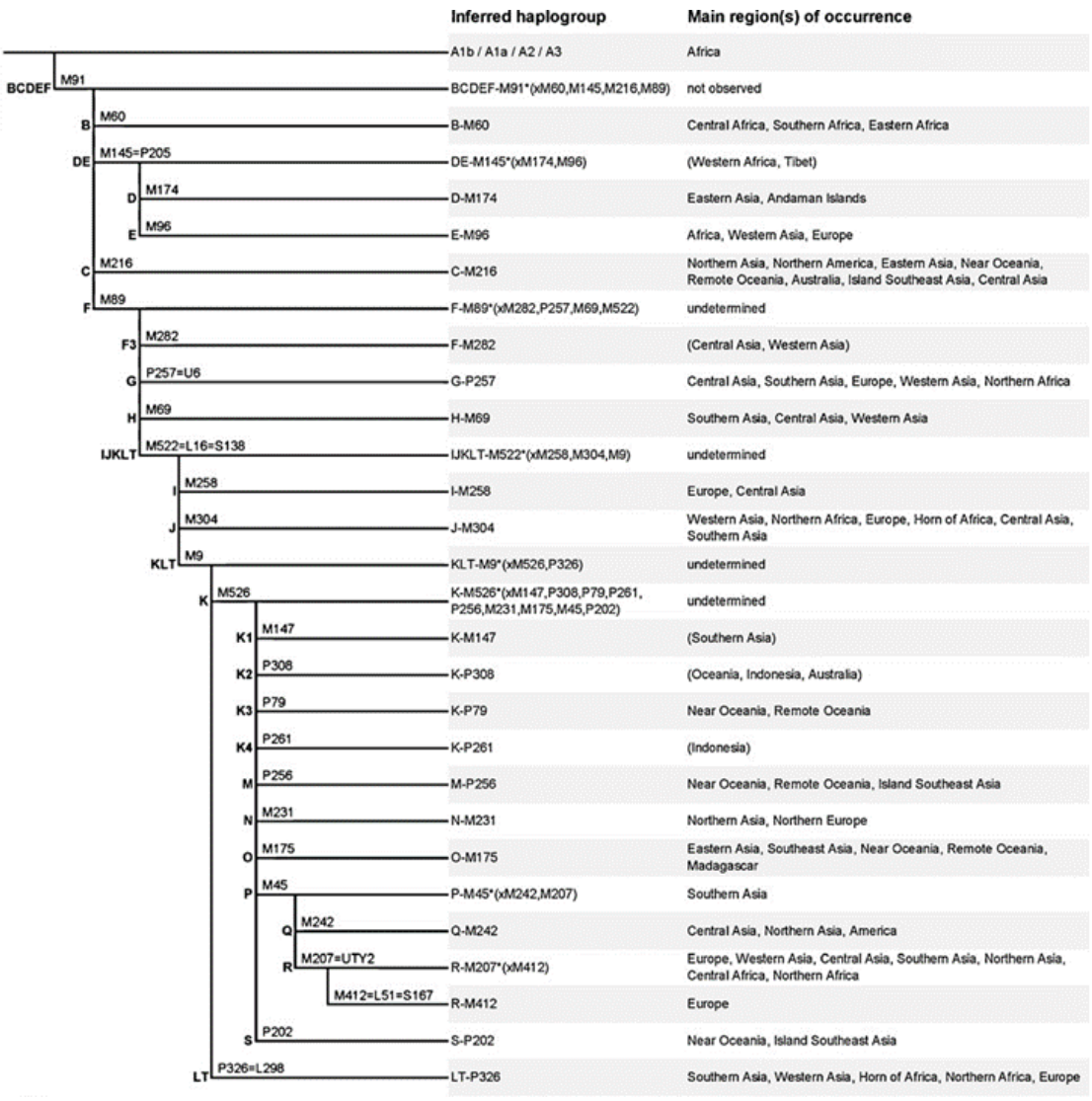


FONTE: (MELTON, 2008)

A filogenia do cromossomo Y já se encontra bem estabelecida, porém é importante considerar que ela se encontra em contínua expansão à medida que novas UMEs são descobertas através do advento de novas tecnologias (VAN

OVEN *et al.*, 2014). Sua representação costuma ser feita por meio de diagramas ou árvores filogenéticas (**Figura 3**) (ISOGG, 2020; KNIJFF, 2022).

Figura 3 - Representação da filogenia dos haplogrupos Y.



Legenda: Cada intersecção ou nó representa um evento mutacional, e cada braço da árvore filogenética representa uma linhagem que carrega a mesma mutação. Á direita são dispostos os haplogrupos dos mais antigos (que carregam a mutação a partir de um ancestral mais distante) aos mais recentes.

Fonte: VAN OVEN *et al.*, 2014.

Dentre os modelos mutacionais propostos para a dinâmica das mutações dos STRs, o modelo de mutação passo a passo (SMM, do inglês, *stepwise mutation model*) oferece a melhor representação da geração da diversidade nestes marcadores. Como a maioria das mutações nos STRs ocorrem com a perda ou ganho de uma unidade de repetição (KIMURA, 1978), depreende-se que os haplótipos resultantes representariam pequenas diferenças e menores distâncias genéticas, sendo relacionáveis entre si, o que admitiria seu agrupamento (BELEZA; LOPES; CARRACEDO, 2003; NEBEL *et al.*, 2001).

Possivelmente por esta razão, as análises dos marcadores genéticos citados (Y-STRs e Y-SNPs) demonstram que os indivíduos de um mesmo continente são mais similares geneticamente quando em comparação com aqueles de outros continentes (COURT, 2021).

Cada haplogrupo inclui cromossomos Y que compartilham o mesmo UME (mesmos Y-SNPs, por exemplo) (DE KNIJFF, 2000), porém pode conter muitos haplótipos Y-STR distintos. Isso decorre das diferentes taxas mutacionais apresentadas por Y-SNPs e Y-STRs de forma que, dentro de um haplogrupo, pode-se identificar uma rede (linhagens) abrangendo diversos haplótipos (DE KNIJFF, 2000; JANNUZZI *et al.*, 2020). Portanto, haplótipos Y-STR podem ser aplicados à inferência de haplogrupos (JANNUZZI *et al.*, 2020).

A diferenciação genética de populações humana como um todo se deve à fatores de dinâmica populacional, incluindo, entre outros, a deriva genética e casamento preferencial, associados ao expressivo aumento da população ao longo da história (COURT, 2021). Embora os marcadores STR do cromossomo Y possibilitem somente a *inferência* da ancestralidade biogeográfica, restrita à linhagem patrilinear, oferecem um tipo de polimorfismo pouco susceptível à seleção natural (COURT, 2021), e particularmente sujeito à deriva genética, principalmente ao efeito fundador (KAYSER, 2017; UNDERHILL *et al.*, 2001). Tais características conferem aos Y-STRs a capacidade de refletir a história da mobilidade da espécie humana desde sua dispersão a partir da África, milhares de anos atrás (COURT, 2021), elegendo-os como uma abordagem muito propícia para acessar as origens da diversidade populacional contemporânea (UNDERHILL *et al.*, 2001). Logo, a caracterização da diversidade haplotípica de STRs representa amplo interesse da comunidade científica (DE KNIJFF, 2000).

Por outro lado, é importante mencionar que a análise filogenética realizada através de haplótipos de Y-STRs apresenta, como limitação, a dificuldade na diferenciação de homoplasia, ou seja, entre haplótipos idênticos por estado ou por descendência, o que pode comprometer a compreensão das relações filogenéticas (BELEZA; LOPES; CARRACEDO, 2003). Isso reforça o fato de que, para inferências de haplogrupo, os marcadores Y-SNPs sejam considerados o padrão ouro.

2.5 Inferência de haplogrupos a partir de haplótipos Y-STRs

Como visto, os polimorfismos de SNPs e STRs do cromossomo Y, por sua capacidade de formar agrupamentos por similaridade (haplogrupos e haplótipos) podem ser utilizados para se inferir a ancestralidade biogeográfica de indivíduos. No entanto, as baixas taxas mutacionais conferem aos Y-SNPs maior adequação que os haplótipos Y-STR para a inferência de ancestralidade biogeográfica. Todavia, um haplogrupo pode apresentar correlação significativa com a diversidade de haplótipos de Y-STRs associados a ele (JANNUZZI *et al.*, 2020; KAYSER, 2017). A relação entre alelos de STR e haplogrupos foi descrita como sendo forte o suficiente para determinar a distribuição geográfica da população com base nos haplótipos, embora a definição com precisão de ramos filogenéticos de linhagem parental ainda requeira a validação através da análise de SNPs (PETREJČÍKOVÁ *et al.*, 2014). Assim, haplótipos de Y-STRs podem ser utilizados para inferir haplogrupos (JANNUZZI *et al.*, 2020; PETREJČÍKOVÁ *et al.*, 2014).

No entanto, a inferência de haplogrupos através de marcadores Y-SNPs é um processo mais dispendioso, demorado e trabalhoso, requerendo o uso de múltiplos marcadores para se chegar à definição do haplogrupo ao qual um indivíduo pertence. A rotina forense, por exemplo, não contempla a genotipagem Y-SNP devido a, entre outros fatores, quantidades limitadas de DNA obtido a partir de vestígios. Abordagens alternativas se concentram no uso do conjunto de marcadores Y-STR, ou seja, nos haplótipos (JORDAMOVIC´ *et al.*, 2021; KAYSER, 2017). Uma visão mais abrangente das informações geográficas contidas em um haplótipo de Y-STRs pode ser obtida através da busca pelo haplótipo de Y-STRs mais próximo no banco de dados de referência (como o YHRD), pois essa abordagem leva em consideração as etapas de mutação (KAYSER, 2017).

Em virtude de suas histórias evolutivas distintas, com base nos padrões contrastantes de fluxo gênico e deriva genética, os marcadores Y-STR apresentam variações em suas frequências alélicas, que têm o potencial de

conferir, a cada sublinhagem haplotípica, um padrão típico de polimorfismo, gerando diferenças entre haplótipos Y-STR (TARAZONA-SANTOS *et al.*, 2001; WATAHIKI *et al.*, 2019).

Durante a década de 2000, os cientistas buscaram definir os alelos mais frequentes de um conjunto de marcadores Y-STR no intuito de representar um haplótipo denominado haplótipo modal, que representaria o padrão de alelos para cada haplogrupo e poderia ser considerado o mais representativo de uma população, como os descritos haplótipo modal Cohen (*Cohen Modal Haplotype*, dentre a população judaica) e o haplótipo modal do Atlântico (*Atlantic Modal Haplotype*). Seria esperado que o haplótipo modal estivesse obrigatoriamente presente na população, podendo ser também seu haplótipo mais frequente (BORTOLINI *et al.*, 2003; THOMAS *et al.*, 2000; WILSON *et al.*, 2001). No entanto, compete referir que tais haplótipos modais foram estabelecidos para poucos marcadores, conforme permitido pelas tecnologias disponíveis na época. Atualmente, o número de marcadores genotipados aumentou em tal proporção que, mais do que configurar um haplótipo mais frequente na população, este identificaria um indivíduo. Assim, a ideia de definir um haplótipo padrão, característico de uma população, não apresentou continuidade.

Duas abordagens propostas com a finalidade de se obter a inferência do haplogrupo a partir de um conjunto de Y-STRs, como a medida da distância genética entre eles e o ajuste pela frequência alélica dentro de cada haplogrupo, são descritas a seguir (ATHEY, 2005, 2006).

Na inferência da ancestralidade obtida através da distância genética, o haplótipo questionado é comparado aos haplótipos de um banco de dados. Se for encontrado um haplótipo minimamente coincidente (a distância genética não deve ser maior que algum valor especificado, notadamente de duas diferenças), então o haplótipo questionado é atribuído ao mesmo haplogrupo do haplótipo referência. Na ausência de uma coincidência que atenda ao critério estabelecido, nenhuma estimativa de haplogrupo é feita. Embora a taxa de sucesso de predição relatada para esta abordagem seja alta (80%), a principal desvantagem do método é a possibilidade de que a predição não seja feita, mesmo que seja evidente que alguns haplogrupos poderiam ser descartados ou que o haplótipo

pertenceria a um haplogrupo de menor representatividade. Isto é, esse método não fornece uma inferência dos haplogrupos mais prováveis. Foi então elaborado um “ajuste” realizado através das frequências alélicas para cada haplogrupo e na compatibilidade entre o haplótipo questionado e o padrão de alelos de cada haplogrupo. Um exemplo desta aplicação pode ser constatado no caso do haplogrupo I1a, em que todo haplótipo identificado como I1a apresenta o alelo DYS455*8 (ou, raramente, os alelos 7 ou 9, mas nunca igual ou superior a 10). Esta abordagem se baseia em um algoritmo implementado em um programa publicamente disponibilizado desde 2004 (http://www.hprg.com/hapest5/), que retorna um índice de adequação aos haplogrupos inseridos na versão vigente. O acréscimo de marcadores Y-STRs adicionais não necessariamente oferece um índice de valor mais alto pois este estabiliza após cerca de algumas dezenas de marcadores (ATHEY, 2005).

Embora o método que aplica o "ajuste” da frequência alélica tenha sido razoavelmente bem-sucedido em indicar o haplogrupo, ele não oferece, de fato, uma predição ou probabilidade de que um haplótipo pertença a um determinado haplogrupo. A probabilidade de que um haplótipo de Y-STR esteja em um haplogrupo específico, pode ser melhor obtida através de uma abordagem bayesiana baseada nas frequências alélicas. A teoria bayesiana aplicada neste contexto possibilita calcular a probabilidade de um haplótipo específico estar associado a um dos diversos haplogrupos possíveis, com base em diferentes resultados de testes de marcadores Y-STRs, levando em consideração as probabilidades de obter cada resultado de teste, dado o haplótipo e as probabilidades prévias (*a priori*) de cada haplogrupo. Diferentemente da abordagem anteriormente descrita, com a abordagem bayesiana, mais marcadores geralmente melhoram a probabilidade (desde que os marcadores adicionados forneçam poder discriminatório real). Tipicamente, com 10 a 20 marcadores é possível elevar a probabilidade de um dos haplogrupos a um valor superior a 99% (ATHEY, 2006).

Naturalmente esta abordagem requer a disponibilização do maior número possível de dados dos haplótipos e das frequências alélicas em cada haplogrupo. Bancos de dados públicos por vezes disponibilizam dados de um haplótipo modal mínimo para um haplogrupo, a partir de estudos de Y-SNPs já

publicados, porém o cumprimento deste requisito representa o maior obstáculo para a plena implementação desta metodologia (ATHEY, 2005). Atualmente, para os principais haplogrupos, os dados disponíveis são abundantes. No entanto, para os haplogrupos minoritários e não europeus, os dados são muitas vezes insuficientes para incluir o haplogrupo, especialmente para os marcadores que não são frequentemente avaliados (ATHEY, 2006).

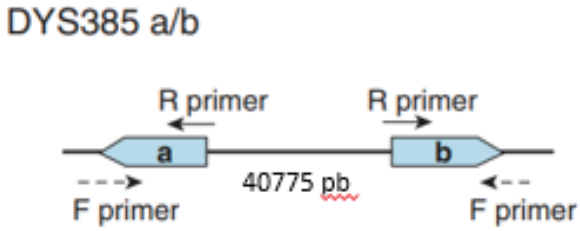
Originalmente, os primeiros conjuntos de marcadores Y-STR recomendados para uso para identificação humana eram constituídos por poucos marcadores, a exemplo do “haplótipo mínimo” composto por nove marcadores (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393 e DYS385a/b) (PURPS *et al.*, 2014). Uma vez que o acréscimo de marcadores adicionais pode aumentar o poder de discriminação do sistema e por recomendação do SWGDAM, marcadores adicionais foram gradativamente propostos (BALLARD *et al.*, 2006; PURPS *et al.*, 2014; YANG *et al.*, 2018). Atualmente, os *kits* de genotipagem de Y-STR com finalidade forense oferecem conjuntos constituídos por diversos marcadores, como, por exemplo, o sistema PowerPlex® Y23 (Promega), que contempla 23 marcadores Y-STR.

2.5.1 Marcador DYS385 a|b

Aqui convém mencionar que, devido às regiões duplicadas e palindrômicas do cromossomo Y, alguns locos Y-STR ocorrem mais de uma vez, como no caso do loco DYS385. Nesta região, um único conjunto de primers pode gerar dois produtos de amplificação, que podem ser interpretados como sendo 'dois locos' para um haplótipo do cromossomo Y. Especificamente, o loco Y-STR DYS385 está duplicado em duas regiões no braço longo do cromossomo Y, cerca de 40.000 pares de bases de distância (BUTLER, 2005). Cada região se apresenta de forma invertida com relação à outra e cada produto de PCR não pode ser inequivocamente atribuído a um loco específico (GUSMÃO *et al.*, 2006) (Figura 4). Ao ser amplificado com um único conjunto de primers, essas regiões geram dois alelos, rotulados como “a” (por convenção, referindo-se ao alelo menor) e “b” (**Figura *4*** e **Figura *5***). Devido a essa duplicação, o loco é chamado

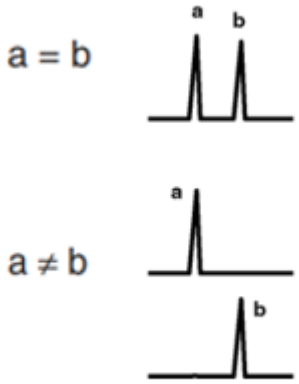
DYS385 a/b, sendo, no entanto, possível, que “a” e “b” tenham o mesmo tamanho, resultando em um único pico em um eletroferograma. Os alelos “a” e “b” podem ser amplificados separadamente através de uma abordagem de PCR realizada a partir dos produtos de uma PCR anterior (“*Nested PCR”)* (BUTLER, 2005; GUSMÃO *et al.*, 2006).

Figura 4 - Marcador DYS385a|b, ilustrando as duas regiões invertidas do cromossomo Y, separadas por aproximadamente 40 kb.



FONTE: BUTLER (2005)

Figura 5 - Demonstração dos resultados possíveis para genotipagem dos alelos do loco DYS385a|b.



FONTE: GUSMÃO (2006)

2.6 Ferramentas de predição de haplogrupos: o programa STRUCTURE como possibilidade

A inferência de haplogrupos a partir de Y-STRs pode ser realizada através de softwares de livre acesso especialmente desenvolvidos com este fim. Embora vários preditores tenham sido propostos, a preocupação com a precisão de cada software permanece, pois erros têm sido frequentemente reportados, inclusive para a população brasileira (EMMEROVA *et al.*, 2017; JANNUZZI *et al.*, 2020; PETREJČÍKOVÁ *et al.*, 2014). Preditores desta natureza calibram as frequências com base nos perfis depositados em bancos de dados, sendo os haplogrupos então inferidos com base no banco de dados de referência (MUZZIO *et al.*, 2011). Assim, os erros verificados podem estar relacionados ao número de marcadores Y-STR contemplados, a limitações decorrentes de alelos raros não aceitos pelos softwares, às populações testadas, a semelhanças entre haplótipos Y-STR de diferentes haplogrupos, as taxas mutacionais dos Y-STR, à profundidade de ramificações de haplogrupos, à convergência, à representatividade dos haplótipos nos bancos de dados ou, ainda, ao fato de que cada preditor utiliza um algoritmo próprio (EMMEROVA *et al.*, 2017; JANNUZZI *et al.*, 2020; MEDINA *et al.*, 2020).

O programa Haplogroup Predictor (disponível em http://www.hprg.com/hapest5/) (ATHEY, 2005, 2006) utiliza a abordagem Bayesiana de frequências alélicas para estimar a probabilidade de um haplótipo pertencer a um haplogrupo. Uma vez que as frequências haplotípicas podem variar de acordo com a região geográfica, a abordagem requer que esta seja fornecida. Há, entretanto, a possibilidade de selecionar a opção na qual todas as regiões apresentariam igual prioridade, sendo então consideradas em igual proporção. Este programa foi avaliado em algumas publicações, tendo apresentado bom desempenho para populações da República Tcheca (100%) e da Eslováquia (98,80%) (EMMEROVA *et al.*, 2017; PETREJČÍKOVÁ *et al.*, 2014). Na população brasileira, este software inferiu corretamente o haplogrupo em 83,89% das amostras, ao analisar os 26 marcadores Y-STR kit Yfiler™Plus kit (Thermo Fisher Scientific, Inc., Waltham, MA, USA) (JANNUZZI *et al.*, 2020).

O software *NevGen* Y DNA Predictor (CETKOVIC GENTULA; NEVSKI, 2015), disponível em https://www.*nevgen*.org/, utiliza dados do projeto FTDNA (do inglês *Family-tree DNA*), sendo, entretanto, somente considerados haplótipos compostos por Y-SNPs bem estabelecidos e profundos na árvore filogenética. Este programa não requer que seja indicada a provável região de

origem do haplótipo inserido(JANNUZZI *et al.*, 2020). Na população Tcheca, o software *NevGen* apresentou desempenho inferior ao *Haplogroup Predictor Hapest* (EMMEROVA *et al.*, 2017). Entretanto, na população brasileira, o desempenho foi superior (94,71%) (JANNUZZI *et al.*, 2020). É interessante registrar que o desempenho dos softwares preditores apresenta variações relacionadas, não somente ao algoritmo próprio de cada um dos softwares, como também aos conjuntos de marcadores contemplados pelos kits Y-STR. Este fato se justifica porque as predições são baseadas na associação de frequências haplotípicas dentro dos haplogrupos e cujos dados são disponibilizados pelos bancos de dados consultados. A baixa representatividade de alguns haplogrupos e de marcadores Y-STR de alguns kits nos bancos de dados impacta diretamente na precisão desta atribuição (JANNUZZI *et al.*, 2020).

O programa *STRUCTURE,* de livre acesso, é um sistema estatístico bastante difundido para estudos de estrutura populacional aplicável à análise de ancestralidade biogeográfica. O sistema baseia-se em dados populacionais de referência, incluindo genótipos multilocos, contra os quais os genótipos de ancestralidade desconhecida são comparados, permitindo a detecção de variações que admitem a realização de inferências de ancestralidade (KIDD *et al.*, 2011; PHILLIPS, 2015; PRITCHARD; STEPHENS; DONNELLY, 2000).

Este baseia-se em um algoritmo que utiliza a sistemática Bayesiana de *Monte Carlo via Cadeia de Markov* para inferir o número de grupos (K), desconhecido, em que um conjunto de populações pode ser organizado, a partir de análises das diferenças na distribuição de variantes genéticas entre eles (CHEVITARESE, 2009; PRITCHARD; STEPHENS; DONNELLY, 2000). Os indivíduos da amostra em análise são distribuídos probabilisticamente para cada K, de acordo com sua similaridade genética ou podem ser ainda atribuídos para mais de um K, caso os genótipos indiquem que sejam híbridos entre os grupos (PHILLIPS, 2015; PORRAS-HURTADO *et al.*, 2013; PRITCHARD; STEPHENS; DONNELLY, 2000).

O número de grupos (K) formados pode ser predeterminado, porém são os dados fornecidos que farão o programa definir este número (BATINI; JOBLING, 2017; PRITCHARD; STEPHENS; DONNELLY, 2000). Isto é, no cenário mais comum que consiste numa análise não supervisionada (na qual o algoritmo do programa é quem analisa e agrupa os dados), as amostras não são

rotuladas por região, sendo os agrupamentos atribuídos somente pelos padrões de similaridade genética detectados entre as amostras (PHILLIPS, 2015). Assim, os indivíduos em investigação são atribuídos de forma probabilística a um ou mais grupos, representando estes uma ou mais populações (BATINI; JOBLING, 2017; PRITCHARD; STEPHENS; DONNELLY, 2000).

Cada execução do *STRUCTURE* gera uma matriz com os coeficientes que indicam a origem populacional de cada amostra (referência ou desconhecida). De acordo com o modelo selecionado, os valores dos coeficientes são interpretados como probabilidades ou como frações de cada indivíduo atribuído a cada K, sendo passíveis de análise e comparação. Se os agrupamentos corresponderem bem à região de origem da amostra, então o painel de marcadores pode ser considerado informativo para os grupos analisados. Por essa razão, este processo é frequentemente utilizado para testar a eficiência de um conjunto de marcadores para diferenciar comparações de grupos populacionais específicos (EARL; VONHOLDT, 2012; KOPELMAN *et al.*, 2015; PURPS *et al.*, 2014).

Em análises de ancestralidade no programa *STRUCTURE*, nem sempre a estimativa de K que melhor se ajusta aos dados obtidos é concluída facilmente. Além disso, a interpretação da relação de K com a real estrutura genética nas populações analisadas pode ser mais complexa, sendo por vezes influenciável por informações *a priori* sobre a amostra em análise, como numa análise supervisionada (em que as amostras são rotuladas previamente à sua inserção). Tais situações requerem eventuais ajustes ao utilizar o programa ou a interpretar dados por ele gerados (PHILLIPS, 2015; PORRAS-HURTADO *et al.*, 2013). Questões pertinentes a respeito das análises desta natureza já foram pontuadas, como: (1) Amostragens de tamanhos heterogêneos e a distribuição das populações amostradas influenciam a formação dos agrupamentos e das probabilidades de melhor ajuste de K; e (2) As análises realizadas pelo *STRUCTURE* são estocásticas e, portanto, sujeitas a resultados frequentemente distintos entre corridas replicadas, mesmo quando a escolha de modelo e parâmetros é mantida inalterada (KIDD *et al.*, 2011; KOPELMAN *et al.*, 2015). Além disso, resultados são sensíveis ao tipo de marcadores genéticos usados (por exemplo, microssatélites), ao número de locos pontuados, ao número de populações amostradas e ao número de indivíduos genotipados em cada

amostra (EVANNO; REGNAUT; GOUDET, 2005). Em teoria, almeja-se um bom ajuste de agrupamentos aos dados populacionais. A recomendação é que, ao invés de buscar a melhor corrida ou o melhor padrão encontrado, se busque a primeira probabilidade mais estável, com um menor K, não permitindo que a interpretação seja influenciada por suposições sobre as populações amostradas (PORRAS-HURTADO *et al.*, 2013). De qualquer forma, é comum que o usuário se depare com o desafio de resumir e comparar centenas, e às vezes milhares, de execuções, dentro e entre os valores de K.

O valor de K mais provável costuma ser estimado através de programas complementares, como o *CLUMPAK - Clustering Markov Packager Across K*, o qual oferece uma ilustração gráfica dos resultados obtidos pelo *STRUCTURE* (KOPELMAN *et al.*, 2015). Recursos adicionais permitem a seleção do valor preferido de K (EVANNO; REGNAUT; GOUDET, 2005).

O programa *STRUCTURE* permite identificar as populações a partir de dados ou também atribuir indivíduos a populações que representem o melhor ajuste para os padrões de variação apresentados, uma vez que sua metodologia permite comparar as frequências alélicas que definem as populações com os alelos encontrados nos indivíduos (PORRAS-HURTADO *et al.*, 2013).

Quando comparado a outros programas de análise populacional, o *STRUCTURE* pode não garantir, por exemplo, a maior celeridade nas análises, porém oferece flexibilidade mais expressiva por se adaptar a diferentes demandas de análise, como, por exemplo, sua capacidade de lidar simultaneamente com genótipos de STRs e SNPs (PHILLIPS, 2015; PORRAS-HURTADO *et al.*, 2013).

A relevância do programa *STRUCTURE* nas investigações forenses inclui a verificação da ausência ou presença de estrutura populacional, a atribuição de indivíduos a populações e a identificação de migrantes (PORRAS-HURTADO *et al.*, 2013).

2.7 Comportamento de Misturas Forenses em Termos de Diversidade

Como mencionado anteriormente, o processo de colonização do território brasileiro culminou com a prevalência de uma população miscigenada

constituída essencialmente pelas ancestralidades europeia, africana e ameríndia, em diferentes proporções distribuídas pelo país. Tais populações ancestrais apresentam histórias demográficas próprias que refletiram em sua diversidade genética.

Estudiosos concordam que a linhagem do *Homo sapiens* moderno teve origem na África há 500-300 mil anos. Por este motivo, considera-se que os genomas africanos abrigam a maior diversidade genética e fenotípica da humanidade (PEREIRA *et al.*, 2021).

Medidas de diferenciação genética entre populações são menores na Europa do que em outros continentes, evidenciando que os europeus, de maneira geral, se apresentam como um povo mais homogêneo. Está também descrito um gradiente genético que se inicia no Oriente Médio e se direciona ao noroeste através dos Bálcãs e da Europa central até a França, Inglaterra e Escandinávia, resultado de mistura contínua e parcial dos agricultores em expansão com os caçadores-coletores locais (CAVALLI-SFORZA; PIAZZA, 1993).

Estudos envolvendo os povos indígenas indicam que estes apresentam a menor diversidade genética nas Américas, em relação a outras regiões continentais, o que sugere uma redução populacional associada a um estabelecimento geograficamente mais restrito, representando gargalos recentes ocorridos durante as migrações humanas a partir da Ásia (WANG *et al.*, 2007). No contexto da colonização primordial na América do Sul, os primeiros grupos indígenas brasileiros colonizaram ambientes muito variados, como a floresta amazônica, florestas atlânticas e cerrado, dentre outros (PENA; SANTOS; TARAZONA-SANTOS, 2020). Embora as características genéticas destes povos nativos ainda não tenham sido plenamente estabelecidas, diversos estudos apontam que isso se refletiu em uma ampla diversidade inicial dos indígenas, inferida a partir de mais de mil línguas indígenas distintas originalmente faladas. No entanto, expressiva parte desta variação inicial pode ter se perdido, como sugerido, com extinção de alguns grupos indígenas e suas línguas, em virtude dos efeitos advindos da colonização (JOERIN *et al.*, 2022; PENA; SANTOS; TARAZONA-SANTOS, 2020; SALZANO; SANS, 2014). Este

fato se sobrepõe aos efeitos da colonização humana a partir da África (WANG *et al.*, 2007).

Com base nessas considerações, pressupõe-se que a diversidade variável dos povos ancestrais possa influenciar a forma como se visualiza seus marcadores genéticos quando estes se encontram em uma mistura de material biológico proveniente de dois indivíduos. Por exemplo, espera-se que misturas genéticas formadas por dois indígenas de uma mesma região geográfica evidenciem menor diversidade do que misturas formadas por dois africanos, cuja diversidade genética é maior. Neste contexto, seria legítimo questionar se as diferenças nas diversidades, quando constatadas nas misturas, poderiam sugerir o haplogrupo Y-SNP de seus contribuintes. Este questionamento torna-se particularmente relevante ao se considerar populações nas quais a miscigenação pode refletir a diversidade dos povos originais.

Ao considerar a miscigenação da população brasileira e as consequências da diversidade genética das populações ancestrais, bem como a frequência de crimes sexuais que frequentemente resultam em amostras contendo material biológico de mais de um indivíduo, é relevante analisar como se comportam as misturas genéticas em amostras forenses sob a perspectiva da diversidade. Isso inclui examinar como a diversidade se manifesta e potencialmente revela padrões nas misturas, tanto naquelas envolvendo dois indivíduos da mesma população quanto naquelas de populações diferentes. Em essência, interessa investigar se a análise dos efeitos da diversidade genética em misturas de indivíduos miscigenados com ancestralidades paternas iguais ou distintas pode proporcionar percepções significativas que auxiliem na inferência da origem de cada indivíduo.

2.8 Correlação de fenótipos e origem biogeográfica no Brasil

Por várias razões, incluindo fatores demográficos ocorridos no território brasileiro, houve uma intensa mistura genética entre colonizadores, colonizados e escravizados desde os primeiros contatos. A extensão dessa mistura tem sido avaliada ao longo dos séculos, frequentemente utilizando critérios relacionados

à aparência física, que passou a ser considerada um dos pilares da discussão sobre a origem geográfica humana. Isso envolve associar características físicas externas, como cor da pele, textura do cabelo e formatos de lábios e de nariz, a grupos específicos, como os negros ou brancos. No Brasil, o termo "cor" é comumente usado nesse contexto. Tais diferenças fenotípicas correlacionam-se bem com o continente de origem quando se tratam de populações mais homogêneas. No entanto, não refletem variações genômicas que possam ser generalizadas entre os grupos populacionais miscigenados (PENA; BORTOLINI, 2004).

Reformulando a afirmação de Pena et al (2004), é evidente que seria difícil confundir pessoas típicas oriundas de localidades como Botswana, Noruega ou nativos americanos. No caso descrito anteriormente, ocorrido na Holanda, a inferência de ancestralidade do criminoso fez bastante sentido, pois envolve uma população com predominância fenotípica marcante. No entanto, estudos filogeográficos com populações urbanas brasileiras, através de marcadores uniparentais, revelou que a maior parte das patrilinhagens é europeia enquanto a maioria das matrilinhagens é ameríndia ou africana, evidenciando um nítido padrão de reprodução assimétrico (JOERIN et al., 2022; PENA; BORTOLINI, 2004; SALZANO; SANS, 2014), o que se reflete de maneira muito aleatória nas características externamente visíveis da população. Tais fatores impactam, inclusive, nos critérios utilizados conforme o senso comum para a autodeclaração de raça, utilizada por exemplo, nos censos nacionais (IBGE, 2023; JOERIN et al., 2022; KAYSER et al., 2023; PENA; BORTOLINI, 2004).

Dessa forma, enquanto marcadores genéticos autossômicos oferecem uma classificação mais precisa da ancestralidade genômica a nível individual nas populações miscigenadas, o mesmo não se observa com os marcadores de ancestralidade uniparentais ou nas características físicas consideradas marcantes da ancestralidade dos indivíduos (PENA; BORTOLINI, 2004). Em outras palavras, em populações que apresentam alto grau de miscigenação, como a brasileira, as características externas visíveis não deveriam ser consideradas como evidências da origem africana, europeia ou indígena dos indivíduos. Nesse contexto, os haplogrupos de Y-SNPs na população miscigenada brasileira refletem uma diversidade proveniente de várias origens

continentais, tornando-os indicadores imprecisos da origem específica dos indivíduos, a menos que o agressor apresente um haplogrupo ainda não relatado na população brasileira.

Em um cenário ideal, a implementação de um método de inferência de ancestralidade utilizando marcadores uniparentais acrescentaria informações indicativas de ancestralidade compartilhada às já obtidas por meio dos marcadores autossômicos de ancestralidade (KAYSER et al., 2023).

JUSTIFICATIVA E HIPÓTESES

A inferência de haplogrupos de Y-SNPs a partir de haplótipos compostos por Y-STRs pode ser realizada por *softwares* desenvolvidos com este propósito, provendo informações de patrilinhagem de indivíduos. Contudo, erros de predição são inerentes a todas as ferramentas e já foram reportados para diferentes populações. Considerando-se a história de formação e miscigenação da população brasileira, esta se torna especialmente desafiadora para os propósitos de predição aqui propostos.

Para amostras de fonte única, isto é, constituídas pelo perfil genético de apenas um indivíduo, duas ferramentas de predição de livre acesso já foram testadas e avaliadas em um subconjunto de brasileiros (*HAPEST Haplogroup Predictor* e *NevGen*), mas apresentam taxas de erro elevadas para serem utilizadas em análises forenses. Portanto, a proposta do atual trabalho é avaliar, neste tipo de amostra, a performance de um outro programa, o *STRUCTURE*, já amplamente utilizado em estudos populacionais, mas que até o momento não foi utilizado como preditor de haplogrupos a partir de haplótipos de Y-STR.

Além disso, a presença frequente de perfis genéticos múltiplos em amostras forenses, provenientes simultaneamente de dois indivíduos do sexo masculino, representa um desafio constante na identificação de seus contribuintes individuais. Esse cenário é particularmente comum em vestígios relacionados a agressões sexuais, cuja incidência é consideravelmente elevada no país. A complexidade aumenta quando não há indicação de quem possam ser os possíveis contribuintes, resultando de agressões perpetradas por desconhecidos, o que dificulta as investigações criminais na identificação de suspeitos. Os bancos de perfis genéticos, criados para possibilitar a identificação de suspeitos ou de outras vítimas do mesmo autor em amostras de autoria desconhecida, nem sempre são efetivos em virtude da falta de representatividade dos perfis genéticos de criminosos em seus arquivos. Nestes casos, não apenas torna-se inviável identificar a fonte desses perfis, mas também impossibilita inferir qual a haplogrupo a cada contribuinte pertenceria. No entanto, a aplicabilidade do programa *STRUCTURE* para perfis múltiplos,

assim como de outros programas existentes, é limitada, uma vez que essas ferramentas operam exclusivamente com perfis únicos.

Naturalmente eventuais suspeitos apontados através destas abordagens deverão ser submetidos a confronto de seus perfis genéticos com aqueles perfis obtidos dos vestígios. Isto é, a inferência do haplogrupo não configura a identificação definitiva de um suspeito. A confirmação de seu vínculo genético com um vestígio deve seguir os ritos de confronto genético já amplamente estabelecidos e praticados na rotina forense.

Assim, a hipótese deste trabalho é de que a ferramenta *STRUCTURE,* por trabalhar com método e algoritmo diferente dos demais preditores, possa mostrar um desempenho superior e ser eficientemente utilizada para predizer o haplogrupo paterno a partir de Y-STRs na população urbana e indígena brasileira.

Para casos de perfis múltiplos encontrados em vestígios forenses, constituídos por dois contribuintes simultâneos, nos quais a aplicação do programa *STRUCTURE* não seria viável, a proposta consiste em explorar a diversidade populacional nas regiões geopolíticas brasileiras estudadas. Este conhecimento forneceria subsídios para detecção de pertencimento de um agressor eventualmente pertencente a um grupo populacional que não faça parte daquela população brasileira.

Logo, a hipótese inclui a possibilidade de extrair informações da diversidade de populações de origem parental (haplogrupo de Y-SNP) europeia, africana ou ameríndia (as mais prevalentes no Brasil), a partir das médias da diversidade observadas em vestígios apresentando perfis múltiplos que envolvam dois homens de origem desconhecida.

3. OBJETIVOS

4.1 Objetivo Geral

Avaliar se o programa *STRUCTURE* se mostra um bom preditor de haplogrupos a partir de haplótipos de Y-STR em amostras de fonte única, de maneira tal que possibilite seu uso na inferência de haplogrupos Y-SNPs em indivíduos da população urbana e indígena brasileira.

4.2 Objetivos Específicos

- Caracterizar haplotipicamente indivíduos oriundos de populações urbanas brasileira (Mato Grosso do Sul, São Paulo e Paraná) e indígena (Guarani e Kaingang) para o conjunto de 23 marcadores Y-STR inclusos no sistema *PowerPlex® Y23* (Promega);
- Avaliar a performance do programa *STRUCTURE* na inferência de haplogrupos Y-SNP a partir dos haplótipos Y-STR para o referido conjunto amostral, cujos haplogrupos encontram-se estabelecidos, comparando-a com a performance de preditores de livre acesso disponíveis (*HAPEST Haplogroup Predictor* e *NevGen Y-DNA Predictor*);
- Estimar as frequências alélicas, diversidade gênica e haplotípica para os 23 marcadores STRs;
- Avaliar estatisticamente a eficiência forense (probabilidade de coincidência e poder de discriminação) do referido sistema nas populações citadas;
- Avaliar a possibilidade de, baseando-se em padrões de diferenças intra-locus encontradas em misturas, inferir o haplogrupo de pertencimento dos contribuintes de um perfil múltiplo contendo dois contribuintes do sexo masculino, para auxiliar na elucidação por abordagem qualitativa de casos de misturas forenses;
- Contribuir com bancos de dados públicos de cromossomo Y para fins de uso acadêmico e forense.

5. MÉTODOS

5.1 Amostras populacionais

Fizeram parte deste estudo 180 indivíduos do sexo masculino, sendo 172 urbanos e 8 indígenas. Os urbanos são provenientes das regiões Centro-Oeste (Mato Grosso do Sul, n=125), Sudeste (São Paulo, n=17), Sul (Paraná, n=24), Norte (Rondônia e Roraima, n=2), Paraguai (n=1) e 3 não possuem origem geopolítica informada. Os indivíduos foram subdivididos em afrobrasileiros, eurobrasileiros e indígenas com base em trabalho anterior de nosso grupo. Um indivíduo foi classificado como asiático. Compuseram ainda a amostra oito indígenas oriundos de reservas indígenas do Paraná (região Sul), das etnias Kaingang (n=4) e Guaraní Mbya (n=4) das Reservas do Rio das Cobras (25°18'S, 52°32'O) e Ivaí (24°30'S, 51°40'O) (AUGUSTO *et al.*, 2021; PETZL-ERLER; LUZ; SOTOMAIOR, 1993; TSUNETO *et al.*, 2003).

As coletas de material biológico e os procedimentos de extração de DNA foram realizados pela equipe do Laboratório de Genética Molecular Humana da Universidade Federal do Paraná, conforme descrito detalhadamente por Joerin et al. (2022). Este estudo foi aprovado pelo Comitê Nacional de Ética em Pesquisa com Seres Humanos (CONEP-Comissão Nacional de Ética em Pesquisa), sob o protocolo número 02727412.4.0000.0096, conforme Leis Federais Brasileiras, sendo que todos os indivíduos forneceram autorização expressa para a coleta segundo a Declaração de Helsinki (AUGUSTO *et al.*, 2021). Para a utilização do material biológico dos indígenas, o projeto também foi aprovado pela Fundação Nacional do Índio (FUNAI) e pelos líderes das populações indígenas estudadas. As amostras das populações indígenas foram coletadas anteriormente à resolução CNS 196/96 do Ministério da Saúde, quando ainda não se obtinha o consentimento informado por escrito, e seguiram o código de conduta de informação prévia dos participantes e de seus líderes comunitários, participação voluntária e garantia de anonimato (PETZL-ERLER; LUZ; SOTOMAIOR, 1993; TSUNETO *et al.*, 2003).

Também foram usadas neste trabalho informações genotípicas das amostras que compõem o painel do Projeto Diversidade do Genoma Humano (HGDP, do inglês, *Human Genome Diversity Project)* do Centro de Estudos do

Polimorfismo Humano (CEPH, do francês, *Centre D'Êtude du Polymorphisme Humain*). O HGDP-CEPH inclui 1.064 indivíduos oriundos de 52 populações mundiais (distribuídas em 7 regiões geográficas: África, Europa, Norte da África, Oriente Médio, Ásia (Central/Sul/Oeste), Oceania e Américas (Nativos) (CANN, 2002) (**Figura *6***). Foram coletados dados geográficos (país de coleta) e genéticos, já publicados, referentes a 573 indivíduos do sexo masculino deste painel. Os dados genéticos se referem àqueles gerados a partir do DNA originalmente extraído de linhas de células linfoblastoides humanas e contemplam tanto os microarranjos de SNPs de alta resolução, quanto os haplótipos obtidos pela análise de 23 marcadores Y-STR (BERGSTROM *et al.*, 2020; HALLAST *et al.*, 2021).

Figura 6 - Origem geográfica das 54 populações do painel HGDP-CEPH.

Fonte: BERGSTROM *et al.* (2020).

5.2 Genotipagem de Y-SNPs e classificação dos haplogrupos

A genotipagem dos Y-SNPs nas 180 amostras brasileiras foi feita em estudo prévio do nosso grupo que analisou SNPs em genoma total através de microarranjos de alta resolução da plataforma Illumina CoreExome-24 v1.1 (Illumina, San Diego, CA) (AUGUSTO *et al.*, 2021). A filtragem e seleção do conjunto de SNPs do cromossomo Y, a lista completa dos 2009 Y-SNPs utilizados, bem como o procedimento para classificação dos haplogrupos foram feitas anteriormente e estão descritos em Joerin et al. (2022). De maneira breve, a filtragem dos SNPs do cromossomo Y foi realizada com PLINK v1.9.0 (CHANG *et al.*, 2015). A classificação do haplogrupo Y foi realizada por meio de yHaplo™ (POZNIK, 2016) considerando as mutações definidoras de haplogrupo descritas pela Sociedade Internacional de Genealogia Genética (ISOGG, do inglês, *International Society of Genetic Genealogy - Y-DNA Haplogroup Tree* 2019) (ISOGG, 2020").

Para fins de desambiguação, o termo clado será usado para designar as linhagens principais (por exemplo, E, L, M, N, Q, R), e o termo haplogrupo (Hg) indicará seus derivados (por exemplo, R1a, E1b1). As proporções de ancestralidade uniparental foram estimadas diretamente com base na especificidade biogeográfica de cada haplogrupo, de acordo com trabalho previamente (JOERIN *et al.*, 2023).

5.3 Genotipagem dos Y-STRs e determinação dos haplótipos

A genotipagem dos Y-STR nas amostras brasileiras foi realizada nas dependências do Laboratório de Genética Molecular Forense da Polícia Científica do Estado do Paraná. Foram analisados 23 locos STRs contemplados no *kit* PowerPlex® Y23 (Promega) (DYS19, DYS385a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS481, DYS533, DYS549, DYS570, DYS576, DYS635, DYS643 e YGATAH4). A reação foi efetuada segundo as recomendações do fabricante do *kit*. Brevemente, as amostras foram ajustadas à concentração de

0,25ng/µL de DNA e foram amplificadas em reação multiplex em termociclador Veriti™ Dx Thermal Cycler (Applied Biosystems). Todas as reações de amplificação foram devidamente acompanhadas de um controle positivo (2800M Control DNA, fornecido pelo fabricante do *kit*) e de um controle negativo (água ultrapura que acompanha o *kit* de amplificação). Alíquotas de 1µL das amostras amplificadas foram diluídas em 9,6µL de formamida Hi-Di™ e 0,4µL do marcador interno WEN Internal Lane Standard 500 Y23, desnaturadas à 95ºC em termociclador Veriti™ Dx Thermal Cycler (Applied Biosystems), e, então, submetidas à corrida eletroforética capilar com polímero POP-4, seguida de detecção da fluorescência laser-induzida no analisador automático de DNA ABI PRISM 3500 (Applied Biosystems). A análise individual e comparativa dos eletroferogramas resultantes foi realizada pelo Genemapper ID-X Software v. 1.4 (Applied Biosystems) com o uso do PowerPlex® Y23 Allelic Ladder Mix (Promega) correspondente. Frente à eventual perda de amplificação dos marcadores de uma amostra, uma tentativa de reamplificação foi realizada.

5.4 Classificação dos haplogrupos

As mutações no cromossomo Y que definem os haplogrupos constituem variações nucleotídicas únicas (SNVs) (HALLAST *et al.*, 2014). Quando definem haplogrupos principais são consideradas “mutações definidoras” ou SNVs terminais, enquanto as demais mutações marcam linhagens dentro de um haplogrupo principal, sendo consideradas “mutações internas” (KARAFET *et al.*, 2008). Todos os indivíduos do painel HGDP-CEPH já apresentavam a informação de genotipagem de múltiplas variantes do Y. Portanto, a partir destas, foram identificadas as SNVs terminais, as quais foram utilizadas para identificar o respectivo haplogrupo a partir do painel disponibilizado pelo ISOGG (ISOGG, 2020). Na sequência, o haplogrupo do cromossomo Y de cada amostra foi utilizado para estimar a origem biogeográfica da linhagem paterna conforme publicações disponíveis (ASHRAF A EWIS, JUWON LEE, TOSHIKATSU SHINKA, 2002; AUTON *et al.*, 2015; BERGER *et al.*, 2013; BERGSTROM *et al.*, 2020; ELKAMEL *et al.*, 2021; FEHÉR *et al.*, 2015; GRUGNI *et al.*, 2019; HALLAST *et al.*, 2014; HAMMER *et al.*, 2009; HUANG *et al.*, 2018; ILUMÄE *et al.*, 2016; KARAFET *et al.*, 2008; MENDEZ *et al.*, 2013; MIZUNO *et al.*, 2010;

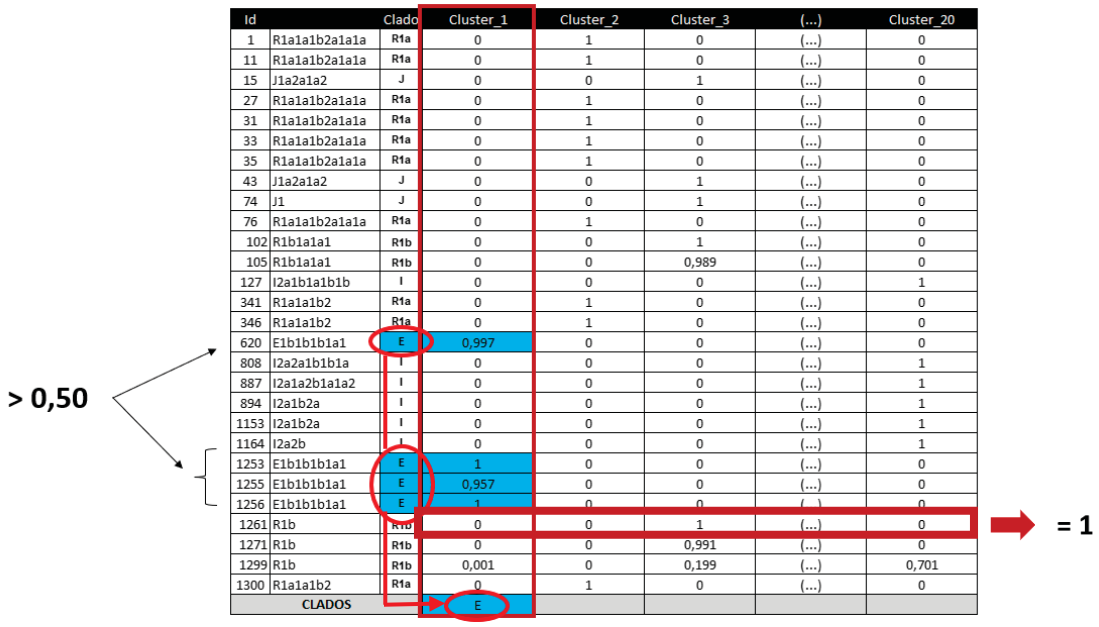
MOUSSA *et al.*, 2018; NAIDOO *et al.*, 2010; PURPS *et al.*, 2014; ROHRLACH *et al.*, 2021; ROOTSI *et al.*, 2004, 2007; SAHAKYAN *et al.*, 2021; SCOZZARI *et al.*, 2012; SOLÉ-MORATA *et al.*, 2017; TROMBETTA *et al.*, 2011, 2015; UNDERHILL *et al.*, 2015; VAN OVEN *et al.*, 2014; YANG *et al.*, 2010; ZHONG *et al.*, 2011).

5.5 Inferência de haplogrupos Y-SNPs a partir dos haplótipos de Y-STRs

A inferência de haplogrupos de todos os participantes brasileiros deste estudo (POPFLAG = 0), tomando como referência os indivíduos do HGDP-CEPH organizados de acordo com os haplogrupos pré-definidos (POPFLAG = 1), foi feita com o programa *STRUCTURE* (PRITCHARD; STEPHENS; DONNELLY, 2000) sendo realizada com base em resultados multilocos assumindo frequências alélicas correlacionadas no modelo que não prevê miscigenação (no *admixture model*), habilitando a função LOCPRIOR, com *burn-in* de 200.000 iterações coletadas via Cadeia de Markov (MCMC), 200.000 repetições e com randomização desativada. Nesta análise o valor de *K* variou de dois a 30, com 15 iterações independentes para cada valor. Para a análise do melhor K foi considerado o valor de K variando de 14 a 30. Foram empregados somente os 21 marcadores de Y-STRs (DYS19, DYS385a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS481, DYS533, DYS549, DYS570, DYS576, DYS635, DYS643 e YGATAH4) que apresentam um alelo apenas enquanto o loco DYS385a|b, por apresentar duas regiões amplificadas foi, portanto, excluído destas análises. A escolha do valor mais provável de *K* foi realizada utilizando o programa *CLUMPAK - Cluster Markov Packager Across K* (KOPELMAN *et al.*, 2015), segundo os métodos de Evanno e Pritchard (EVANNO; REGNAUT; GOUDET, 2005; PRITCHARD; STEPHENS; DONNELLY, 2000) a partir das cinco melhores corridas por apresentarem a maior probabilidade de acerto, conforme o maior valor de LnP(D) obtido. O programa *Distruct* foi empregado para produzir gráficos em barras coloridas apenas com a melhor corrida para cada K. A verificação das atribuições individuais de haplogrupos foi realizada utilizando-se o programa *Excel*.

Os resultados apresentados pelo *STRUCTURE* indicam probabilidades de pertencimento de cada indivíduo a diferentes clusters, variando de 0 a 1. Inicialmente, desconhecia-se qual o clado que estaria representado por cada cluster. As probabilidades acima de 0,50 foram consideradas como fortemente indicativas do pertencimento do indivíduo ao cluster onde este valor foi obtido. Neste contexto, uma vez que o clado de cada indivíduo já era conhecido, tornou-se possível identificar qual clado estava sendo representado por cada cluster (**Figura 7**).

Figura 7 - Representação dos resultados apresentados pelo *STRUCTURE* e sua interpretação.



Id		Clado	Cluster_1	Cluster_2	Cluster_3	(...)	Cluster_20
1	R1a1a1b2a1a1a	R1a	0	1	0	(...)	0
11	R1a1a1b2a1a1a	R1a	0	1	0	(...)	0
15	J1a2a1a2	J	0	0	1	(...)	0
27	R1a1a1b2a1a1a	R1a	0	1	0	(...)	0
31	R1a1a1b2a1a1a	R1a	0	1	0	(...)	0
33	R1a1a1b2a1a1a	R1a	0	1	0	(...)	0
35	R1a1a1b2a1a1a	R1a	0	1	0	(...)	0
43	J1a2a1a2	J	0	0	1	(...)	0
74	J1	J	0	0	1	(...)	0
76	R1a1a1b2a1a1a	R1a	0	1	0	(...)	0
102	R1b1a1a1	R1b	0	0	1	(...)	0
105	R1b1a1a1	R1b	0	0	0,989	(...)	0
127	I2a1b1a1b1b	I	0	0	0	(...)	1
341	R1a1a1b2	R1a	0	1	0	(...)	0
346	R1a1a1b2	R1a	0	1	0	(...)	0
620	E1b1b1b1a1	E	0,997	0	0	(...)	0
808	I2a2a1b1b1a	I	0	0	0	(...)	1
887	I2a1a2b1a1a2	I	0	0	0	(...)	1
894	I2a1b2a	I	0	0	0	(...)	1
1153	I2a1b2a	I	0	0	0	(...)	1
1164	I2a2b	I	0	0	0	(...)	1
1253	E1b1b1b1a1	E	1	0	0	(...)	0
1255	E1b1b1b1a1	E	0,957	0	0	(...)	0
1256	E1b1b1b1a1	E	1	0	0	(...)	0
1261	R1b	R1b	0	0	1	(...)	0
1271	R1b	R1b	0	0	0,991	(...)	0
1299	R1b	R1b	0,001	0	0,199	(...)	0,701
1300	R1a1a1b2	R1a	0	1	0	(...)	0
CLADOS			E				

Legenda: Cada linha representa um indivíduo (ID), e cada coluna representa um agrupamento (*cluster*). O software estima probabilidades de pertencimento a cada agrupamento para cada indivíduo, sendo a soma dessas probabilidades equivalente a 1. Consideraram-se como corretas as atribuições com probabilidades superiores a 0,50. A partir desse limiar, foi possível determinar qual clado corresponde a cada *cluster*.

Foi também realizada a inferência dos haplogrupos para cada haplótipo composto por 23 marcadores Y-STR obtido neste trabalho, através das ferramentas online *Haplogroup Predictor* (http://www.hprg.com/hapest5/) (Athey (2005, 2006) e *NevGen* Y-DNA Predictor (http://*NevGen*.org), de acordo com as instruções constantes nos respectivos *websites*.

5.6 Estimativa de diversidade e parâmetros forenses

Os principais parâmetros estatísticos utilizados para a validação forense do uso de marcadores Y-STR para as populações são: diversidade gênica, diversidade haplotípica, além de probabilidade de coincidência haplotípica e capacidade de discriminação.

- A **diversidade gênica** (DG) representa a probabilidade de dois indivíduos, escolhidos ao acaso na população, apresentarem alelos diferentes em um determinado loco. Sua estimativa é feita a partir das frequências alélicas estimadas para a população, correspondendo, nos STR autossômicos, à medida de heterozigosidade esperada. A DG é obtida pela seguinte expressão matemática (PURPS *et al.*, 2014):

$$DG = n\,(1-\sum p_i^2)\,/\,(n-1)$$

onde n = número de alelos e p = frequências alélicas relativas.

- A **diversidade haplotípica** (DH) se refere à probabilidade de encontrar dois haplótipos distintos na população, sendo estimada com a seguinte fórmula (EXCOFFIER; LAVAL; SCHNEIDER, 2005):

$$DH = n\,(1 - \sum p_{Hi}^2)\,/\,(n-1)$$

Onde n = número de haplótipos e p = frequências haplotípicas relativas.

- A medida estatística que avalia a probabilidade de dois indivíduos aleatoriamente amostrados compartilharem o mesmo haplótipo é denominada **probabilidade de coincidência haplotípica** (PCH), sendo calculada pela soma do quadrado das frequências haplotípicas (PURPS *et al.*, 2014). Quanto menor seu valor, maior será o peso da evidência.
- O poder ou **capacidade de discriminação** (CD) reflete a medida da eficiência do conjunto de marcadores para a distinção de indivíduos não aparentados, sendo determinado pela razão entre o número de

haplótipos distintos observados na amostra pelo número total de haplótipos obtidos (PURPS *et al.*, 2014). Quanto maior seu valor, mais apropriado o conjunto de Y-STR utilizado.

As frequências alélicas e haplotípicas para todos os 23 locos, bem como as medidas de diversidade, gênica e haplotípica, foram estimadas utilizando o programa ARLEQUIN v. 3.1 (EXCOFFIER; LAVAL; SCHNEIDER, 2005). O Poder de Discriminação foi calculado dividindo-se o número de haplótipos diferentes obtidos pelo número total de haplótipos da amostra. A Probabilidade de Coincidência foi obtida pela soma do quadrado das frequências haplotípicas (PURPS *et al.*, 2014). Os parâmetros forenses foram calculados para Afrobrasileiros, eurobrasileiros e indígenas na amostra populacional estudada. Os dois alelos do marcador DYS385a/b foram tratados como alelos individuais pois nenhum destes pode ser atribuído de forma inequívoca a um loco definido (**Figura 5**).

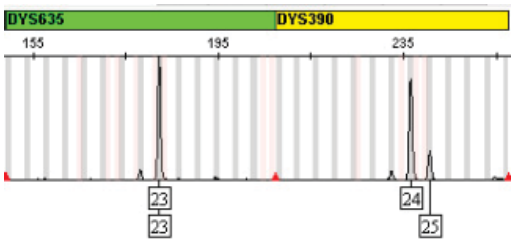
5.7 Simulações e análise de misturas genéticas compostas por indivíduos com haplótipos de Y-STR distintos

Foram produzidas misturas genéticas de pares de indivíduos, por meio de simulações computacionais de misturas dos respectivos perfis de Y-STR, para os 178 indivíduos que compõem a amostra. Para tanto foi desenvolvido um script utilizando o Pacote R para Windows versão 4.2.3 (disponível em https://www.r-project.org/) e o software auxiliar RStudio 2023.03.0+386 (R CORE TEAM., 2020). Nesta etapa, as amostras do indivíduo asiático e do indivíduo cuja procedência geopolítica e cujo haplogrupo não são conhecidos foram excluídas. Os resultados foram exportados para análise detalhada utilizando o programa Excel.

A análise de cada uma das 15.931 simulações envolvendo todos os pares possíveis a partir dos 178 indivíduos consistiu na investigação da presença de alelos diferentes entre os haplótipos dos dois indivíduos envolvidos. Para facilitar a compreensão, adaptamos as expressões "falsa heterozigose" indicando diferenças e "falsa homozigose" referindo-se a igualdades, embora estes termos

essencialmente não se apliquem ao cromossomo Y. No exemplo de eletroferograma de caso real apresentado abaixo (**Figura *8*Figura *8***), ao observarmos o loco DYS635, ambos os indivíduos contribuintes da mistura apresentam o mesmo alelo, DYS635*23, traduzido como uma falsa homozigose. Em contraste, no loco DYS390, os indivíduos possuem alelos diferentes (DYS390*24 e DYS390*25). Nesse contexto, nos referimos a "falsa heterozigose". Todas as falsas heterozigoses, ou diferenças de alelos, foram contabilizadas em cada uma das comparações.

Figura 8 - Fragmento de eletroferograma de dois locos Y-STRs, representando possibilidade de cenários em uma mistura entre dois indivíduos.



Legenda: O loco DYS365 demonstra a “falsa homozigose” e o loco DYS390 indica a “falsa heterozigose”.

Fonte: A Autora (2023).

Dada a complexidade das análises efetuadas, a **Tabela 2** ilustra como foi conduzido o processo, através da representação de 19 simulações fictícias, utilizando indivíduos e dados imaginários. As simulações entre cada par de indivíduos foram listadas e numeradas, de maneira a representar, cada qual, eventos independentes.

Na simulação fictícia 1 representada na **Tabela 2**, dois indivíduos europeus (EUR3 e EUR2, fictícios) foram comparados. A comparação foi feita entre os haplótipos de cada um deles, sendo os Y-STRs comparados um a um como explicado na **Figura *8***. Foram contabilizadas 12 falsas heterozigoses (diferenças) entre os alelos dos Y-STRs de EUR3 e de EUR2. Isto é, os indivíduos EUR3 e EUR2 compartilhavam alguns alelos de Y-STR, porém 12 Y-

STRs se mostraram diferentes entre eles. As 12 diferenças correspondem a 54,55% dos locos efetivamente amplificados para estes dois indivíduos.

Nas simulações 4 e 5 representadas (**Tabela 2**), envolvendo outros pares de indivíduos europeus imaginários, foram observadas 14 falsas heterozigoses em cada simulação. As 14 diferenças correspondem a 63,64% dos locos efetivamente amplificados para as duas simulações.

Tabela 2 – Representação exemplificativa, com indivíduos e dados fictícios, para descrição do procedimento de comparação de alelos Y-STR entre indivíduos.

Número da Simulação	Id1	Linhagem Patrilínea 1	Id2	Linhagem Patrilínea 2	22 LOCOS COMPARADOS																						Soma das Diferenças	Locos Amplificados	Porcentagem de Diferença
					Loco 1	Loco 2	Loco 3	Loco 4	Loco 5	Loco 6	Loco 7	Loco 8	Loco 9	Loco 10	Loco 11	Loco 12	Loco 13	Loco 14	Loco 15	Loco 16	Loco 17	Loco 18	Loco 19	Loco 20	Loco 21	Loco 22			
1	EUR3	European	EUR2	European	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0	0	1	1	0	1	1	0	12	22	54,55
2	EUR10	European	EUR9	European	1	1	0	1	0	1	1	1	0	1	1	-	1	0	1	0	-	1	0	1	1	-	13	19	68,42
3	AMR4	Amerindian	AFR7	African	1	1	0	1	0	1	1	1	0	0	1	1	0	0	1	0	1	1	0	1	0	1	13	22	59,09
4	EUR2	European	EUR1	European	0	1	0	1	0	1	1	1	0	0	1	1	1	1	1	0	0	1	1	1	0	1	14	22	63,64
5	EUR6	European	EUR5	European	1	1	0	1	0	1	1	1	0	0	1	1	1	1	0	0	1	1	0	1	0	1	14	22	63,64
6	EUR9	European	EUR8	European	1	1	-	1	0	1	1	1	-	0	1	1	-	1	1	0	1	1	-	1	1	1	15	18	83,33
7	AMR5	Amerindian	AFR8	African	0	1	0	1	0	1	1	1	0	-	1	1	1	1	1	1	1	1	0	1	0	1	15	21	71,43
8	AMR6	Amerindian	AMR2	Amerindian	1	1	0	1	0	1	1	1	-	0	1	1	0	1	0	0	1	1	1	1	1	1	15	21	71,43
9	EUR1	European	EUR15	European	1	1	-	1	0	1	1	1	0	1	1	1	1	1	0	1	1	0	1	0	1	0	16	21	76,19
10	EUR7	European	EUR6	European	1	1	0	1	1	1	1	1	1	0	1	1	0	1	0	0	1	1	0	1	1	1	16	22	72,73
11	AFR8	African	EUR11	European	1	1	0	1	0	1	1	1	1	1	1	1	-	1	0	0	1	1	0	1	1	1	16	21	76,19
12	AMR9	Amerindian	EUR14	European	1	1	0	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	0	16	22	72,73
13	AFR9	African	AMR1	Amerindian	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	1	1	1	1	17	22	77,27
14	AMR8	Amerindian	ERU13	European	1	1	0	1	0	1	1	1	1	0	1	1	1	0	1	0	1	1	1	1	1	1	17	22	77,27
15	EUR4	European	EUR3	European	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	1	0	1	18	22	81,82
16	EUR5	European	EUR4	European	1	1	0	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	18	22	81,82
17	AFR7	African	EUR10	European	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	0	1	18	22	81,82
18	AMR7	Amerindian	EUR12	European	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	1	1	1	18	22	81,82
19	EUR8	European	EUR7	European	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	19	22	86,36
TOTAL					17	19	3	19	6	19	19	19	9	8	19	18	10	13	10	5	17	18	6	18	12	16	-		

Legenda: Cada linha representa o resultado da simulação de comparação entre indivíduo 1 e indivíduo 2. Id1 e Id2: identificação única dos dois indivíduos. Nas colunas dos locos comparados, o número 1 representa a falsa heterozigose (presença de diferença no mesmo loco dos dois indivíduos) e o número 0 representa a falsa homozigose (ausência de diferença no mesmo loco dos dois indivíduos). Soma das diferenças: somatória do número de falsas heterozigoses em cada simulação. Locos amplificados: número de Y-STRs cuja amplificação foi bem-sucedida. A falha de amplificação por loco é identificada por "-". Porcentagem de diferença: número de diferenças observadas com relação ao número de locos amplificados. Os valores e identificações utilizados nesta tabela são fictícios.

Considerando-se eventuais falhas de amplificação observadas para alguns dos Y-STRs dos indivíduos da amostra populacional, as diferenças encontradas na comparação entre os Y-STRs dos haplótipos individuais foram ajustadas em percentual, de forma a representar a proporção do número de marcadores que efetivamente foram genotipados em ambas as amostras da mistura simulada. Assim, nas simulações fictícias 9 e 10 representadas (**Tabela 2**), envolvendo outros pares de indivíduos europeus, foram observadas 16 falsas heterozigoses para cada simulação. No entanto, como pode ser observado, na

simulação 9 houve perda de amplificação de um Y-STR, o qual não pode ser, portanto, comparado. Logo, as 16 diferenças correspondem a 72,73% dos locos efetivamente amplificados para a simulação 10, e a 76,19% dos locos efetivamente amplificados na simulação 9. Por conseguinte, podemos afirmar que a média de eventos correspondente a 16 falsas heterozigoses equivale a 74,46%.

Reunindo-se os resultados para todas as simulações entre europeus na **Tabela 2**, contabilizou-se, desta forma, as médias de eventos para todas as quantidades de falsas heterozigoses: uma ocorrência (ou evento) para 12 diferenças (54,55%), um evento para 13 diferenças (68,42%), dois eventos para 14 diferenças (média de 63,64%), um evento para 15 diferenças (83,33), dois eventos para 16 diferenças (média de 74,46%), e assim por diante.

O procedimento foi reproduzido nas simulações realizadas para todas as demais populações isoladamente (AFR, EUR e AMR) e depois misturando-se também indivíduos de cada duas populações (EUR/AFR; AFR/AMR; AMR/EUR). Dessa forma obteve-se um panorama de todas as diferenças observadas em comparações dentre e entre populações.

Os dados ainda possibilitaram, como também mostrado na **Tabela 2**, que o total das diferenças observadas por cada um dos Y-STRs pudesse ser computado. No exemplo, o loco 1 apresentou ao todo 17 diferenças no conjunto das 19 simulações, o loco 2 apresentou ao todo 19 diferenças, o loco 3 teve 3 diferenças, e assim por diante.

Cada Y-STR apresenta um número de repetições de STRs, ou seja, um alelo. O script utilizado para a comparação dos marcadores entre os indivíduos, foi desenvolvido, portanto, para comparar alelos. Por exemplo, ao comparar a região DYS19 entre dois indivíduos, o resultado pode ser a identificação de dois alelos iguais ou dois alelos diferentes (**Figura *8***). Para o marcador DYS385a|b, como já mencionado, são duas as regiões (“a” e “b”) amplificadas, não sendo possível distinguir se o fragmento (alelo) obtido se refere à região “a” ou “b”. Por convenção, e unicamente por este motivo, o alelo que indica o menor número de repetições é referido como “a”. Devido a essa particularidade de duas regiões e à incerteza sobre a qual alelo pertence a qual região específica no marcador, foi

necessário realizar uma adaptação na comparação desta região entre dois indivíduos. Essa adaptação simplificadamente considerou qualquer diferença observada entre os dois alelos genotipados neste marcador como falsa heterozigose, mesmo que um dos alelos fosse igual. Em outras palavras, se o indivíduo A tivesse DYS385*11,14, e o outro indivíduo B tivesse DYS385*11,14, o script considerou isso como falsa homozigose. Por outro lado, se o indivíduo B tivesse DYS385*12,14, consideramos que este marcador era inteiramente diferente, mesmo que o alelo 14 fosse igual entre eles.

Por este motivo, o loco DYS385 foi considerado um loco único embora represente dois alelos. Assim, o total de locos comparados foi de 22 Y-STRs, embora o kit comercial utilizado contemple 23 Y-STRs.

6. RESULTADOS E DISCUSSÃO

6.1 Caracterização da diversidade de haplótipos (Y-STR) na amostra brasileira

As diversidades haplotípicas e gênicas observadas para as regiões geopolíticas estudadas (Mato Grosso do Sul, São Paulo e Paraná) são altas, com média de 0,9999 para diversidade haplotípica e 0,6920 para diversidade gênica, quando considerado simultaneamente o conjunto das populações (**Tabela 3**). Embora o tamanho amostral possa influenciar os valores obtidos, a alta diversidade haplotípica obtida neste estudo reproduz resultados semelhantes aos encontrados em populações urbanas brasileiras (JANNUZZI *et al.*, 2020). Além disso, esses resultados são consistentes com grupos de populações urbanas de diversas origens, como africanos, asiáticos, europeus, latinos, norte-americanos, e indígenas americanos, considerando-se a genotipagem de 23 Y-STRs, como proporcionado pelo kit utilizado (PURPS *et al.*, 2014). A observação de valores mais elevados para diversidade haplotípica também está associada ao aumento no número de marcadores testados. Isso foi evidenciado pela obtenção de resultados ligeiramente menores ao analisar 17 STRs em comparação com 26 Y-STRs (JANNUZZI *et al.*, 2020). Sendo assim, entende-se que o tamanho amostral não comprometeu a análise do desempenho do kit comercial utilizado.

Tabela 3 - Diversidade haplotípica e gênica nas populações estudadas.

PARÂMETROS	POPULAÇÃO MISCIGENADA				AMERÍNDIOS DO PARANÁ
	MATO GROSSO DO SUL	SÃO PAULO	PARANÁ	TOTAL	
Número de indivíduos	124	17	24	165	8
Diversidade Haplotípica	0,9999	0,9999	0,9999	0,9999	0,9999
Diversidade Gênica Média	0,6895	0,6659	0,7046	0,6920	0,5838

Foi observado um haplótipo coincidente entre dois indivíduos do Mato Grosso do Sul, dois haplótipos coincidentes entre dois pares de indígenas no Paraná e um haplótipo coincidente entre um indivíduo de São Paulo e outro do Paraná. A alta diversidade encontrada decorre da pequena proporção de compartilhamento de haplótipos dentro das populações. Cabe esclarecer que, nos casos forenses, haplótipos coincidentes demandam a complementação de

avaliações estatísticas que incluam taxas de mutação e distribuições de frequências haplotípicas dos Y-STRs nas populações de interesse. Tais dados são disponibilizados, por exemplo, por bancos de dados como o YHRD (do inglês *Haplotype Reference Database*). Além disso, a complementação com STRs autossômicos sempre é desejável, por ser mais informativa e conclusiva. Em estudos populacionais, não é incomum o relato de haplótipos Y-STR coincidentes entre indivíduos de uma população (JANNUZZI *et al.*, 2020; PURPS *et al.*, 2014). No entanto, é crucial destacar que tal achado não é conclusivamente indicativo de parentesco, o que não constitui o foco da pesquisa. Portanto, é simplesmente relatado como uma ocorrência que influencia a diversidade haplotípica da população estudada, não requerendo que as populações sejam avaliadas separadamente.

Os valores de diversidade haplotípica obtidos para o Mato Grosso do Sul e São Paulo corroboram dados já publicados (JANNUZZI *et al.*, 2020). O Paraná apresentou diversidade haplotípica semelhante à relatada para 17 Y-STRs anteriormente estudados (ALVES, 2012).

Foi observado que o Paraná ainda apresentou a maior diversidade gênica (0,7046) dentre os três estados, em contraste com a menor diversidade gênica (0,5838) encontrada em seus indígenas. Este último resultado está em conformidade com a menor diversidade esperada para populações isoladas (PENA; SANTOS; TARAZONA-SANTOS, 2020; SALZANO; SANS, 2014). A diversidade gênica reflete a capacidade do conjunto de marcadores em distinguir indivíduos dentro de uma população. Na amostra analisada, esse valor está alinhado com a alta diversidade haplotípica obtida. No entanto, a possibilidade de viés de amostragem nos valores obtidos para o Paraná e São Paulo deve ser considerada, pois a representatividade da amostra pode influenciar significativamente os resultados. Portanto, a amostra populacional pode não refletir totalmente a diversidade genética de toda a população do Paraná e São Paulo.

6.2 Caracterização de haplogrupos (Y-SNPs) na amostra brasileira

Conforme mencionado anteriormente, as amostras analisadas no presente estudo (indivíduos da população urbana do Mato Grosso do Sul, São Paulo e Paraná, além de indígenas do Paraná), foram recentemente submetidas a uma caracterização genética baseada em marcadores uniparentais (JOERIN *et al.*, 2022). Com relação ao cromossomo Y, os haplogrupos inferidos pelos Y-SNPs pertencem a oito linhagens principais (clados), dentro dos quais foram divididos em haplogrupos de acordo com as mutações que os originaram. Dentre os haplogrupos identificados e descritos até o momento, conforme *a* ISOGG (ISOGG, 2020) os clados E, G, I, J, L, R, Q e T estão presentes na amostra (**Figura *9***) razão pela qual serão brevemente descritos de acordo com os achados de Joerin et al., (2022):

6.2.1 Clado E

Presente em 22,34% dos indivíduos da amostra populacional do presente trabalho, o clado E é considerado o mais diversificado de todos os clados do cromossomo Y. Os polimorfismos de seus marcadores definem haplogrupos presentes em alta frequência na África, em moderada quantidade no Oriente Médio e no sul da Europa, e com baixa ocorrência na Ásia (KARAFET *et al.*, 2008). Na amostra estudada, são identificados oito subgrupos distintos, sendo os indivíduos classificados como haplogrupos E1b1a1, E1a2 e E2b1 (e seus subhaplogrupos) atribuídos à origem africana, e os indivíduos classificados como E1b1b (e seus subhaplogrupos), à origem europeia.

6.2.2 Clado G

Este clado apresenta uma distribuição geográfica de menor amplitude, sendo encontrado principalmente no Oriente Médio, no Mediterrâneo e nas Montanhas do Cáucaso (KARAFET *et al.*, 2008). A população urbana brasileira que apresentou a mutação G (M201) representa 3,91% da amostra e foi atribuída como sendo de origem europeia.

6.2.3 Clado I

Representando, junto com o clado R, os dois maiores clados europeus, está virtualmente ausente de outras regiões geográficas. Os indivíduos assim

classificados foram atribuídos à origem europeia, e constituem 7,82% da amostra do presente estudo, que contempla os haplogrupos I1 (presente no norte da Europa) e I2 (mais frequente haplogrupo no leste Europeu e nos Balcãs) (KARAFET *et al.*, 2008).

6.2.4 Clado J

O clado J, na amostra do presente trabalho, contempla indivíduos que apresentam mutações distintas que permitem que sejam minimamente classificados como J2a1 e J2b2, todos atribuídos a origem europeia (8,38%). Sua distribuição mundial compreende linhagens do haplogrupo J encontradas em altas frequências no Oriente Médio, Norte da África, Europa, Ásia Central, Paquistão e Índia (KARAFET *et al.*, 2008).

6.2.5 Clado L

O clado L está presente em 0,55% da amostra, sendo aqui atribuído à origem asiática. Segundo publicações anteriores, este clado está presente, em sua maioria, no subcontinente indiano, no Oriente Médio, Asia, Norte da África e costa mediterrânea da Europa (KARAFET *et al.*, 2008; VAN OVEN *et al.*, 2014).

6.2.6 Clado R

6.2.6.1 Haplogrupo R1a

Este haplogrupo é uma ramificação do haplogrupo R1, identificado pelas mutações M173 e R1a-M420, apresentando-se mais frequente no leste europeu (KING *et al.*, 2011). No presente trabalho, está presente em 1,67% dos indivíduos.

6.2.6.2 Haplogrupo R1b

Assim como o haplogrupo R1a, este haplogrupo contempla europeus, porém mais especificamente do oeste da Europa, sendo uma ramificação do haplogrupo R1 identificado pela mutação M173 e R1b-M343 e R1b-M269 (KING *et al.*, 2011). A maioria dos europeus pertence a este haplogrupo (KARAFET *et al.*, 2008) e, no presente trabalho, está presente em 44,13% dos indivíduos. O haplótipo modal do Atlântico faz parte deste haplogrupo.

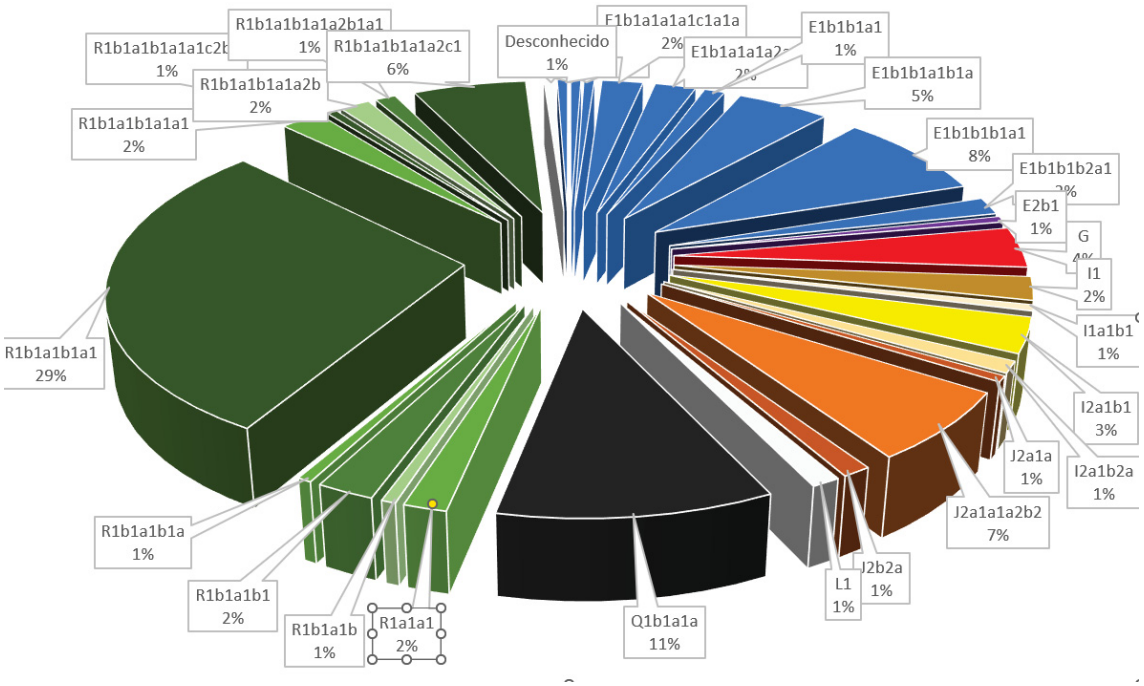
6.2.7 Clado Q

O haplogrupo Q é a principal linhagem dentre os nativos americanos, sendo também vastamente distribuído na Ásia (central e norte) (KARAFET *et al.*, 2008; VAN OVEN *et al.*, 2014). O haplogrupo Q-M3 é praticamente restrito às Américas (KARAFET *et al.*, 2008) e, no presente trabalho, representa 10,61% da população estudada, sendo encontrado em todos os indígenas da amostra.

6.2.8 Clado T

O haplogrupo T está representado por 0,55% da amostra, tendo-lhe sido atribuída a origem europeia. Publicações anteriores atribuem este haplogrupo ao Oriente Médio, África e Europa, com baixa frequência (KARAFET *et al.*, 2008).

Figura 9 – Ilustração da proporção dos haplogrupos presentes na amostra populacional brasileira urbana e indígena.



Legenda: Haplogrupos pertencentes ao mesmo clado são representados pela mesma cor, variando em tonalidades. Haplogrupos que representam um clado único são indicados por uma única cor.

Fonte: A Autora, 2023.

Em suma, nos indígenas (AMR) está presente apenas o haplogrupo (Q-M3). Na população urbana, foram identificados vários haplogrupos europeus (EUR) (clados E, G, I, J, L, R) e africanos (AFR) (clado E), conforme ilustrado na **Figura 9**. O haplogrupo R1b foi o mais frequente, seguido do haplogrupo E1b1b1b1a1.

Os haplogrupos da amostra populacional apontam para a alta frequência das sublinhagens dos haplogrupos R1b, E1b1, G-M201, I e J, nas regiões em estudo, como também já publicado anteriormente (RESQUE *et al.*, 2016).

Considerando a diversidade de haplogrupos dentro das regiões geopolíticas estudadas (Mato Grosso do Sul, São Paulo e Paraná) a composição dos haplogrupos é sumarizada na **Tabela 4**. Para as análises populacionais iniciais optou-se por não considerar os indivíduos originários da região Norte (n=2) e do Paraguai (n=1). Isso em virtude do reduzido número amostral e sua consequente baixa representatividade, bem como do desvio do foco das regiões que, por apresentarem maior número de indivíduos, foram efetivamente estudadas do presente trabalho.

A predominância de haplogrupos EUR (84% a 100%) é evidente. No estado de São Paulo, provavelmente em função do baixo número amostral (n=17), apenas indivíduos de haplogrupos EUR compuseram a amostra. No Mato Grosso do Sul observa-se como segunda maior presença o componente AMR (9,6%) e, por fim, os de origem AFR (5,6%). O segundo maior componente no Paraná é de origem AFR (12,50%). Dentre os oito AMR no Paraná, um indivíduo apresentou haplogrupo AFR.

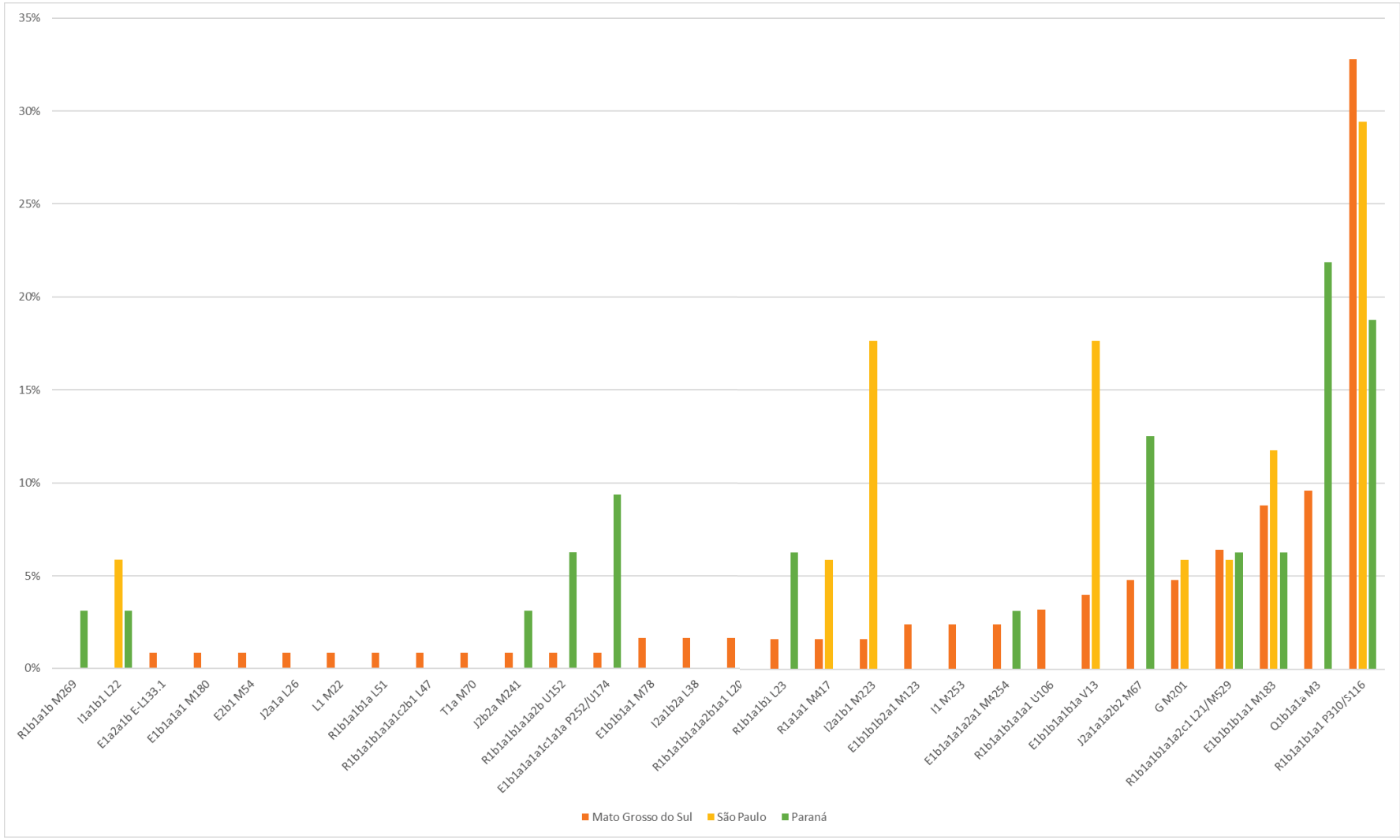
Tabela 4 - Composição dos haplogrupos Y-SNPs nos três estados avaliados

ORIGEM ANCESTRAL (Y-SNP)	POPULAÇÃO MISCIGENADA								AMERÍNDIOS	
	MATO GROSSO DO SUL		SÃO PAULO		PARANÁ		TOTAL		PARANÁ	
EUROPEIA	105	84,0%	17	100,00%	21	87,50%	143	86,14%	0	0,00%
AFRICANA	7	5,6%	0	0%	3	12,50%	10	6,02%	1	12,50%
AMERÍNDIA	12	9,6%	0	0%	0	0,00%	12	7,23%	7	87,50%
ASIÁTICA	1	0,8%	0	0%	0	0%	1	0,60%	0	0,00%
TOTAL	**125**	100%	**17**	100%	**24**	100%	166	100%	**8**	100%

Na distribuição dos haplogrupos que compõem as populações nos três estados (**Figura 10**), o haplogrupo R1b (EUR) é o predominante nos três estados, seguido pelos haplogrupos E1b1b e E1b1a. O Estado do Mato Grosso do Sul exibe ampla diversidade de haplogrupos devido à sua representatividade significativa na amostragem, bem como à sua história demográfica única, na qual a descoberta das minas atraiu migrantes de diferentes origens, incluindo portugueses, e ocasionou a realocação de escravizados do norte/nordeste do Brasil e de países vizinhos. Possui a segunda maior população AMR do país, sendo que tanto os AMR locais quanto os realocados foram essenciais para o desenvolvimento regional (JOERIN *et al.*, 2022). Em São Paulo, os haplogrupos I2a e E1b1b1 são os mais prevalentes, compondo, junto com o haplogrupo R1b, mais da metade da amostra e suportando a respectiva ancestralidade paterna EUR.

A caracterização mostrada reflete os dados demográficos da população geral e também das três regiões estudadas (IBGE, 2023). A predominância europeia e as menores proporções africana e indígena, corroboram publicações recentes (RESQUE *et al.*, 2016) e o histórico de ocupação dos estados estudados no processo da colonização brasileira (CALLEGARI-JACQUES *et al.*, 2003; JOERIN *et al.*, 2022; PENA; SANTOS; TARAZONA-SANTOS, 2020; SOUZA *et al.*, 2019).

Figura 10 - Distribuição dos haplogrupos determinados por Y-SNPs nos três estados brasileiros avaliados.



6.3 Caracterização da diversidade haplotípica (Y-STR) dentro dos haplogrupos inferidos

Dos indivíduos da amostra investigada, 15,55% apresentaram falha parcial de genotipagem Y-STR, com perda de detecção de um a seis alelos do sistema Y-23, sendo a maioria por perda de um a três marcadores (8,89% da população total). Avaliou-se que esta perda não comprometeu o número de locos viáveis para oferecer resultados conclusivos, uma vez que o painel Y23 configura um haplótipo amplo, e, portanto, não representa impacto sobre a estimativa das frequências alélicas e haplotípicas, tampouco sobre as medidas de diversidade e demais parâmetros forenses. Ressalta-se que o resultado de falha de genotipagem foi um pouco maior que o observado na população referência (9,42% da população total) do CEPH-HGDP. Porém em nenhuma das populações (HGDP e questionada) excedeu 16% dos indivíduos. Perdas de amplificação são relativamente comuns no universo forense, devido à baixa qualidade e quantidade do material genético obtido. Não é o caso de amostras referência, como das populações estudadas. Mas exemplifica que a perda de amplificação pode ocorrer e não impossibilita a realização de análises pertinentes. De toda forma, os laboratórios de genética forense preconizam estudos de validação que definem o número mínimo de marcadores com desempenhos dentro dos limiares esperados e que, portanto, permitem compor o que se considera um perfil minimamente viável, bem como quais os perfis que não são considerados analisáveis.

As frequências alélicas dos 23 Y-STR encontram-se listadas, para cada um dos grupos de ancestralidade parental estudados (**Tabela *5***). Dentre os indivíduos com haplogrupos EUR determinados pelos Y-SNPs, o alelo DYS393*13 (0,7095) foi o mais frequente; e aqueles com os haplogrupos AFR e AMR compartilharam, com valores semelhantes, a maior frequência do alelo DYS438*11 (0,9091 e 0,8947, respectivamente).

Tabela 5 - Frequência relativa dos alelos de cada marcador Y-STR na amostra estratificada de acordo com a origem biogeográfica dos haplogrupos inferidos (europeu, ameríndio e africano).

DYS19				DYS385a				DYS385b			
Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio
12	0,0068		0,0526	9	0,0068			12	0,0068		
13	0,1689		0,8421	10	0,0203			12.2	0,0068		
14	0,5405	0,0909	0,0526	11	0,4054			13	0,0878		
15	0,1419	0,2727		12	0,1216			14	0,5135		0,4211
16	0,0743	0,1818	0,0526	13	0,2027		0,1579	15	0,1622	0,0909	
17	0,0135	0,2727		14	0,1014	0,2727	0,7895	16	0,0473	0,0909	0,3158
				15	0,0743	0,2727	0,0526	17	0,0811	0,1818	0,1579
				16	0,0541	0,3636		18	0,0743	0,2727	0,0526
				17	0,0068	0,0909		19	0,0135	0,0909	
				19	0,0068			20	0,0068	0,2727	0,0526

DYS390				DYS391				DYS392			
Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio
21	0,0068	0,8182		9	0,0946			11	0,4122	0,9091	
22	0,0338	0,0909		10	0,4730	0,7273	0,6316	12	0,0473		
23	0,2365		0,1053	11	0,4257	0,2727	0,3684	13	0,4392		
24	0,5743		0,7368	12	0,0068			14	0,0608		0,6316
25	0,1419	0,0909	0,1579					15			0,3158
26	0,0068										

DYS438				DYS439				DYS448			
Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio
9	0,0541	0,0909		10	0,0878			17	0,0203		
10	0,3851		0,1053	11	0,3446	0,2727	0,2632	18	0,1014		
11	0,0473	0,9091	0,8947	12	0,4324	0,4545	0,3684	19	0,4459	0,0909	0,2632
12	0,4595			13	0,1149	0,2727	0,2105	20	0,2905		0,7368
13	0,0203			14	0,0068		0,1053	21	0,1284	0,8182	
				15			0,0526	23	0,0068		

DYS481				DYS533				DYS549			
Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio
21	0,0878			9	0,0270			10	0,0068		
22	0,4324			10	0,0405		0,0526	11	0,1689	0,3636	
23	0,1959		0,2632	11	0,2500	0,6364	0,3158	12	0,4257	0,3636	0,5263
24	0,0743	0,1818	0,5789	12	0,5405	0,2727	0,5263	13	0,3108	0,0909	0,4737
25	0,0608	0,2727	0,0526	13	0,1149	0,0909	0,0526	14	0,0743	0,0909	
26	0,0338	0,1818		14	0,0203			15	0,0068		
27	0,0676	0,0909	0,0526	15	0,0068						
28	0,0203	0,1818	0,0526								
29	0,0203	0,0909									
31	0,0068										

DYS635				YGATAH4				DYS643			
Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio
19	0,006757			10	0,0203		0,1053	9	0,094595	0,263158	
20	0,067568		0,090909	11	0,3176	0,5455	0,4737	10	0,418919	0,526316	0,090909
21	0,216216		0,545455	12	0,5135	0,2727	0,3684	11	0,135135	0,105263	0,181818
22	0,121622	0,473684	0,272727	13	0,0743	0,1818	0,0526	12	0,222973	0,105263	0,181818
23	0,506757	0,526316	0,090909	14	0,0135			13	0,047297		0,272727
24	0,081081			11.2	0,0068			14			0,181818
				11.3	0,0068						

.

DYS389I				DYS389II				DYS456			
Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio
11			0,0526	**26**	0,0068			**12**	0,0068	0,0909	
12	0,2027	0,3636	0,2632	**27**	0,0068			**13**	0,0203		
13	0,6081	0,3636	0,4211	**28**	0,1351	0,1818	0,0526	**14**	0,0608	0,0909	
14	0,1757	0,2727	0,2632	**29**	0,4459	0,1818	0,3158	**15**	0,4189	0,5455	0,4737
15	0,0135			**30**	0,2838	0,1818	0,3684	**16**	0,3378	0,1818	0,4737
				31	0,0811	0,4545	0,2632	**17**	0,0946		
				32	0,0203			**18**	0,0270		

DYS393				DYS437				DYS458			
Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio
11			0,1579	**14**	0,3378	0,8182	0,7368	**14**	0,0203		
12	0,1554	0,0909	0,2632	**15**	0,4865		0,2105	**15**	0,1216		0,2105
13	0,7095	0,3636	0,4737	**16**	0,0946	0,0909		**16**	0,2568	0,6364	0,1053
14	0,0946	0,1818	0,1053					**17**	0,2365	0,1818	0,1579
15	0,0338	0,3636						**18**	0,2432	0,1818	0,1579
								19	0,0541		0,3158
								20	0,0068		0,0526
								17.2	0,0203		
								18.2	0,0270		
								21.2	0,0068		

DYS570				DYS576			
Alelo	Europeu	Africano	Ameríndio	Alelo	Europeu	Africano	Ameríndio
12		0,052632		**14**	0,02027		0,181818
14	0,006757		0,090909	**15**	0,013514	0,105263	0,181818
15	0,013514	0,052632		**16**	0,108108		0,181818
16	0,094595	0,157895	0,090909	**17**	0,243243	0,105263	0,363636
17	0,290541	0,578947	0,272727	**18**	0,358108	0,315789	0,090909
18	0,222973	0,157895		**19**	0,195946	0,263158	
19	0,121622			**20**	0,040541	0,210526	
20	0,074324		0,454545	**21**	0,02027		
21	0,040541		0,090909				
22	0,101351						
23	0,027027						
25	0,006757						

A diversidade gênica de um marcador específico indica sua capacidade de distinguir indivíduos de uma população, sendo obtido pela fórmula GD=n(1-pi^2)/(n-1) (NEI; TAJIMA, 1981). A diversidade gênica média avalia este potencial para o conjunto de marcadores. Em outras palavras, a diversidade gênica média, calculada para uma população, informa a probabilidade de que dois de seus indivíduos, selecionados de forma aleatória, possuam alelos diferentes em cada loco analisado. A diversidade gênica estimada para cada marcador Y-STR e por haplogrupo, bem como a diversidade gênica média, considerando as três ancestralidades, estão apresentadas na **Tabela 6**.

Tabela 6 - Diversidade gênica estimada pelas frequências alélicas dos 23 marcadores Y-STR na amostra populacional brasileira estratificada por ancestralidade paterna e total.

Y-STR	Haplogrupo Europeu	Haplogrupo Africano	Haplogrupo Ameríndio	Amostra Populacional Total
DYS19	0,62	0,81	0,30	0,68
DYS385a	0,76	0,78	0,37	0,80
DYS385b	0,68	0,87	0,73	0,72
DYS389I	0,56	0,73	0,72	0,59
DYS389II	0,69	0,76	0,73	0,71
DYS390	0,60	0,35	0,44	0,63
DYS391	0,59	0,44	0,49	0,58
DYS392	0,62	0,18	0,53	0,69
DYS393	0,45	0,76	0,71	0,52
DYS437	0,62	0,35	0,43	0,63
DYS438	0,63	0,18	0,20	0,70
DYS439	0,68	0,71	0,78	0,69
DYS448	0,69	0,35	0,41	0,70
DYS456	0,70	0,71	0,58	0,68
DYS458	0,80	0,58	0,84	0,81
DYS481	0,75	0,89	0,62	0,80
DYS533	0,63	0,56	0,65	0,63
DYS570	0,83	0,76	0,64	0,82
DYS576	0,76	0,84	0,81	0,78
DYS635	0,66	0,67	0,53	0,68
DYS643	0,73	0,89	0,67	0,75
YGATAH4	0,60	0,65	0,66	0,61
DYS549	0,68	0,73	0,53	0,67
Diversidade Gênica Média	**0,67**	**0,63**	**0,58**	**0,69**

Foi evidenciada maior diversidade gênica média (0,67) no grupo EUR e intermediária para o grupo AFR (0,63), enquanto a menor foi constatada para o grupo AMR (0,58). Haplótipos que apresentam marcadores com uma maior diversidade gênica, apresentam variabilidade haplotípica mais ampla e, consequentemente, terão uma maior capacidade de discriminação. Logo, os resultados obtidos sugerem eficácia discretamente maior do conjunto de 23 Y-STRs em distinguir indivíduos na população ancestral EUR. No entanto, os valores de diversidade e as diferenças encontradas entre eles, entre os haplogrupos AFR (n=11) e EUR (n=143), pode ser o reflexo do discrepante tamanho amostral destas populações.

Nos haplogrupos europeus, a maior diversidade encontrada foi no marcador DYS570 (0,83) e a menor no DYS393 (0,45). Este comportamento se refletiu também na amostra total, provavelmente devido à influência do tamanho da população EUR em comparação com as demais. Seis marcadores revelaram diversidade comparativamente mais elevada (superior a 0,70). Não foi detectado nenhum outro marcador além do DYS393 com diversidade inferior a 0,50, sugerindo a confirmação da eficiência do painel de Y23 para discriminação de indivíduos de haplogrupos EUR e na população urbana brasileira.

O componente AFR apresentou a maior diversidade para os marcadores DYS643 e DYS481 (0,89) e a menor para os DYS392 e DYS438 (0,18), sendo esses os valores extremos de diversidade encontrados em todos os haplogrupos. Seis marcadores revelaram baixa diversidade (inferior a 0,50). No entanto, a presença de treze marcadores apresentando alta diversidade (superior a 0,70), revela que o conjunto do painel apresenta bom desempenho de informatividade para AFR. A variação nos valores de diversidade entre populações com ascendência AFR, aqui relatada, poderia ser atribuída ao fato de que os kits Y-STR, originalmente concebidos para o haplótipo mínimo em populações EUR e AMR. No entanto, tais kits, como o kit comercial utilizado neste trabalho, foram posteriormente aprimorados com a inclusão de mais marcadores Y-STR e a incorporação de marcadores de mutação rápida, proporcionando uma representatividade relatada como sendo mais precisa para essas populações (SHABALALA; GUAI; OKPEKU, 2022).

O haplogrupo AMR apresentou a maior diversidade para o marcador DYS458 (0,84) e a menor para DYS438 (0,20). Sete marcadores revelaram diversidade inferior a 0,50, e outros sete apresentaram diversidade superior a 0,70. Conforme a literatura na área, o kit comercial Y23 apresenta bom desempenho para populações isoladas (SHABALALA; GUAI; OKPEKU, 2022).

O marcador DYS576 apresentou simultaneamente diversidade elevada (0,76 a 0,84) e com pouca variação quando comparados os três componentes ancestrais EUR, AFR e AMR. O marcador DYS19 apresentou diferenças nos valores de diversidades entre os haplogrupos EUR, AFR e AMR (0,62, 0,81 e 0,30 respectivamente). Estes valores destacam notável disparidade entre os

dados para cada haplogrupo, quando considerado apenas este marcador isoladamente. Os marcadores DYS385a, DYS390, DYS393 e DYS448 apresentaram variações nas similaridades entre os haplogrupos. Para um marcador, EUR e AFR mostraram semelhanças, enquanto AMR variou. Em outros casos, AFR e AMR foram semelhantes, enquanto EUR variou. Além disso, houve situações em que AMR e EUR eram semelhantes, mas AFR variava. Essa variação destaca a diversidade nos padrões genéticos para cada marcador.

Considerando-se, entretanto, a prática forense, na qual habitualmente não se dispõe de informação de ancestralidade paterna para vestígios analisados, convém analisar os dados obtidos de diversidade genética sob a perspectiva da população total que efetivamente representa a miscigenação brasileira. A diversidade da amostra total (0,69) se mostrou maior que se consideradas as populações de ancestralidades isoladas. Além disso, foram observados dez marcadores com expressiva diversidade (superior a 0,70) e somente um (DYS393) com diversidade mais baixa (0,52), indicando que a unificação das populações tende a aumentar a diversidade dos marcadores nela contidos.

Observou-se variações de diversidade gênica conforme o Y-STR, especificamente entre os grupos populacionais. Observando-se a diferença entre os valores máximos e mínimos de diversidade gênica para os marcadores Y-STR, a maior amplitude foi observada no grupo africano (0,71), variando de 0,18 (locos DYS392 e DYS438) a 0,89 (locos DYS643 e DYS481) e justificando a obtenção de haplótipos únicos nesse conjunto de haplogrupos, embora o reduzido número de indivíduos possa ter contribuído para esse resultado. No haplogrupo AMR, a amplitude (0,64) variou de 0,20 (loco DYS438, assim como nos haplogrupos AFR) a 0,84 (loco DYS481), possivelmente justificando a ocorrência de dois pares de haplótipos idênticos. O haplogrupo EUR apresentou a menor amplitude (0,38), variando de 0,45 (loco DYS393) a 0,83 (loco DYS570).

Uma vez que valores mais elevados de diversidade gênica são mais interessantes para promover melhor capacidade discriminativa, foram selecionados os Y-STRs que apresentassem os maiores valores detectados para cada uma das ancestralidades estudadas na população miscigenada

brasileira. Isto é, dentre os eurobrasileiros, selecionamos os marcadores com diversidade gênica superior a 0,70. O mesmo se fez para a população afro-brasileira e indígena brasileira. Dessa forma foi possível selecionar 17 marcadores que apresentaram desempenhos elevados. Assim, supõem-se que o conjunto minimamente constituído por estes 17 marcadores, DYS390, DYS392, DYS437, DYS438, DYS448 (identificados com os mais informativos para o haplogrupo EUR), DY385a, DYS635, DYS393, DYS456, DYS481, DYS576, DYS643 e DYS549 (identificados para o haplogrupo AFR) e DYS439, DYS458, DYS533 e YGATAH4 (identificados para o haplogrupo AMR) demonstre, em teoria, potencial discriminatório eficiente para a população miscigenada brasileira. O ranqueamento dos marcadores mais e menos informativos foi semelhante ao relatado em publicações anteriores (BALLANTYNE *et al.*, 2012; NASCIMENTO *et al.*, 2014; PURPS *et al.*, 2014). É interessante registrar que parte dos marcadores selecionados (DYS576, DYS458, DYS385a/b e DYS570) são classificados como marcadores de mutação rápida (taxas mutacionais inferiores a 10^{-2}) e apresentam, portanto, maior poder de diferenciação em linhagens masculinas (BALLANTYNE *et al.*, 2012; NASCIMENTO *et al.*, 2014).

6.4 Estimativa dos parâmetros forenses

Com relação à distribuição dos haplótipos de Y-STRs, dentro dos clados descritos, 178 haplótipos foram gerados (Tabela 6), sendo 174 distintos e quatro compartilhados entre indivíduos (dois pares de indivíduos com haplogrupos EUR e dois pares com haplogrupos AMR). Os indivíduos dos haplogrupos AFR produziram 11 haplótipos distintos, sem registro de compartilhamento entre eles (**APÊNDICE 1, APÊNDICE 2 e APÊNDICE 3**).

A estimativa dos parâmetros forenses encontra-se representada na **Tabela 7**. Observa-se diversidade haplotípica elevada, registrando valores superiores a 99%, tanto nos três grupos populacionais quanto na população unificada, em conformidade com resultados publicados para outras amostras da população brasileira (JANNUZZI *et al.*, 2020). O poder de discriminação de cada grupo apresentou valores elevados para os grupos EUR (98,65%) e AFR (100%),

sendo este último devido à ausência de haplótipos repetidos dentro desse grupo. O grupo AMR apresenta igualmente um número reduzido de indivíduos, porém menor diversidade gênica que o grupo parental AFR, como esperado, tendo ainda apresentado dois haplótipos iguais, o que contribuiu para reduzir o valor deste parâmetro neste grupo (89,47%). Foi observada baixa probabilidade de coincidência para os três grupos populacionais.

Assim, através da avaliação dos parâmetros forenses, que indicam alta diversidade gênica para os marcadores testados, alta diversidade haplotípica, alta capacidade de discriminação e baixa probabilidade de coincidência na amostra populacional examinada, confirmou-se a eficácia do painel de 23 marcadores avaliado nas populações brasileiras urbana e indígena estudadas. O resultado concorda com publicações anteriores, tanto para populações semelhantes (NASCIMENTO *et al.*, 2014) como para amostras populacionais mais abrangentes (PURPS *et al.*, 2014).

Tabela 7 - Parâmetros forenses obtidos pelo painel de 23 marcadores Y-STR para os três haplogrupos do presente estudo.

PARÂMETROS	TODAS	HAPLOGRUPOS EUROPEUS	HAPLOGRUPOS AFRICANOS	HAPLOGRUPOS AMERÍNDIOS
Número de Indivíduos	178	148	11	19
Total de Haplótipos	178	148	11	19
Haplotipos distintos	174	146	11	17
Diversidade Haplotípica	0,9991	0,9999	1,0000	0,9948
Probabilidade de Coincidência	6,60E-03	6,94E-03	9,09E-02	6,37E-02
Poder de Discriminação	0,9775	0,9865	1,0000	0,8947

6.5 Inferência da Origem Biogeográfica de Haplogrupos das Populações Urbana e Indígena Brasileiras

O custo e o tempo requeridos, além da complexidade envolvida para a genotipagem de um conjunto mínimo de Y-SNPS com a finalidade de obter a classificação de um indivíduo em um haplogrupo configuram um substancioso obstáculo para os pesquisadores e mais ainda para peritos forenses em suas rotinas. A predição do haplogrupo através dos Y-STRs, assim como estudos de fenotipagem forense, tem se estabelecido como uma alternativa promissora de

pesquisa (JANNUZZI *et al.*, 2020; KAYSER *et al.*, 2023; MUZZIO *et al.*, 2011; PETREJČÍKOVÁ *et al.*, 2014; SCHLECHT *et al.*, 2008; WANG *et al.*, 2015), ainda que a primeira opção (Y-SNPs) seja mais precisa e confiável (JANNUZZI *et al.*, 2020; MUZZIO *et al.*, 2011). No entanto, tais linhas de investigação ainda não rotineiramente são aplicadas no Brasil

Embora existam diversas opções de programas para análises populacionais, todos apresentam vantagens e desvantagens, abrangendo desde a agilidade e velocidade de operação, a capacidade de trabalhar com volumes variáveis de dados e com marcadores específicos, bem como de contemplar particularidades das populações, além de apresentar variações no desempenho observado (PORRAS-HURTADO *et al.*, 2013). Variações nos algoritmos próprios de cada programa também exercem influência em seus resultados (ATHEY, 2005; EMMEROVA *et al.*, 2017; JANNUZZI *et al.*, 2020; PETREJČÍKOVÁ *et al.*, 2014). De fato, erros têm sido reportados para programas disponíveis para a inferência de haplogrupo de pertencimento a partir de marcadores Y-STR em diversas populações (EMMEROVA *et al.*, 2017; JANNUZZI *et al.*, 2020; MUZZIO *et al.*, 2011; PETREJČÍKOVÁ *et al.*, 2014). Considerando-se populações miscigenadas como a brasileira, torna-se interessante buscar alternativas de programas que possam oferecer ou complementar inferências de haplogrupo.

O programa *STRUCTURE* apresenta uma flexibilidade ímpar para se adaptar a diferentes demandas como marcadores distintos (PORRAS-HURTADO *et al.*, 2013), o que suscitou sua escolha para avaliação de sua aplicabilidade para predição de haplogrupos utilizando Y-STRs.

Foram então simultaneamente submetidos à análise pelo programa 573 indivíduos das populações do HGDP-CEPH e 180 indivíduos da população brasileira. Para as populações do HGDP, a análise realizada foi na modalidade supervisionada, sendo as amostras rotuladas de acordo com seu haplogrupo conforme informado nas respectivas publicações (BERGSTROM *et al.*, 2020; HALLAST *et al.*, 2021). A população miscigenada brasileira foi analisada de maneira não supervisionada, de forma a permitir que o algoritmo do programa efetuasse a análise e distribuição dos dados, atribuindo os indivíduos aos

agrupamentos (ou, no inglês, *clusters*) formados. Uma vez que a análise contemplou populações cuja classificação em haplogrupos já era conhecida previamente, a adequação da inferência foi feita por comparação direta. Isto é, para a classificação efetuada pela ferramenta poder ser contabilizada como correta, o clado ou conjunto de clados da inferência efetuada pelo *software* deveria incluir o haplogrupo já conhecido para aquele indivíduo.

Ressalta-se que as atribuições realizadas pelo *STRUCTURE* contemplam, em geral, somente as linhagens principais (clados), motivo pelo qual estes são mencionados nas análises envolvendo essa ferramenta.

Nas análises de inferência de haplogrupos foram reintroduzidos na amostra populacional estudada os indivíduos de origem asiática e de origem desconhecida que foram desconsiderados em análises anteriores, pois incrementam a oportunidade de testar a eficácia do programa *STRUCTURE* frente a outros softwares preditores.

6.5.1 Escolha do melhor K

Na análise de agrupamentos efetuada pela ferramenta *STRUCTURE*, o número mais provável de *clusters* em que a população questionada foi dividida, foi verificado pela estimativa do ΔK, para K variando de 14 até 30 (Evanno et al. (2005). Estes valores foram selecionados por serem aqueles a partir dos quais o K demonstrava-se mais bem organizado. O maior incremento de probabilidades posteriores [LnP(D)] foi obtido para K=20 (**Figura *11***), no qual as probabilidades individuais de agrupamento permitiriam, de maneira geral, uma separação clara dos haplótipos questionados em 20 agrupamentos genéticos, os quais se mostraram correspondentes aos 20 clados (**Figura *12***).

Figura 11 - Comparação da eficiência [LnP(D)] para cada K no programa *STRUCTURE*.

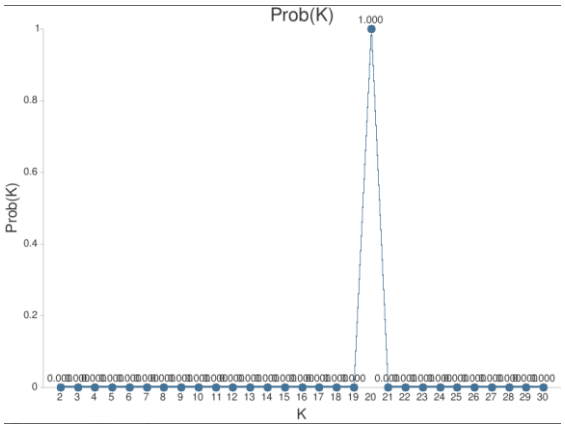
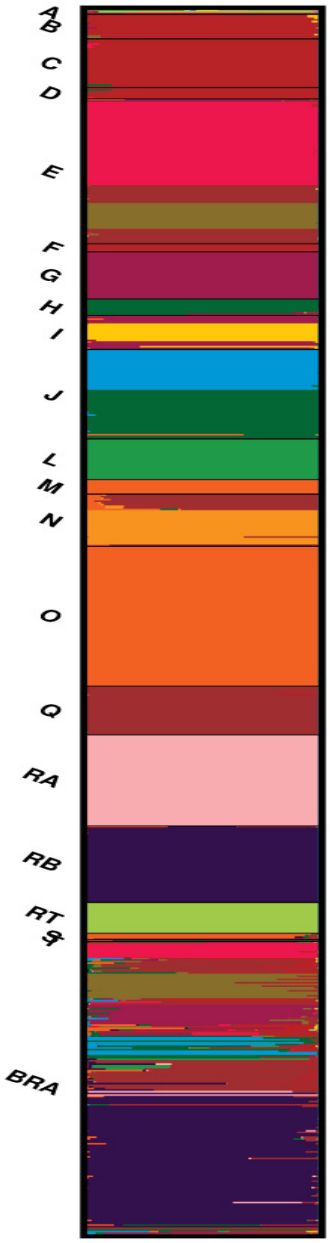


Figura 12 – Representação da composição genética de Y-STRs da amostra referência (CECH-HGDP) e da amostra brasileira questionada urbana e indígena (BRA), obtida utilizando o *STRUCTURE*, com K=20.



Legenda: Análise de agrupamentos obtidos com o programa *STRUCTURE* dos 573 indivíduos do painel de referência (análise supervisionada) e dos 180 indivíduos da população brasileira urbana e indígena (BRA) (análise não supervisionada). As letras A, B, C, D, E, F, G, H, I, J, L, M, N, O, Q, RA, RB, RT, S e T representam os clados do cromossomo Y e seus haplogrupos na população referência. As letras RA, RB e RT agrupam três categorias contemplando o clado R (R1a, R1b e R2, respectivamente, e seus subhaplogrupos) do cromossomo Y. Os vinte grupos genéticos identificados na população referência estão representados em cores distintas representando as populações pertencentes a cada haplogrupo. Os haplótipos identificados na população brasileira receberam as cores correspondentes aos haplogrupos coincidentes da população referência. Indivíduos são representados como linhas verticais finas, divididas em segmentos correspondentes à sua associação aos agrupamentos genéticos indicados pelas cores.

Fonte: A Autora (2023).

Foram estipuladas três possibilidades de interpretação para os resultados gerados para cada indivíduo: (1) atribuições de indivíduos a um clado com coeficiente acima de 50%, desde que concordassem com o clado conhecido pela análise de Y-SNPs, foram consideradas acertos; (2) os erros foram definidos pelas atribuições com coeficiente acima de 50%, porém inconsistentes com os dados de Y-SNPs conhecidos; (3) casos em que indivíduos não foram atribuídos a nenhum clado com coeficiente acima de 50% foram considerados incertezas de atribuição. Incertezas de atribuição não foram interpretadas como erros por não resultarem na alocação efetiva do indivíduo a um clado específico.

Observou-se ainda que um único *cluster* determinado pelo *STRUCTURE* pode contemplar um ou vários clados, que podem ser interpretados como contendo similaridades nos Y-STRs que compõem os haplótipos. Por outro lado, os casos em que isso foi observado sugerem a existência de uma ancestralidade compartilhada. De qualquer modo, em geral a atribuição aos clados foi coerente. Frente a isso, podemos entender que, para fins de definição de haplogrupos, os Y-STR seriam informativos. Além disso, alguns *clusters* podem se apresentar sem atribuição nenhuma de clado, tendo sido, portanto, considerados “vazios”. A existência de clusters vazios indicaria de que o número de agrupamentos excedeu o número de conjuntos em que os indivíduos podem ser organizados pelo algoritmo. De fato, à medida em que o número de K foi mais elevado, maior foi a quantidade de clusters vazios.

Para a melhor corrida selecionada para K=20, foram registrados, a partir desta metodologia, um total de 3,91% de erros de atribuição, o que incitou que fosse realizada a investigação do desempenho para os demais K (14 a 30), buscando confirmar que K=20 era efetivamente a escolha mais apropriada (**Tabela 8**). Foram verificadas as taxas de erro para a melhor corrida de cada K (variando de 14 a 30), as quais oscilaram entre 2,79% e 9,50%. Em um primeiro processo de filtragem, foram desconsiderados aqueles K que apresentaram taxa de erro acima da média de todos os K (isto é, 4,29%, que inclui os K=14, 15, 16, 18, 25 e 30). O K=20 foi mantido em virtude de ter sido sugerido pelo programa como aquele que se encontrava melhor organizado, como também pode ser visualizado na **Figura *12***.

Tabela 8 - Erros de atribuição e percentual de clusters vazios na população questionada apresentados pelo STRUCTURE para K variando de 14 a 30 na melhor corrida para cada K.

Parâmetro	Melhor corrida para cada K																	
	K=14	K=15	K=16	K=17	K=18	K=19	K=20	K=21	K=22	K=23	K=24	K=25	K=26	K=27	K=28	K=29	K=30	MÉDIA
% Erros de Atribuição	6,70	4,47	5,03	3,91	4,47	3,91	4,47	3,35	3,35	2,79	3,35	4,47	2,79	3,91	3,91	3,35	9,50	4,34
% Clusters Vazios	14,29	20	18,75	17,65	22,22	26,32	30	33,33	36,36	34,78	37,50	44	68,75	70,59	72,22	78,95	53,33	39,94

Em seguida, foram avaliadas as 15 repetições dentro de cada K com taxa de erros abaixo da média (17, 19, 20, 21, 22, 23, 24, 26, 27, 28 e 29). As taxas de erro na população questionada mostraram-se bastante semelhantes entre si (média de 4,38%), excetuando-se os erros contabilizados para K=17 (7,78%) (**Tabela 9**).

Tabela 9 - Média de erros de atribuição (obtidos pela análise do conjunto de 15 corridas para cada K) utilizados como critérios de escolha de K=20 dentre os demais K remanescentes após a primeira seleção.

Parâmetro	População	K=17	K=19	K=20	K=21	K=22	K=23	K=24	K=26	K=27	K=28	K=29	MÉDIA
Média de Erros de Atribuição (%)	Questionada	7,78	4,62	4,20	4,13	3,69	4,10	4,80	3,95	3,80	3,47	3,69	4,38
	Referência	0,76	0,76	0,69	1,03	0,79	0,95	1,48	1,05	1,10	1,26	0,71	0,96
Clusters "vazios" (%)		27,84	36,49	37,00	39,37	43,03	42,90	43,06	41,21	43,46	41,57	54,68	40,96

Em virtude da homogeneidade generalizada, isto é, observada para vários K, da taxa de erros de atribuição na população analisada, optou-se por utilizar um critério adicional que envolveu avaliar os erros de atribuição na população referência. Nesta população, também foram consideradas como erros aquelas atribuições que, mesmo atingindo coeficiente de probabilidade de 50%, demonstraram incoerências com os haplogrupos determinados pelos Y-SNPs. Isso incluiu situações como clusters com número reduzido de indivíduos (por exemplo, clado A, com três indivíduos; o clado F, com cinco; ou o clado T, com apenas um). Tais situações admitem que a falta de representatividade de

populações de referência podem comprometer a inferência realizada. Também foram consideradas potencialmente incoerentes as atribuições que não se alinhavam sistematicamente às observadas em análises correlatas. Por exemplo, a maioria das atribuições de indivíduos do clado Q foi encontrada em *clusters* relacionados ao mesmo clado Q. No entanto, em alguns K, indivíduos desse clado Q foram atribuídos a *clusters* de outros clados, o que foi interpretado como um alerta de erro. Uma terceira situação também considerada incoerente envolveu a classificação solitária de um a dois indivíduos de um mesmo clado em um cluster que abrangia diversos indivíduos, porém de outro(s) clado(s). Por exemplo, em um cluster com a presença predominante de indivíduos dos clados M-O-S-T, identificou-se a inclusão de um indivíduo do clado J. Esse tipo de discrepância ficou particularmente evidente após a avaliação de todos os K, uma vez que puderam ser identificados padrões na composição de alguns dos *clusters* constituídos por diversos clados, como aqueles repetidamente constituídos pelos clados M-O-S-T ou E-N-Q, por exemplo.

A análise do desempenho dos K restantes em relação à classificação dos clados da população referência permitiu constatar o menor valor de erros de atribuição para K=20 (0,69%) (**Tabela 9**). Considerando que a análise de K=20 revelou uma taxa média de erro na população questionada (4,20%) inferior à média de erro do conjunto dos demais K (4,38%), além de apresentar a menor taxa de erro na população de referência (0,52%), bem como um número relativamente baixo de clusters definidos como “vazios” na comparação com o restante, esse valor de K pode ser confirmado como o mais apropriado para a identificação e agrupamento dos clados da população em estudo.

6.5.2 Avaliação das atribuições de haplogrupos pelo *STRUCTURE*

Durante a análise das atribuições de haplogrupos pelo *STRUCTURE* (K=20) observou-se que os *clusters* podem abranger um único clado, sendo, entretanto, frequente que agrupem simultaneamente múltiplos clados. Por exemplo, a classificação de um indivíduo pertencente ao clado E poderia ser em um *cluster* composto exclusivamente pelo clado E ou, alternativamente, em um *cluster* inesperadamente composto pelo conjunto de clados E-Q-N, sugerindo

um possível padrão de compartilhamento de alelos entre os haplótipos de tais clados. Em situações em que um agrupamento contempla mais de um clado, contudo, não é possível determinar com precisão a qual clado pertencem os indivíduos nele alocados. Tal situação pode ser interpretada como uma limitação da ferramenta. Por outro lado, a inferência de clados mistos pode ser interpretada como resultante de divergência recente entre eles (PORRAS-HURTADO *et al.*, 2013). De fato, observou-se que os clados agrupados em um *cluster* frequentemente indicam uma possível relação evolutiva mais recente entre eles, como por exemplo os conjuntos E-G-I, E-N-Q, B-C-D-E, B-C-D-F e M-O-S-T. Além disso, a formação de *clusters* mistos pode indicar a ocorrência de convergência na população, que significa a possibilidade de encontrar os mesmos alelos e, portanto, haplótipos iguais, ou semelhantes, em haplogrupos distintos (JANNUZZI *et al.*, 2020; WANG *et al.*, 2015). A convergência se deve às altas taxas de mutação dos STRs em geral, bem como à profundidade temporal das ramificações dos haplogrupos (WANG *et al.*, 2015). Ademais, as os mecanismos de mutação dos Y-STRs de diferentes linhagens de Y-SNPs podem ser diferentes entre si (XU *et al.*, 2015). Vale ressaltar que a ocorrência de convergência não reflete a qualidade dos softwares de predição de haplogrupo, como STRUCTURE, HAPEST e NevGen, embora comprometa a acurácia da predição (WANG *et al.*, 2015).

Conforme a programação efetuada no *software STRUCTURE*, a análise (com K=20) de cada um dos indivíduos da população em estudo foi repetida até completar 15 corridas, gerando uma oportunidade de inferência de clado em cada repetição. Dessa maneira, o total de 15 inferências de clados foram obtidas para cada indivíduo. Considerando toda a amostra populacional, portanto, as análises totalizaram 2685 inferências, nas quais foram detectados 113 erros (4,20%) e 140 incertezas de atribuição (5,21%). Os erros foram distribuídos em um grupo composto por apenas 22 indivíduos, os quais foram triados para análises mais detalhadas descritas adiante. As incertezas de atribuição se distribuíram de maneira aleatória ao longo de todo o conjunto de inferências realizadas.

Embora ofereça o resultado de agrupamento somente em linhagens principais, contudo, o *STRUCTURE* também demonstrou habilidade para

separar indivíduos que pertencem a algumas linhagens secundárias, embora não nominalmente, como observado nos resultados envolvendo os indivíduos de diferentes haplogrupos do Clado E. Estes indivíduos, na amostra populacional testada, pertenciam a três haplogrupos distintos, E1b1a1a1 (origem parental africana), E1b1b1b1 e E1b1b1b2 (origem parental europeia), tendo cada grupo sido alocado em um *cluster* separadamente. No caso do clado E, essa habilidade do *STRUCTURE* ficou, portanto, bastante evidente. Em outro exemplo, como na alocação de indivíduos do clado J, que contemplava dois haplogrupos distintos (J2a1a e J2b2a), a capacidade de alocar haplogrupos foi mais sutil. Nesse caso, os indivíduos do haplogrupo J2a1a foram distribuídos em dois *clusters* distintos. Porém, a título de curiosidade, verificou-se que, em todas as 15 corridas, o agrupamento seguiu consistentemente o mesmo padrão de divisão entre os grupos destes indivíduos. Isto é, os grupos constituídos pelos mesmos indivíduos mantiveram distribuição igual em todas as corridas. Esse fato revela uma característica de constância no resultado oferecido pelo *STRUCTURE* e que pode indicar fortemente que um clado está constituído por indivíduos de diferentes haplogrupos derivados. Da mesma forma, indivíduos dos haplogrupos R1a e R1b do clado R (e seus subhaplogrupos) foram sistematicamente separados em agrupamentos distintos.

Foram identificados oito indivíduos erroneamente alocados em haplogrupos, quando analisada a melhor corrida de K=20, o que corresponde a 4,47% da população questionada e configura a taxa de erros do *STRUCTURE*. Essas alocações errôneas afetaram quatro indivíduos pertencentes ao clado E (com 4 erros), e outros quatro indivíduos pertencentes, cada qual, a um dos clados Q, I e R (R1a e R1b). Foram ainda identificados 10 indivíduos classificados como incertezas de atribuição (5,58%).

6.5.3 Comparação do desempenho do *STRUCTURE* com o de outros programas

Como parte do objetivo desta investigação consiste em avaliar o desempenho do programa *STRUCTURE* na alocação de indivíduos a haplogrupos, a população questionada foi também submetida à predição dos

haplogrupos através de haplótipos Y-STR utilizando, paralelamente, dois *softwares* de livre acesso, *Whit Athey's Haplogroup Predictor* (*HAPEST)* (ATHEY, 2005, 2006) e *Haplogroup Predictor NevGen* (CETKOVIC GENTULA; NEVSKI, 2015). Ambos os *softwares* foram escolhidos com base em publicações anteriores que atestaram sua eficácia para classificação de haplogrupos através de haplótipos Y-STR, ainda que de maneira menos eficiente da apresentada pelo método padrão ouro (Y-SNPs) (EMMEROVA *et al.*, 2017; JANNUZZI *et al.*, 2020; PETREJČÍKOVÁ *et al.*, 2014). Os resultados obtidos com o uso das três ferramentas foram comparados, em termos de taxas de acertos, erros e ainda de incertezas de atribuição, tomando como referência as definições realizadas com base na genotipagem de Y-SNPs (**Tabela 10**).

Foram aplicados os mesmos critérios anteriormente definidos para o *STRUCTURE* para detecção e contagem dos acertos, erros e incertezas de atribuição nos três programas. No entanto, a interpretação dos resultados precisou de dois ajustes.

Em primeiro lugar, considerando que existem 15 corridas no *STRUCTURE* para K=20, optou-se por utilizar somente os resultados informados por uma única corrida, escolhendo-se para tanto a indicada como mais provável. Essa escolha se baseia no fato de que a maior parte das corridas apresenta concordância entre si. No entanto, alguns casos podem apresentar discrepâncias, as quais, por sua vez, representam os erros e as incertezas de atribuição, os quais foram adequadamente considerados.

Em segundo lugar, enquanto o *STRUCTURE* se restringe a uma classificação em clados, ambos os softwares, *HAPEST* e *NevGen,* oferecem a possibilidade de classificação em linhagens secundárias (haplogrupos ou mesmo subhaplogrupos), um detalhamento maior do que se considerarmos somente as linhagens principais. Embora essa possa ser considerada uma vantagem dos dois programas sobre o *STRUCTURE*, também implicou em que, sobre eles, fosse aplicado maior rigor no rastreio de erros e acertos. Assim, para os resultados de *HAPEST* e *NevGen*, ainda que o clado estivesse corretamente inferido, qualquer discrepância nos níveis mais ramificados da filogenia foi contabilizada como erro (**Tabela 10**). Ressalta-se que a avaliação destes

programas por outros autores também condicionou a interpretação ao nível mais restrito dos haplogrupos e subhaplogrupos (EMMEROVA *et al*., 2017). Por exemplo, se o haplogrupo determinado pelos Y-SNPs sob investigação fosse J1a2a, o *STRUCTURE* estaria correto se tivesse feito a alocação ao clado J. Entretanto, para *HAPEST* e *NevGen*, considerou-se corretas apenas as predições exatas se, ao menos, contemplassem uma ramificação precursora próxima da hierarquia, como, no caso do exemplo, J1a. O critério do detalhamento foi utilizado com o único objetivo de comparar o desempenho dos softwares testados. Contudo, para serem efetivamente úteis no contexto forense, o nível de clado seria suficiente e poderia ser utilizado desta forma em qualquer um dos softwares preditivos.

Tabela 10 - Resultados das inferências nos clados realizadas pelas três ferramentas para a população questionada a partir de haplótipos de Y-STRs.

HAPLOGRUPO Y-SNP	N	STRUCTURE			HAPEST			NEVGEN		
		CORRETOS (%)	INCORRETOS (%)	INCERTEZAS (%)	CORRETOS (%)	INCORRETOS (%)	INCERTEZAS (%)	CORRETOS (%)	INCORRETOS (%)	INCERTEZAS (%)
E	40	85,00	10,00	5,00	92,50	7,50	0,00	90,00	2,50	7,50
G	7	100,00	0,00	0,00	100,00	0,00	0,00	71,43	0,00	28,57
I	14	42,86	7,14	50,00	57,14	42,86	0,00	57,14	14,29	28,57
J	15	93,33	0,00	6,67	13,33	86,67	6,67	40,00	13,33	46,67
L	1	100,00	0,00	0,00	100,00	0,00	0,00	100,00	0,00	0,00
Q	19	94,74	5,26	0,00	100,00	0,00	0,00	52,63	0,00	47,37
RA	3	66,67	33,33	0,00	66,67	33,33	0,00	66,67	33,33	33,33
RB	79	98,73	1,27	0,00	97,47	0,00	1,27	97,47	0,00	1,27
T	1	100,00	0,00	0,00	0,00	100,00	0,00	100,00	0,00	0,00

Legenda: Clado original segundo (Joerin et al. (2022). RA: representa o haplogrupo R1a e seus subhaplogrupos. RB: representa o haplogrupo R1b e seus subhaplogrupos.

Todos os três programas de predição apresentaram erros e incertezas de atribuição. Sugere-se que o tamanho amostral destes haplogrupos impacte no número de erros e acertos. Entretanto, é possível observar que os clados I, J e Q indicam maiores discrepâncias entre os programas, com as maiores taxas de erros e de incertezas.

No *STRUCTURE*, as maiores taxas de erros (7,14%) foram na inferência do clado I, que ainda mostrou quantidade elevada de incertezas de atribuição (50%), seguido pelo clado Q (5,26% de taxa de erros). Para o programa *HAPEST*, a maior taxa de erros (86,67%) e de incertezas (6,67%) foi para o clado J, seguida do clado I (42,86% de erros). O programa *NevGen* resultou em taxas de incerteza de medição mais elevadas e distribuídas em vários clados (G, I, J, Q e RA), justificadas pelo próprio *software* como devidas a alguma incompatibilidade com a versão vigente do preditor ou devido a eventual baixa

representatividade haplotípica no sistema (ATHEY, 2006). No entanto, clados constituídos por poucos indivíduos, como L, T, G e RA, não ofereceram dados efetivamente representativos da eficiência de cada programa (**Tabela 10**).

A partir dos resultados obtidos, foi verificado o desempenho dos três programas em possibilitar a inferência de ancestralidade na população brasileira analisada (**Tabela 11**). Embora os três tenham apresentado um bom desempenho, *STRUCTURE* e *HAPEST* se destacaram pelos maiores índices de acertos, registrando taxas de 89,94% e 85,47%, respectivamente. *NevGen* e *STRUCTURE* apresentaram as menores taxas de erros (3,35% e 4,47%, respectivamente). Vale ressaltar que as taxas de incertezas, embora não sejam consideradas erros, desempenham um papel informativo relevante ao indicar a falha do programa em oferecer resultados definitivos. Neste quesito, *NevGen* obteve o maior valor (15,08%). Ressalta-se que, comparando-se erros e incertezas, as últimas são mais desejáveis que os erros em si, por não induzirem a interpretações incorretas.

Tabela 11 - Desempenho de cada programa testado na realização de inferências de ancestralidade paterna para a população questionada a partir de haplótipos compostos por Y-STRs.

MA	PREDIÇÕES CORRETAS (%)	PREDIÇÕES INCORRETAS (%)	INCERTEZAS (%)
STRUCTURE	89,94	4,47	5,59
HAPEST	85,47	13,41	1,12
NEVGEN	81,56	3,35	15,08

Segundo Muzzio et al (2011), os *software*s *HAPEST* e *NevGen* são programas que podem ser denominados preditores, pois utilizam as informações genéticas, como os padrões de repetições de Y-STRs correlacionados a haplogrupos específicos, para então prever o haplogrupo do indivíduo.

Por outro lado, existem programas que podem ser denominados como “classificadores”, pois agrupam os indivíduos com base em padrões

semelhantes em seus dados genéticos, muitas vezes coincidindo com grupos de pessoas com afinidades linguísticas ou geográficas, mas não preveem necessariamente os haplogrupos (MUZZIO *et al.*, 2011; PORRAS-HURTADO *et al.*, 2013; SCHLECHT *et al.*, 2008). Entende-se que o programa *STRUCTURE* possa ser compreendido como um classificador. No caso dos Y-STRs, os indivíduos são classificados em grupos genéticos semelhantes, ou *clusters*, com base em similaridades genéticas identificadas, por exemplo, em uma população predefinida que fornece um conjunto de padrões genéticos conhecidos. Estas similaridades genéticas envolveriam, por exemplo, as frequências alélicas dos Y-STRs em cada população (PORRAS-HURTADO *et al.*, 2013). O treinamento com uma população de referência instrui o programa a identificar e agrupar padrões genéticos compartilhados, agrupando-os conforme o algoritmo apreendeu a partir da população referência. O programa *STRUCTURE* requer portanto que lhe sejam fornecidos previamente modelos de haplótipos com haplogrupos conhecidos para que possa realizar a classificação de uma população questionada (MUZZIO *et al.*, 2011), o que justifica a necessidade de que populações de referência sejam suficientemente representativas (PORRAS-HURTADO *et al.*, 2013; ROSENBERG *et al.*, 2005). No presente estudo, foram utilizados dados do painel HGDP-CEPH, vastamente empregado em pesquisas genéticas e estudos populacionais, sendo acessível para laboratórios forenses, mediante contato prévio com o Projeto de Diversidade do Genoma Humano.

Embora os haplogrupos da maior parte da população tenham sido corretamente inferidos pelos três programas simultaneamente, tais diferenças entre cada um dos algoritmos podem influenciar os resultados gerados, impactando de maneiras distintas no desempenho individual da ferramenta (EMMEROVA *et al.*, 2017), justificando as taxas diferenciadas de erros, acertos e incertezas constatados entre os três programas.

6.5.4 Detalhamento das inferências realizadas pelo *STRUCTURE*

Os resultados das inferências realizadas pelo *STRUCTURE*, juntamente com aquelas geradas pelos demais programas encontram-se compiladas no **APÊNDICE 5**.

Ao explorar a distribuição dos *clusters* conduzida pelo programa *STRUCTURE*, considerando a melhor corrida de K=20, foi possível ainda constatar que:

- O clado E foi classificado em três grupos (*clusters*) distintos, evidenciando sua divisão em diferentes haplogrupos (e seus subhaplogrupos). Os haplótipos do clado E classificados como Afrobrasileiros (*Cluster* 1: E1b1a1a1) foram alocados em um único *cluster* isolado. Os haplótipos do clado E classificados como euro-brasileiros foram distribuídos em dois *clusters* diferentes (*Cluster* 2: E1b1b1a1 e *Cluster* 3: E1b1b1b1 e E1b1b1b2). Um indivíduo do haplogrupo E2b1, originalmente classificado como sendo de origem afro-brasileira, foi atribuído ao *cluster* 2. Essa separação é relevante uma vez que permite distinguir, a partir dos haplogrupos do clado E, Afrobrasileiros e euro-brasileiros.

- Os indivíduos do clado J, foram classificados em dois *clusters* diferentes, evidenciando a presença de dois subhaplogrupos distintos (J2a1a e J2b2a). Os treze indivíduos que representam o subhaplogrupo J2a1a foram distribuídos entre os dois clusters. Os dois indivíduos do haplogrupo J2b2a foram mantidos em um mesmo *cluster.*

- Três indivíduos compunham o subhaplogrupo R1a do clado R. Dois foram corretamente classificados, enquanto um indivíduo (P179ASC) foi sistematicamente classificado em um *cluster* do haplogrupo R1b. Observou-se a reprodução deste padrão de predição de haplogrupo pelos softwares *HAPEST* e *NevGen* para este mesmo indivíduo. Este evento não está alinhado com o objetivo de classificação em haplogrupos.

- Os indivíduos do clado R, haplogrupo R1b (e subhaplogrupos) constituíram um cluster único. Um indivíduo (640UCX) da população questionada (urbana brasileira), cuja origem constava como indefinida por falha de genotipagem dos Y-SNPs, foi alocado neste grupo, tanto pelo *STRUCTURE* como pelo *HAPEST e NevGen*.

- Um indivíduo (CP449VVF), cujo haplogrupo é conhecido (E1b1b1a1b1a V13), porém cuja origem geopolítica não está disponível, foi alocado em *clusters* compartilhando clados com o clado E, pelo *STRUCTURE*. Os softwares

HAPEST e *NevGen* também designaram este indivíduo para o clado E, subhaplogrupo E1b1b.

- Na população questionada, 26 indivíduos (16,56%), apresentaram haplótipos incompletos, com perda de amplificação de um a três locos (53,84%) e quatro a seis locos (46,15%). No entanto, as falhas de genotipagem destes indivíduos não comprometeram a precisão da classificação correta dos haplogrupos.

6.5.5 Análise dos erros de atribuição gerados pelo *STRUCTURE*

Na melhor execução do programa STRUCTURE com K=20, foi observado que oito indivíduos foram incorretamente alocados em clados, o que representa 4,47% de taxa de erro na população em análise. No entanto, o *STRUCTURE* é um método em que cada corrida oferece uma iteração diferente, gerando resultados possivelmente diferentes. Assim, também investigamos os erros observados nas outras execuções do K=20, procurando entender a natureza desses erros nesta ferramenta. Nesta investigação estendida, o registro de erro do *STRUCTURE* foi efetuado sempre que, em ao menos uma das 15 repetições, houvesse alguma atribuição incorreta. Observamos que os erros em todas as execuções do K=20 permaneceram consistentemente dentro de uma faixa semelhante, com uma média de 4,43% (**Tabela 12**) e se concentraram em 22 indivíduos (**APÊNDICE 4**).

Tabela 12 - Taxas de erros de atribuição do *STRUCTURE* em cada uma das 15 corridas efetuadas para K=20.

CORRIDAS (K=20)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	MÉDIA
% Erros de Atribuição	4,47	5,59	3,91	3,35	2,79	3,35	6,70	5,03	5,59	4,47	6,15	3,91	2,79	3,91	1,12	4,43

Analisando-se detalhadamente os erros reportados para o *STRUCTURE* para os 22 indivíduos triados, constatou-se que foram registradas

inconsistências em todos os clados da amostra questionada, com exceção do clado L (representado por um único indivíduo). Foi possível observar a sobreposição de 10 indivíduos com erros pelo *HAPEST* (45,45%) e de dois indivíduos pelo *NevGen* (9,09%). Considerando simultaneamente os erros constatados pelos três programas, dois erros foram registrados para os mesmos indivíduos (9,09%). De forma similar, ao se avaliarem as sete incertezas de atribuição observadas para o *STRUCTURE*, observou-se concomitantemente três incertezas na comparação com *HAPEST* (42,86%), e duas na comparação com *NevGen* (28,57%). Não houve sobreposição de incertezas quando os três programas foram comparados simultaneamente. Dentre os 22 erros do *STRUCTURE*, cinco (22,72%) indivíduos apresentaram incertezas de atribuição pelo *NevGen*, enquanto o *HAPEST* os classificou corretamente, com probabilidades entre 99,9-100%, aos clados E, G, I e Q. Seis erros do *STRUCTURE* (27,27%), distribuídos em indivíduos do clado E (n=3) e Q (n=3) foram avaliados corretamente pelo *HAPEST* e pelo *NevGen*, com probabilidades entre 79,09% e 100% (**APÊNDICE 4 e APÊNDICE 5**).

Um dos erros constatados para os três programas se refere à alocação do indivíduo P179ASC ao haplogrupo R1b, embora sua classificação original pelos Y-SNPs tenha classificado ao haplogrupo R1a1a1. O outro erro concomitante se refere ao indivíduo CP335JPS, originalmente classificado como I2a1b1. Ambos os programas *HAPEST* e *NevGen* o classificaram à linhagem principal I, porém o detalhamento da classificação no haplogrupo (I2b1 e I2a2a, respectivamente) justificou os erros registrados.

Com o objetivo de compreender a dinâmica que levou o programa *STRUCTURE* a alocar os 22 indivíduos em haplogrupos de maneira inconsistente com a expectativa original, foram investigados fatores que poderiam ter exercido influência nesse processo:

1) Possíveis Falhas de Genotipagem ou atribuição de haplogrupo

Foi individualmente verificada a genotipagem dos Y-SNPs dos indivíduos triados, através dos resultados fornecidos pela análise com o software *yHaplo* e que contém os SNPs do cromossomo Y derivados de cada amostra. Todos os

22 indivíduos apresentam diversos SNPs específicos que definem os respectivos haplogrupos, não sendo constatadas incoerências.

Foi então verificada a possibilidade de que falhas de amplificação dos marcadores Y-STR que pudessem comprometer a inferência. Constatou-se perda de marcadores no conjunto, com a perda variando de 1 a 5 marcadores. O indivíduo CP445LPS teve o maior número de falhas, com perda de cinco locos (21,74% dos marcadores). No entanto, esta perda não impactou no resultado da inferência realizada pelos outros dois programas. Por outro lado, dois indivíduos para os quais foi verificada a perda de apenas um loco (4,34% dos marcadores) pode ter havido impacto na precisão da alocação aos respectivos haplogrupos: o indivíduo C649OPS (perda de DYS448) foi erroneamente classificado pelo *STRUCTURE* e pelo *HAPEST*, e apresentou incerteza pelo *NevGen*. O indivíduo CP439JCO (perda de YGATAH4) somente foi classificado corretamente pelo *NevGen*. Os demais indivíduos apresentaram 100% de sucesso de amplificação para todos os 23 marcadores testados (**Tabela 13**).

Por outro lado, vale salientar que 20 indivíduos corretamente atribuídos pelo *STRUCTURE* também apresentaram eventual perda de amplificação (1 a 6 perdas). Os marcadores não amplificados variaram, havendo apenas a coincidência do DYS 643. Provavelmente as perdas se devem à qualidade da amostra preservada, uma vez que tentativas de reamplificação não foram bem-sucedidas. Três casos de incertezas pelo *NevGen* e um caso pelo *HAPEST,* abrangem estes indivíduos. Além destes, o indivíduo P123JNO (com quatro perdas de amplificação) foi atribuído com erro pelo *HAPEST*, apresentou incerteza pelo *STRUCTURE* e atribuição correta pelo *NevGen*.

Na literatura consta a recomendação de que a inferência de haplogrupos não deveria ser realizada em casos nos quais poucos Y-STRs (por exemplo, menos do que 12 locos) tenham sido efetivamente genotipados (EMMEROVA *et al.*, 2017). Não foi registrada nenhuma perda superior a seis locos na amostra populacional. O fato de as falhas de genotipagem constatadas não terem sido concentradas em marcadores específicos sugere que os erros de atribuição não estariam diretamente associados a elas. Em geral, para a maior parte dos indivíduos, as falhas de genotipagem de Y-STRs constatadas neste trabalho

realmente não parecem ter comprometido significativamente a capacidade preditiva dos programas. No entanto, os resultados obtidos demonstram o potencial impacto das diferenças nos algoritmos dos três programas na precisão das inferências. O uso concomitante de, no mínimo, dois programas, parece minimizar este efeito.

Tabela 13 – Haplótipos conforme Y-SNPs dos 22 indivíduos que apresentaram incoerências na alocação de haplogrupos pelo STRUCTURE.

Identificação	HAPLOGRUPO (*)	Amplificação (%)	DYS19	DYS385a	DYS385B	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS481	DYS533	DYS549	DYS570	DYS576	DYS635	DYS643	YGATAH4
C649OPS	E1a2a1b-L133.1	95,65	15	14	15	12	28	22	11	11	12	16	9	12	NA	12	17	26	13	14	14	16	20	10	11
GRC134	Q1b1a1a-M3	100	13	14	16	14	31	24	11	14	11	15	11	14	19	15	15	24	12	13	17	18	22	12	11
GRC145	Q1b1a1a-M3	100	13	14	16	14	31	24	11	14	11	15	11	14	19	15	15	24	12	13	17	18	22	12	11
KRC064	Q1b1a1a-M3	100	13	14	14	12	29	24	10	15	12	14	11	13	20	16	19	23	11	12	17	18	23	9	12
P294NAS	Q1b1a1a-M3	100	13	14	14	11	28	24	10	15	12	14	11	13	20	16	18	23	12	12	17	19	23	9	12
P298MNC	Q1b1a1a-M3	100	13	14	14	13	31	25	11	14	13	15	11	12	20	15	17	24	11	13	12	17	23	10	12
P360JIS	Q1b1a1a-M3	100	13	14	14	12	29	24	10	15	12	14	11	13	20	16	19	23	12	13	17	18	23	9	12
C611AKN	E1b1b1b2a1-M123	100	13	16	16	12	29	25	10	11	12	14	10	12	21	15	19	24	10	12	19	18	21	12	11
P324DPS	I1-M253	100	14	14	14	12	27	23	10	11	13	16	10	11	20	14	17	26	11	11	20	16	22	12	11
P364INS	I2a1b1-M223	100	16	15	15	14	30	23	10	12	14	15	10	12	17	18	16	27	13	12	19	17	20	13	11
P389DVC	I2a1b1-M223	100	15	12	15	13	29	23	10	11	13	14	10	12	18	14	18	23	13	12	16	17	21	12	11
CP331TRR	I2a1b1-M223	100	17	12	17	13	30	24	11	11	13	14	10	11	20	17	17	25	12	11	20	17	21	12	11
CP335JPS	I2a1b1-M223	100	15	15	15	13	30	24	10	12	15	15	10	11	20	14	16	24	10	12	18	19	19	11	11
CP439JCO	J2a1a-L26	95,65	13	14	16	13	32	23	10	11	12	14	9	11	20	16	19	23	11	11	14	15	22	NA	11
CP445LPS	E1b1b1a1b1a-V13	78,26	NA	17	18	12	NA	24	10	NA	NA	14	10	12	20	16	15	22	12	12	20	17	21	NA	12
CP446TEO	E1b1b1b2a1-M123	100	13	16	18	13	32	26	10	11	13	14	10	13	20	15	17	25	10	13	18	18	21	13	11
CP456CLO	G-M201	100	15	13	13	14	32	22	12	11	13	16	10	13	21	13	19	21	11	12	18	18	21	11	11
CP457LCJ	E1b1b1a1-M78	100	17	19	20	13	31	24	10	11	13	14	10	11	20	18	15	24	12	13	21	17	20	12	12
CP458DPD	T1a-M70	100	16	14	16	13	29	23	10	13	13	14	9	11	19	15	18	22	12	13	19	16	22	10	11
P147CAS	J2a1a1a2b2-M67	100	14	12	18	14	30	25	10	11	12	14	10	11	21	14	21.2	26	12	11	19	17	23	9	11
P179ASC	R1a1a1-M417	100	14	11	14	13	29	25	10	13	14	16	12	13	19	15	16	21	14	13	18	16	23	11	11.3
P203JCA	R1b1a1b1a-L51	100	14	13	16	13	30	25	10	11	13	15	10	11	20	16	16	25	12	12	16	18	21	9	10

Legenda: NA: falha de amplificação.

2) Tamanho da população referência

Um fator considerado importante é o tamanho da população referência, uma vez que é a partir das frequência alélicas e haplotípicas representativas desta que o programa executa a alocação dos indivíduos da população questionada (PORRAS-HURTADO *et al.*, 2013; ROSENBERG *et al.*, 2005). Este aspecto é particularmente importante para microssatélites, em virtude de sua variabilidade (PORRAS-HURTADO *et al.*, 2013). Logo, a representatividade da população referência para cada haplogrupo pode ser compreendida como essencial à boa execução das inferências pelo *STRUCTURE*, da mesma forma como a disponibilização de bancos de frequências alélicas e haplotípicas é crucial para o adequado desempenho dos demais programas (JANNUZZI *et al.*,

2020; NÚÑEZ *et al.*, 2012). Embora a população referência fosse numerosa (573 indivíduos), alguns clados apresentam realmente poucos indivíduos (A=3, D=7, F=5, S=4 e T=1), uma limitação deste painel já apontada anteriormente (AUTON *et al.*, 2015; PORRAS-HURTADO *et al.*, 2013). No entanto, destes, somente o clado T estava também presente e sub representado na população questionada sendo, de fato, um dos indivíduos triados por erro (**Tabela 14**).

Tabela 14 - Quantidade de indivíduos que compõem cada clado, conforme Y-SNPs, nas populações questionada (total e com erros de atribuição a haplogrupo) e de referência.

CLADOS	População Referência (n)	População Questionada (n)	Taxa de erros por clado na populaçao questionada (%)		
			STRUCTURE	HAPEST	NEVGEN
A	3	0	0	0	0
B	15	0	0	0	0
C	30	0	0	0	0
D	7	0	0	0	0
E	89	40	6	3	1
F	5	0	0	0	0
G	29	7	3	0	0
H	10	0	0	0	0
I	21	12	24	29	10
J	55	14	4	22	4
L	25	1	0	0	0
O	86	0	0	0	0
Q	30	19	20	0	0
RA	56	3	2	2	2
RB	57	79	2	0	0
S	4	0	0	0	0
T	1	1	100	100	0

Com relação à representatividade da população questionada, também foram verificados poucos indivíduos em parte dos clados (G=7, L=1, RA=3, T=1). Destes, somente o clado L não apresentou erro de atribuição.

3) Particularidades nos haplótipos individuais, de acordo com as origens biogeográficas:

Os haplótipos dos indivíduos triados foram na sequência examinados individualmente. Já foi sugerido que diferentes Y-STRs poderiam ser especialmente informativos para diferentes haplogrupos, mas ainda não existe um grupo de Y-STRs identificado com este fim (SCHLECHT *et al.*, 2008). Ainda assim, uma das estratégias adotadas nesta investigação consistiu em buscar

padrões que pudessem sugerir alguma relação entre os haplótipos erroneamente inferidos com a origem a que se esperava que pertencessem. Essa investigação incluiu a análise dos chamados haplótipos modais das diferentes populações como possíveis referências. Para tanto, foi selecionado o haplótipo modal Bantu (CARVALHO *et al.*, 2010; LEITE *et al.*, 2008; THOMAS *et al.*, 2000), representando a ancestralidade africana, com os alelos de cinco marcadores STR; um conjunto de quatro marcadores STR descritos como alelos modais entre indígenas (BARCELOS, 2006); e o haplótipo modal do Atlântico, referência de origem europeia, composto pelos alelos modais de nove marcadores STR (BORTOLINI *et al.*, 2003; THOMAS *et al.*, 2000; WILSON *et al.*, 2001). Entende-se que a comparação aos haplótipos modais é limitada pelo número de marcadores considerados na época em que foram propostos. Além disso, desde então, o número de marcadores STR genotipados aumentou consideravelmente, tornando improvável que um indivíduo apresente o perfil completo do que seria um haplótipo referência de uma população. Assim, a presente investigação não visa estabelecer uma assinatura característica que configurasse o pertencimento de um indivíduo a uma população de origem. Em vez disso, concentrou-se na investigação de padrões ou comportamentos singulares, conforme discutido a seguir.

3.1) Origem africana

O haplótipo modal Bantu (DYS19*15, DYS390*21, DYS391*10, DYS392*11 e DYS393*13) (THOMAS *et al.*, 2000) foi utilizado como referência para comparações. Para o indivíduo P649OPS foi verificada a perda de amplificação do marcador DYS448. Dentre os 22 locos amplificados, foram observadas apenas duas coincidências (DYS19*15 e DYS392*11) com marcadores do haplótipo modal Bantu, enquanto três outros locos (DYS390*22, DYS391*11 e DYS393*12) apresentaram diferença de um passo de mutação. Em contraste, a maior parte das demais amostras de origem africana apresentou ao menos três coincidências com este haplótipo de referência. Além disso, de 22 marcadores amplificados, nove apresentaram alelos exclusivos para esta amostra, diferenciando-a das demais amostras africanas.

3.2) Origem ameríndia

Os alelos DYS19*13, DYS390*24, DYS391*10 e DYS393*13 foram descritos como alelos modais entre os indígenas e, em parte, em populações urbanas brasileiras (BARCELOS, 2006). Sendo assim, comparou-se os seis indígenas, triados por atribuição errônea no *STRUCTURE*, a este haplótipo. A maioria do conjunto total dos indígenas apresenta o mesmo alelo DYS385a*14 e o mesmo alelo modal DYS19*13 (**Tabela 15**), o que pode dever-se à baixa diversidade da população indígena.

Tabela 15 - Comparação entre o haplótipo modal ameríndio e os haplótipos indígenas das amostras triadas por inconsistências na atribuição de haplogrupo pelo STRUCTURE.

Indivíduos	Correspondências com alelos modais ameríndios	Correspondência (%)	Marcadores correspondentes com alelos modais ameríndios	Número de marcadores com diferenças (+/- 1 SMM)	Alelos exclusivos
P298MNC	1	11,11	DYS19*13	1	DYS570*12
GRC134	2	22,22	DYS19*13 - DYS390*2	1	DYS439*14 e DYS643*12
GRC145	2	22,22	DYS19*13 - DYS390*2	1	DYS439*14 e DYS643*12
KRC064	3	33,33	DYS19*13 - DYS390*24 - DYS391*10	1	-
P294NAS	3	33,33	DYS19*13 - DYS390*24 - DYS391*10	1	DYS389I*11 - DYS389II*28
P360JIS	3	33,33	DYS19*13 - DYS390*24 - DYS391*10	1	-

Legenda: SMM=*stepwise mutation model* (modelo de mutação de passo), corresponde a diferença de uma repetição do motivo do STR. Correspondência (%): indica o grau de semelhança com o haplótipo modal.

3.3) Origem europeia:

A investigação dos indivíduos de origem europeia triados incluiu a análise comparativa com o haplótipo modal do Atlântico (HMA), constituído por nove alelos: DYS19*14, DYS385a*11, DYS385b*14, DYS389I*13, DYS389II*29, DYS390*24, DYS391*11, DYS392*13 e DYS393*13 (**Tabela 16**). Exceto por uma única amostra, todos os demais apresentaram marcadores correspondentes ao HMA (1 a 6). Observou-se uma relação inversamente proporcional entre o número de marcadores que diferem por uma repetição e o número de marcadores coincidentes com o HMA.

Tabela 16 - Comparação entre o haplótipo modal atlântico e os haplótipos europeus das amostras triadas por inconsistências na atribuição de haplogrupo pelo STRUCTURE.

Indivíduos	Correspondências com HMA	Correspondência (%)	Marcadores correspondentes com HMA	Número de marcadores com diferenças (+/- 1 SMM)
P364INS	0	0	-	7
C611AKN	1	11,11	DYS389II*29	5
CP439JCO	1	11,11	DYS389I*13	4
CP445LPS	1 (*)	11,11	DYS390*24	2 (*)
CP456CLO	1	11,11	DYS393*13	4
P147CAS	1	11,11	DYS19*14	6
CP335JPS	2	22,22	DYS389I*13 - DYS390*24	5
CP446TEO	2	22,22	DYS389I*13 - DYS393*13	2
P324DPS	3	33,33	DYS19*14 - DYS385b*14 - DYS393*13	3
P389DVC	3	33,33	DYS389I*13 - DYS389II*29 -DYS393*13	5
CP457LCJ	3	33,33	DYS389I*13 - DYS390*24 -DYS393*13	1
CP458DPD	3	33,33	DYS389I*13 - DYS389II*29 - DYS392*13 - DYS393*13	2
P203JCA	4	44,44	DYS389I*13 - DYS390*24 - DYS391*11 - DYS393*13	2
CP331TRR	4	44,44	DYS389I*13 - DYS390*24 - DYS391*11 - DYS393*13	2
P179ASC	6	66,67	DYS19*14 - DYS385a*11 - DYS385b*14 - DYS389I*13 - DYS389II*29 - DYS392*13	2

Legenda: SMM=*stepwise mutation model* (modelo de mutação de passo), corresponde a diferença de uma repetição do motivo do STR; (*) Perda de amplificação de cinco marcadores, quatro fazem parte dos nove do HMA. Correspondência (%): indica o grau de semelhança com o haplótipo modal.

No entanto, é importante registrar uma observação relacionada a esta análise. O indivíduo P179ASC, classificado de maneira definitiva como haplogrupo R1b através da genotipagem Y-SNP, foi simultaneamente atribuído ao haplogrupo R1a pelos três programas aqui utilizados. Na comparação com o HMA, apresentou seis coincidências com os nove alelos modais. Além disso, apresentou dois marcadores adicionais com diferenças de um passo de mutação com estes alelos. Já foi descrito que a precisão da previsão de haplogrupos pode se dever a similaridades entre os haplótipos Y-STR em diferentes haplogrupos, comprometendo a previsão (EMMEROVA *et al.*, 2017). Embora pareça viável interpretar este evento como sendo devido ao haplótipo Y-STR que ele possui, existe ainda a questão evolutiva a ser considerada. R1a e R1b são ambos haplogrupos derivados do haplogrupo R1, embora com distribuições geográficas distintas no continente europeu (R1a no Leste e R1b no Centro e no Oeste), (KING *et al*., 2011; UNDERHILL *et al*., 2015), não descartando-se a possibilidade de convergência, o que poderia explicar sua alocação errônea como R1b. Neste caso, esta ambiguidade de predição somente poderia ser esclarecida pela genotipagem Y-SNP (WANG *et al*., 2015).

Não foi constatada coincidência completa com os haplótipos modais (**Tabela 15** e **Tabela 16**). Porém, independentemente do número de correspondências observadas com os haplótipos modais, estes indivíduos foram erroneamente atribuídos pelo *STRUCTURE*. Portanto, é possível inferir que a

atribuição errada ocorreu independentemente deste aspecto, ao que prosseguimos na investigação.

6.5.6 Análise da inferência de origem biogeográfica a partir dos haplogrupos atribuídos

Em seguida, buscou-se investigar se, apesar dos erros de atribuição observados, estes teriam impactado na possível inferência das origens biogeográficas feita pelo *STRUCTURE*. Observou-se, contudo, expressiva abrangência nas origens biogeográficas obtidas através da inferência dos clados para a vasta maioria destes indivíduos. Isto é, para os indivíduos com erros de atribuição, os clusters onde foram alocados apresentaram conjuntos de clados em vez de um clado somente. Comparativamente, as origens geográficas inferidas para os indivíduos corretamente alocados se mostraram especificidade, a nível continental, principalmente quando a distribuição dos indivíduos ocorreu em *clusters* constituídos por poucos clados. Assim, pode-se concluir que o *STRUCTURE,* por alocar os indivíduos a clados, oferece a possibilidade de inferir a origem biogeográfica.

No caso dos indivíduos selecionados devido a erros de atribuição, observou-se que parte das corridas havia também resultado em inferências corretas, como mencionado anteriormente. No entanto, as inferências corretas foram agrupadas, na maioria das vezes, em *clusters* que englobam múltiplos clados. Devido a esta ampla variedade de clados representados para cada indivíduo, a inclusão de origem mostrou-se inespecífica, abrangendo várias regiões geográficas de origem. Logo, embora as inferências corretas tenham oferecido uma associação potencial dos indivíduos a suas origens biogeográficas, a natureza abrangente dos dados obtidos não proporcionou clareza suficiente para atribuir, de forma inequívoca, uma região específica a um indivíduo.

Novamente, uma exceção foi observada para o mesmo indivíduo P176ASC já citado, originalmente pertencente ao haplogrupo R1a e aqui realocado como R1b por todos os três programas. Através da inferência

oferecida, a origem biogeográfica pode ser determinada como sendo europeia, estando, portanto, adequada independentemente do haplogrupo ser R1a ou R1b (KING *et al*., 2011; UNDERHILL *et al*., 2015).

6.5.7 Investigação da Concordância entre os Métodos Testados

Foram constatados tanto acertos quanto erros simultâneos entre as inferências de ancestralidade paterna, utilizando haplótipos Y-STR, pelos três programas avaliados (**Tabela 17**). Diante disso, procedeu-se a uma análise de concordância entre os resultados, bem como da capacidade do *STRUCTURE* oferecer resultados efetivamente conclusivos em casos de ambiguidade de atribuição obtida através dos outros programas.

Para tal análise, todas as possibilidades de combinação em que *HAPEST* e *NevGen* geraram resultados foram registradas (**Tabela 17**). Na sequência, investigou-se a potencial contribuição do *STRUCTURE* como ferramenta adicional que possibilitasse a atribuição correta ao cenário.

Cada cenário representado na tabela abrange uma situação específica. Os dois primeiros cenários contemplam situações nas quais se obteve concordância entre os resultados de *HAPEST* e *NevGen*.

No cenário um, é possível verificar que o *STRUCTURE* demonstrou uma taxa de concordância de 92,65% com os resultados corretos obtidos pelos outros dois programas, simultaneamente e de maneira independente. No cenário dois, em que os dois programas convergiram, porém erram na atribuição do haplogrupo, o *STRUCTURE* ofereceu 33,33% de capacidade de contribuir com uma atribuição correta.

Nos cenários três a seis, apresentamos possibilidades de discrepância entre os resultados reportados para os programas *HAPEST* e *NevGen*. No cenário três, em que houve erro para *HAPEST* e acerto para *NevGen*, o *STRUCTURE* possibilitou a conclusão definitiva (correta) em 40% dos casos. No cenário quatro, no qual *HAPEST* produziu resultados corretos, porém *NevGen* apresentou incertezas de atribuição, *STRUCTURE* ofereceu a possibilidade de

resultado definitivo (correto), em concordância com *HAPEST*, para 64,71% das situações. No cenário cinco, em que houve somente registros de incerteza de atribuição para ambos (*HAPEST* e *NevGen*), o *STRUCTURE* ofereceu resultado conclusivo para todos os casos (100%). Por fim, o cenário seis não foi observado nas análises efetuadas.

Tabela 17 - Concordância entre acertos, erros e incertezas de atribuição da ancestralidade paterna obtidos através de combinações entre os resultados dos três programas testados para a população questionada.

CENÁRIO	EVENTOS (N=179)	HAPEST	NEVGEN	STRUCTURE
1	**136**	A	A	-
	126	A	A	A
	6	A	A	E
	4	A	A	I
2	**6**	E	E	-
	2	E	E	A
	2	E	E	E
	2	E	E	I
3	**10**	E	A	-
	4	E	A	A
	6	E	A	E
	6	E	A	I
4	**17**	A	I	-
	6	A	I	E
	11	A	I	A
5	**2**	I	I	-
	2	I	I	A
6	**0**	A	E	-

Legenda: Cada cenário representa uma combinação de resultados obtidos das análises realizadas com HAPEST e NevGen. A linha destacada apresenta o total de casos do respectivo cenário, sem contabilizar resultados do STRUCTURE. As linhas subsequentes informam os resultados gerados pelo STRUCTURE dentro do cenário. A: acertos; E: erros e I: incertezas de atribuição.

A partir destes resultados é possível observar que o *STRUCTURE* levou a resultados corretos:

- em 40% das situações nas quais os dois primeiros programas geraram resultados discordantes.
- em 64,71% das situações em que um dos programas não gerou resultados.
- em 100% das situações nas quais nenhum dos dois outros programas gerou resultados.

6.6 Análise das combinações (Misturas) de Haplótipos Y-STRs

Com o objetivo de verificar o comportamento de misturas de perfis genéticos em termos de diversidade, os dados dos haplótipos de Y-STR foram utilizados para criar, por meio de simulações computacionais, cenários que refletissem resultados obtidos a partir de situações reais de misturas, envolvendo pares de indivíduos.

A análise de misturas genéticas foi realizada, primeiramente, combinando-se indivíduos cujos haplogrupos pertencem a uma mesma população (por exemplo, EUR com EUR). Em seguida, foram simuladas misturas entre indivíduos cujos haplogrupos pertencem a populações diferentes (por exemplo, EUR com AFR). Cada um dos 178 indivíduos, aqui classificados em clados e haplogrupos, foi individualmente submetido a este sistema de simulação, sendo então combinado com todos os 177 restantes. Dessa forma, foram realizadas 15.931 combinações computacionais de haplótipos, isto é, misturas de dois a dois indivíduos. Para a produção das misturas, não foram incluídos o indivíduo asiático (em virtude de sua baixa representatividade) e o indivíduo de procedência e haplogrupo desconhecidos, uma vez que os resultados não poderiam ser considerados informativos.

Em misturas genéticas de dois indivíduos, como nesta simulação, supõe-se que alguns alelos de Y-STRs poderão ser coincidentes, enquanto outros serão diferentes entre eles, existindo ainda a possibilidade de 100% de similaridade ou de divergência alélica. Com exceção de misturas dos quatro haplótipos idênticos encontrados, todas as demais misturas realizadas acusaram diferenças, que foram interpretadas como manifestações da diversidade

genética característica dos subgrupos formados por indivíduos com diferentes haplogrupos.

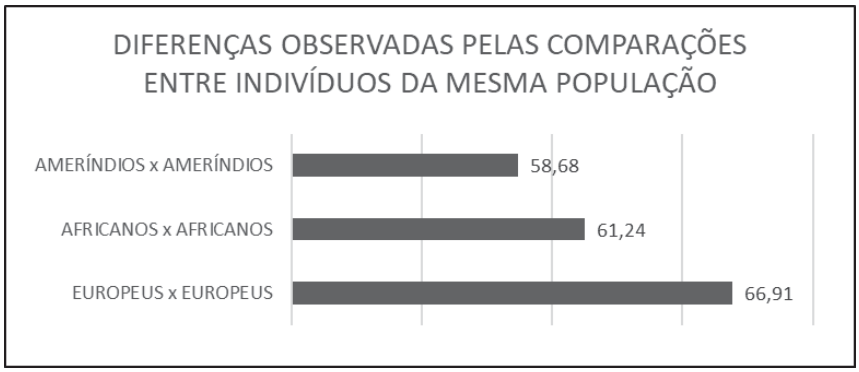
Nesta análise foram, portanto, identificadas, contabilizadas e analisadas todas as diferenças alélicas entre os pares de indivíduos.

O número de marcadores com diferenças alélicas entre cada par de indivíduos e por cada combinação de população de origem dos haplogrupos foi registrada. Foram contabilizadas as diferenças em combinações dentro de uma mesma população (indivíduos de haplogrupos EUR entre si, indivíduos de haplogrupos AFR entre si, indivíduos de haplogrupos AMR entre si) e entre populações distintas (indivíduos de haplogrupos EUR com indivíduos de haplogrupos AFR, indivíduos de haplogrupos EUR com indivíduos de haplogrupos AMR e indivíduos de haplogrupos AFR com indivíduos de haplogrupos AMR). Em outras palavras, todos os indivíduos de clados EUR, quando comparados a todos os demais indivíduos de clados EUR (mesma população), geraram um número médio de diferenças. Por outro lado, estes indivíduos EUR, quando comparados a todos os indivíduos dos clados AFR (duas populações diferentes), geraram um outro valor médio de diferenças. Isto foi feito dessa forma para buscar compreender como se comportam as misturas quando envolvem indivíduos com haplótipos oriundos de uma mesma população ou de populações diferentes. Assim, a dinâmica do comportamento das diferenças nas comparações dentre e entre populações pode ser analisada para cada combinação de população.

As comparações entre os indivíduos dentro da mesma população mostraram menos diferenças alélicas do que entre populações diferentes (**Figura *13*** e **Figura *14***), evidenciando que misturas constituídas por indivíduos pertencentes a haplogrupos distintos apresentam maior média de diversidade, como esperado. As misturas entre haplogrupos EUR apresentaram maior diversidade do que as misturas entre haplogrupos AFR, que apresentaram maior diversidade do que as misturas entre haplogrupos AMR. Esses resultados são consistentes com a história demográfica dos AMR (CAVALLI-SFORZA; PIAZZA, 1993), mas contrastam com as expectativas baseadas na história evolutiva dos AFR (PEREIRA *et al.*, 2021). Esta observação pode se dever à viés da

amostragem populacional, dado o pequeno número de indivíduos AFR e ao expressivo número de indivíduos EUR. Estes resultados ainda reproduzem a análise de diversidade gênica calculada por Y-STRs, porém a diferença no número de haplogrupos originários destas três populações pode ter influenciado nesses resultados.

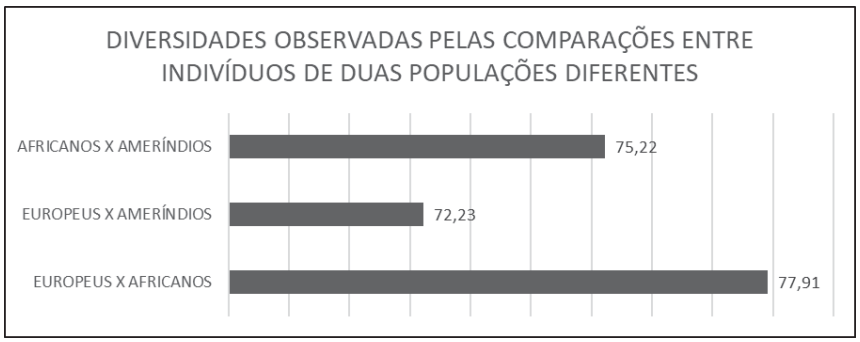
Figura 13 - Média da porcentagem do número total de locos que apresentam diferenças alélicas nas misturas entre indivíduos da mesma população.



Fonte: A Autora, 2023.

As diferenças contabilizadas para os Y-STRs a partir das misturas entre populações mostraram que a variabilidade genética (porcentagem de marcadores com divergências na mistura) se assemelhou à diversidade gênica estimada a partir das frequências alélicas nas análises populacionais, com médias semelhantes para as populações de haplogrupos EUR (66,91% vs. 67%), AFR (61,24% vs. 63%) e AMR (58,68% vs. 58%) (**Tabela 6** e **Figura 13**).

Figura 14 - Média da porcentagem do número total de locos que apresentam diferenças alélicas nas misturas entre indivíduos de populações diferentes.

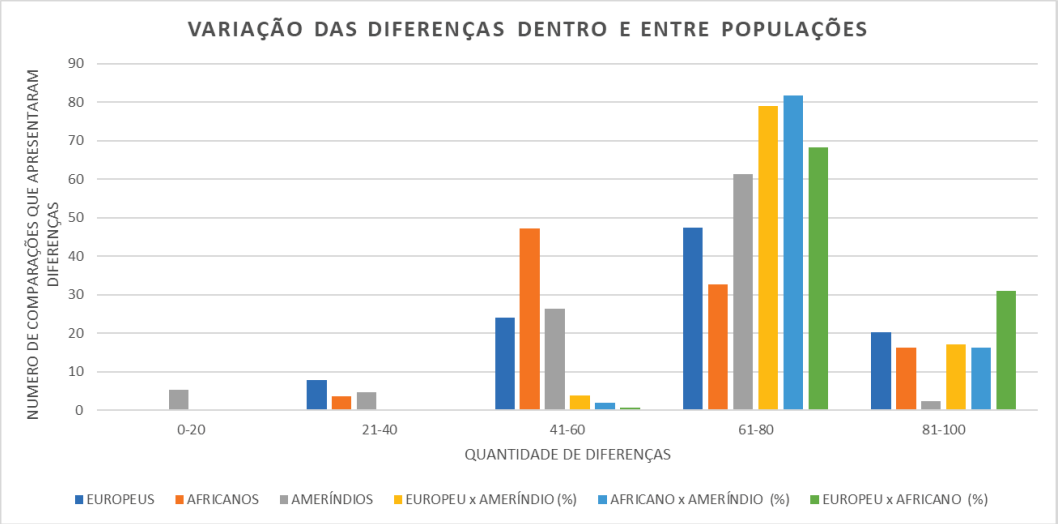


Fonte: A Autora, 2023.

A partir da análise efetuada, foi possível extrair duas informações relevantes para a interpretação do comportamento das misturas: a quantidade de diferenças observadas entre alelos de Y-STR e a frequência com que cada montante de diferenças ocorreu. Em termos simples, ao comparar os Y-STR de dois indivíduos em misturas simuladas, poderíamos encontrar casos em que eles eram totalmente idênticos. No entanto, se houvesse discrepâncias entre os Y-STRs, estas poderiam abranger de um a 23 marcadores. Neste último caso, todos os 23 Y-STRs seriam diferentes. Nossa análise focou-se no número de diferenças observadas entre os Y-STRs e registrou a frequência com que cada quantidade de diferenças ocorreu. Para realizar essa contagem, as discrepâncias foram organizadas em categorias que abrangiam de 0 a 20%, 21 a 40%, 41 a 60%, 61 a 80%, e 81 a 100% de Y-STRs diferentes.

Durante a análise, observou-se que a frequência de eventos com poucas diferenças alélicas (0-20%) foi relativamente baixa. Constatou-se um aumento progressivo na incidência à medida que as diferenças se intensificavam, alcançando picos notáveis na faixa de 61 a 80%. Contudo, essa tendência foi seguida por uma diminuição na frequência, mesmo com o aumento contínuo no número de diferenças. Isso indica uma escassez de eventos com poucas ou muitas diferenças alélicas. Isto é, poucas foram as comparações em que os dois indivíduos foram totalmente iguais. Já o número de comparações em que houve alguma diferença (20% a 100%) ocorreram mais vezes. O comportamento das diferenças foi diferente quando envolveu diferentes haplogrupos. Em todos os cenários, o padrão de frequência das diferenças alélicas foi unimodal, concentrando a maior frequência para diferenças alélicas entre 40-80% do número de Y-STRs comparados (**Figura *15***).

Figura 15 - Variação da porcentagem do número de loci que apresentam diferenças alélicas em Y-STRs observada nas combinações de misturas dentro e entre os haplogrupos de diferentes origens.



Fonte: A Autora, 2023.

Estes achados demonstram que a frequência de diferentes magnitudes de diferenças alélicas em misturas é influenciada pelas populações de origem dos haplótipos envolvidos, com misturas de populações diferentes apresentando mais diferenças do que misturas de populações semelhantes conforme esperado.

Considerando que a variabilidade dos Y-STRs é mais subdividida por haplogrupos do que por populações (MUZZIO *et al.*, 2011), foram aprofundadas ainda mais as investigações, visando aprimorar o grau de detalhamento das misturas e coletar informações mais específicas. Até então, essas misturas haviam sido interpretadas considerando simplesmente as linhagens principais (EUR, AFR e AMR), sem levar em conta os níveis específicos dos haplogrupos que cada população contemplava. A dúvida seria se o sistema de detecção de diferenças poderia ser utilizado para inferir haplogrupos e suas origens. No entanto, a representatividade dos haplogrupos na amostra populacional brasileira mostrou-se muito discrepante (n=1 a 52), o que reduziu consideravelmente o poder informativo que poderia ser extraído destas análises. Idealmente estudos desta natureza requereriam uma distribuição populacional mais proporcional e representativa dos diferentes haplogrupos envolvidos. No entanto, esta situação não é realizável, considerando que a distribuição habitual

dos haplogrupos nas populações é de fato variável, com haplogrupos mais frequentes que outros.

6.6.1 Comportamento dos marcadores Y-STR nos diferentes cenários de mistura

O comportamento de cada marcador Y-STR foi analisado em cada uma das 15.931 combinações efetuadas, como já exemplificado na **Tabela 2**. As diferenças alélicas nas misturas foram então totalizadas para cada par de populações envolvido na mistura. Assim, todas as diferenças que o marcador DYS19, por exemplo, apresentou em diferentes cenários, puderam ser contabilizadas: no total das 15.931 comparações, ou apenas nas comparações realizadas entre haplótipos EUR ou apenas quando comparados haplótipos AFR e AMR, e assim por diante. Procedeu-se dessa forma com os 23 marcadores do sistema, possibilitando analisar individualmente cada um dos marcadores Y-STR nas diferentes combinações de misturas populacionais. Dessa maneira foi possível determinar com precisão onde, exatamente, se localizavam as diferenças encontradas nas análises anteriores.

Foi possível identificar os marcadores que apresentaram as maiores e menores quantidades de diferenças nos diferentes tipos de misturas (**Tabela 18**).

Os resultados demonstram que cada marcador apresentou um padrão de comportamento característico para cada composição de mistura entre as populações estudadas. Esse padrão é medido pelo percentual de diferenças alélicas, calculado dividindo-se o número de diferenças de alelo para um determinado marcador pelo total de comparações efetivas deste mesmo marcador. Os valores encontrados foram denominados taxas de divergência (ou de diferença), sendo possível verificar que alguns marcadores apresentam maiores taxas de divergência do que outros.

As análises das misturas realizadas apontaram o marcador DYS385 como o mais informativo em todas as configurações testadas, reproduzindo o resultado de outras publicações (ALVES, 2012; GRATTAPAGLIA *et al.*, 2005; PALHA *et al.*, 2012; PURPS *et al.*, 2014). Entretanto, como já descrito anteriormente, este

marcador apresenta uma peculiaridade que, além de aumentar sua informatividade, impacta especificamente na configuração feita para análises de misturas nos moldes conduzidos neste trabalho: possui duas regiões alvo distintas, separadas por uma pequena distância de apenas 40pb, o que habitualmente implica que sejam amplificados em conjunto. Assim, diferentemente dos demais marcadores, o DYS385 fornece dois alelos ao invés de um, que podem ser iguais ou diferentes entre si (BUTLER *et al.*, 2006; NIEDERSTÄTTER *et al.*, 2004). Além disso, quando utilizada a PCR tradicional, os fragmentos amplificados não podem ser relacionados de forma indubitável a um loco definido dentro do marcador DYS385, a não ser que sejam idênticos (BUTLER *et al.*, 2006). Como já mencionado, no presente caso das combinações de misturas foi necessário realizar uma adaptação para o loco DYS385, na qual qualquer diferença detectada entre ao menos um dos dois alelos do DYS385 de duas amostras distintas seria contabilizada como uma diferença, mesmo que o outro alelo fosse igual. Essa adaptação possibilitou o uso deste marcador nas análises efetuadas, porém sua característica bialélica apresenta a fragilidade de aumentar a probabilidade de serem detectadas diferenças nas comparações. Provavelmente por este motivo não foi identificado outro marcador que apresentasse comportamento comparável à magnitude das diferenças observadas para o DYS385 em todas as combinações simultaneamente (**Tabela 18**).

O marcador DYS392 é um bom exemplo da variabilidade encontrada entre as taxas de divergência nas diferentes misturas e de como este dado pode ser interpretado em casos de mistura. Em misturas de indivíduos de haplogrupos AFR com indivíduos de haplogrupos AMR, a taxa de divergência deste marcador foi de 100%, sugerindo, em teoria, que ele sempre se apresentaria diferente entre os dois indivíduos. Já em misturas entre dois indivíduos de haplogrupos AFR, sua taxa de divergência é 0%, sugerindo fortemente que ele sempre apresentaria alelos iguais entre os dois indivíduos AFR. Entre indivíduos de haplogrupos AMR, no entanto, sua taxa de divergência é inferior a 50% (47,06%), admitindo, portanto, uma maior taxa de similaridade (52,94%), o que pode ser interpretado como indicativo de que seria ligeiramente mais esperado

que os alelos fossem iguais, do que diferentes, nas misturas entre AMR (**Tabela 18**).

Tabela 18 - Percentual das diferenças observadas em cada Y-STR a partir de misturas simuladas envolvendo haplótipos de diferentes haplogrupos, tomando como base as populações brasileiras urbana e indígena analisadas no presente estudo.

MARCADORES Y-STR	EUROPEUS x EUROPEUS	AFRICANOS x AFRICANOS	AMERÍNDIOS x AMERÍNDIOS	EUROPEUS x AFRICANOS	EUROPEUS x AMERÍNDIOS	AFRICANOS x AMERÍNDIOS
DYS393	45,81	76,36	70,76	69,64	61,05	78,47
DYS437	54,33	20,00	36,60	63,77	57,91	30,00
DYS391	59,01	43,64	49,12	53,99	54,45	44,02
DYS389I	56,19	72,73	71,93	65,72	64,44	67,94
YGATAH4	57,13	69,09	66,08	69,26	62,46	66,51
DYS390	59,69	34,55	44,44	97,85	52,95	98,56
DYS533	63,36	56,36	60,78	68,30	60,77	62,81
DYS19	59,88	80,56	29,82	84,95	80,36	98,25
DYS456	67,96	64,44	52,94	66,25	60,82	60,00
DYS392	60,37	0,00	47,06	57,04	95,77	100,00
DYS549	68,85	73,33	52,63	72,18	62,62	74,21
DYS439	66,89	70,91	77,78	67,37	72,13	70,33
DYS635	67,50	67,27	52,63	79,67	67,57	82,30
DYS643	68,18	70,91	66,67	76,79	68,01	82,30
DYS438	61,33	18,18	19,88	95,04	91,42	18,66
DYS448	69,00	20,00	40,94	83,88	66,63	97,37
DYS389II	68,72	76,36	73,10	80,14	72,05	74,64
DYS576	76,50	83,64	80,70	85,32	79,98	91,39
DYS481	75,46	89,09	61,99	95,21	89,58	86,60
DYS458	80,23	58,18	83,63	74,77	85,32	87,56
DYS570	82,93	76,36	64,33	87,41	78,09	82,78
DYS385ab	88,74	96,36	81,29	99,02	96,94	96,17

Legenda: Os valores indicam a porcentagem das diferenças observadas para cada Y-STR. A escala de cores em gradiente indica menor diferença (tendência para azul) e maior diferença (tendência para vermelho), com valores intermediários passando pela cor branca.

Com base nas taxas de divergência apresentadas na **Tabela 18** procuramos investigar possíveis relações entre os marcadores frente a diferentes cenários de misturas. Isto é, utilizando-se as taxas de divergência, procurou-se verificar se estas poderiam ser interpretadas como indicativos de tendências para que marcadores fossem diferentes ou semelhantes em cada combinação populacional. Em outras palavras, cada taxa de divergência constatada foi interpretada como uma sugestão de padrão ou tendência que poderia oferecer uma estimativa do comportamento dos marcadores Y-STR em uma mistura envolvendo dois indivíduos. Taxas de divergência superiores as

50% foram traduzidas como indicativos de tendência de divergência, enquanto as inferiores a este valor foram interpretadas como indicativos de tendência de similaridade. Na análise das misturas, a identificação das tendências citadas poderia representar uma maneira de identificar os haplogrupos dos indivíduos envolvidos. Estas tendências foram reunidas na **Tabela 19** para proporcionar melhor visualização.

A capacidade do marcador apresentar comportamentos diferenciados nas diferentes misturas é o que lhe confere maior poder informativo. Por exemplo, o marcador DYS391 apresentou-se idêntico entre os indivíduos em metade dos seis cenários de misturas analisados, enquanto os marcadores DYS19 e DYS393 se mostraram iguais, cada um deles, em apenas uma das misturas efetuadas. Assim, pode-se entender que DYS19 e DYS393 seriam mais informativos nas misturas de populações aqui estudadas.

Tabela 19 – Interpretação de tendências baseadas no percentual das diferenças observadas em cada Y-STR a partir de misturas simuladas envolvendo haplótipos, tomando como base as populações brasileiras urbanas e indígenas analisadas no presente estudo.

Y-STR	EUROPEUS x EUROPEUS	AFRICANOS x AFRICANOS	AMERÍNDIOS x AMERÍNDIOS	EUROPEUS x AFRICANOS	AFRICANOS x AMERÍNDIOS	EUROPEUS x AMERÍNDIOS
DYS385ab	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS439	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS389I	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS389II	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS456	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS458	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS481	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS533	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS549	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS570	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS576	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS635	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS643	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
YGATAH4	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS19	DIFERENTE	DIFERENTE	IGUAL	DIFERENTE	DIFERENTE	DIFERENTE
DYS390	DIFERENTE	IGUAL	IGUAL	DIFERENTE	DIFERENTE	DIFERENTE
DYS391	DIFERENTE	IGUAL	IGUAL	DIFERENTE	IGUAL	DIFERENTE
DYS392	DIFERENTE	IGUAL	IGUAL	DIFERENTE	DIFERENTE	DIFERENTE
DYS393	IGUAL	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE	DIFERENTE
DYS437	DIFERENTE	IGUAL	IGUAL	DIFERENTE	IGUAL	DIFERENTE
DYS438	DIFERENTE	IGUAL	IGUAL	DIFERENTE	IGUAL	DIFERENTE
DYS448	DIFERENTE	IGUAL	IGUAL	DIFERENTE	DIFERENTE	DIFERENTE

Legenda: Diferente: Y-STR com taxas de divergência > 50%. Igual: Y-STR taxas de divergência < 50%.

6.6.2 Misturas de haplótipos de Y-STRs compostas por indivíduos de mesma origem

De acordo com os resultados já relatados, espera-se que, em uma mistura envolvendo indivíduos com haplótipos de mesmo haplogrupo, obtenham-se os valores de diferença semelhantes aos reportados na **Figura *13***. Por exemplo, em misturas entre indivíduos da população brasileira pertencentes a haplogrupos AMR, supõe-se que cerca de 58% dos marcadores Y-STR terão alelos diferentes. As análises ainda permitiram organizar as taxas de divergência de maneira a proporcionar um ranqueamento de marcadores mais prováveis de serem diferentes em cada cenário de mistura. A partir desta organização e da expectativa de quantidade de diferenças, foi investigada a presença de padrões nos comportamentos individuais dos marcadores nas misturas envolvendo as mesmas populações ancestrais:

- Cenário 1 (mistura entre dois indivíduos de origem EUR): considerando a diferença atestada para misturas envolvendo indivíduos da população urbana pertencentes a haplogrupos EUR (66,91%), espera-se que neste cenário sejam detectadas diferenças em aproximadamente 15 marcadores, dentre os 23 testados. Os locos que apresentaram as maiores taxas de divergência foram DYS385a|b, DYS570, DYS458, DYS576, DYS481, DYS448, DYS549, DYS389II, DYS643, DYS456, DYS635, DYS439, DYS533, DYS438 e DYS392. No entanto, a expectativa das diferenças para fins de identificação dos haplogrupos dos indivíduos desta mistura não deve se limitar a este subconjunto específico de locos. Isso se justifica porque os dados observados para este cenário indicam que 22 marcadores têm taxas de divergência acima de 60%. O marcador DYS393 foi o único a apresentar, ainda que limítrofe, uma taxa maior de similaridade (54,19%) do que de diferença para os dois indivíduos, numa mistura deste cenário.
- Cenário 2 (mistura entre dois indivíduos de origem AFR): considerando a diferença atestada para misturas envolvendo haplótipos dessa população (61,24%) espera-se que neste cenário sejam detectadas diferenças em aproximadamente 14 marcadores, dentre os 23 testados. Os locos que apresentaram maiores taxas de divergência foram DYS385a|b, DYS481,

DYS576, DYS570, DYS389II, DYS549, DYS643, DYS439, DYS635, DYS456, DYS458, DYS533, DYS448 e DYS438. Seguindo o raciocínio já explicado, a expectativa das diferenças com o propósito de identificação dos haplogrupos dos indivíduos desta mistura não deve se limitar a este subconjunto específico de locos. Na verdade, 16 marcadores ao todo mostraram maiores taxas de divergência (acima de 56,36%) do que de similaridade. Neste cenário, os seis restantes (DYS392, DYS438, DYS437, DYS448, DYS390 e DYS391) têm maior taxa de similaridade (acima de 56%) do que de diferença entre os dois indivíduos da mistura e um loco (DYS392) apresentou taxa absoluta de similaridade (100%) na subamostra envolvendo indivíduos da população urbana pertencentes a haplogrupos AFR (n = 11) analisada no presente estudo.

- Cenário 3 (mistura entre dois indivíduos de origem AMR): considerando a diferença encontrada para misturas envolvendo indivíduos da população urbana pertencentes a haplogrupos AMR (58,68%), espera-se que neste cenário sejam detectadas diferenças em cerca de 13-14 marcadores, dentre os 23 testados. Os locos que apresentaram maiores taxas de divergência foram DYS458, DYS385a|b, DYS576, DYS439, DYS389II, DYS389I, DYS393, DYS643, YGATAH4, DYS570, DYS481, DYS533 e DYS456. Neste cenário também se recomenda evitar a concentração apenas neste grupo de marcadores, uma vez que são 15 os locos com taxas de divergência acima de 52,63%. Sete outros marcadores (DYS438, DYS19, DYS437, DYS448, DYS390, DYS392 e DYS391) tiveram taxas de similaridade maiores (superiores a 50,88%) do que de divergência. A taxa de divergência do marcador DYS438 chega a ser superior a 80%.

O número de marcadores que se espera que sejam diferentes não deveria sugerir um resultado que possa ser correlacionado a nenhum dos cenários com exclusividade, pois as médias apresentam elevado desvio padrão. Além disso, a a probabilidade de uma mistura ser constituída por dois indivíduos da população urbana pertencentes a haplogrupos EUR é provavelmente maior ao se considerar sua frequência na população brasileira. Entre os marcadores mais informativos, foram observadas várias ocorrências coincidentes em diferentes cenários, o que é esperado dada a quantidade limitada do painel de 23 marcadores. Foi possível, no entanto, detectar que alguns marcadores, ou

conjuntos de marcadores, sugerem a possibilidade de contribuir isoladamente como um apoio na inferência do cenário envolvido.

6.6.3 Misturas de haplótipos de Y-STRs compostas por indivíduos de origens distintas

A análise das misturas envolvendo dois indivíduos com haplótipos de diferentes haplogrupos revelou que:

- Cenário 4 (mistura entre um indivíduo de haplogrupo EUR e um indivíduo de haplogrupo AFR): considerando a diferença encontrada para misturas envolvendo indivíduos da população urbana pertencentes a estes haplogrupos (77,91%), espera-se que neste cenário sejam detectadas diferenças em aproximadamente 18 marcadores, dentre os 23 testados. Os locos que apresentaram maiores taxas de divergência foram DYS385a|b, DYS390, DYS481, DYS438, DYS570, DYS576, DYS19, DYS448, DYS389II, DYS635, DYS643, DYS458, DYS549, DYS393, YGATAH4, DYS533 e DYS439. No entanto, todos os 23 marcadores mostraram taxas de divergência superiores a 50%, numa mistura de dois indivíduos neste cenário.
- Cenário 5 (mistura entre um indivíduo de haplogrupo AFR e um indivíduo de haplogrupo AMR): considerando a diferença encontrada para misturas envolvendo indivíduos da população urbana pertencentes a estes haplogrupos (75,22%) espera-se que neste cenário sejam detectadas diferenças em cerca de 17 marcadores, dentre os 23 testados. Os locos que apresentaram maiores taxas de divergência foram DYS392, DYS390, DYS19, DYS448, DYS385a|b, DYS576, DYS458, DYS481, DYS570, DYS643, DYS635, DYS393, DYS389II, DYS549, DYS439, DYS389I e YGATAH4. Ao todo, entretanto, 19 dos marcadores apresentam potencial maior que 60% para serem diferentes. Por outro lado, são três os marcadores (DYS438, DYS437 e DYS391) que apresentaram taxas maiores de similaridade (55,98%, 70% e 81,34%, respectivamente).
- Cenário 6 (mistura entre um indivíduo de haplogrupo EUR e um indivíduo de haplogrupo AMR): considerando a diferença encontrada para misturas envolvendo indivíduos da população urbana pertencentes a estes

haplogrupos (72,23%) espera-se que neste cenário também sejam detectadas diferenças em, ao menos, 17 marcadores, dentre os 23 testados. Os locos que apresentaram maiores taxas de divergência foram DYS385a|b, DYS392, DYS438, DYS481, DYS458, DYS19, DYS576, DYS570, DYS439, DYS389II, DYS643, DYS635, DYS448, DYS389I, DYS549 e YGATAH4. No entanto, todos os 23 marcadores mostraram taxas de divergência superiores a 50%, numa mistura de dois indivíduos neste cenário.

Novamente foi possível constatar que o número de marcadores para os quais se espera que existam diferenças é bastante semelhante nos três novos cenários, destacando-se as coincidências entre os marcadores. Como feito anteriormente, foi investigada a presença de marcadores específicos que pudessem, de forma mais objetiva, sugerir qual o cenário mais provável.

6.6.4 Inferência da origem biogeográfica paterna dos contribuintes de misturas populacionais

Tomando-se como base as análises de misturas nos diferentes cenários descritos, acreditamos que seja possível supor que o comportamento dos marcadores Y-STR poderia ser aplicado para a realização da inferência de haplogrupos dos contribuintes de misturas dentre e entre populações urbanas e indígenas brasileiras. Um exemplo de como tais tendências poderiam ser aplicadas encontra-se descrito a seguir.

Fundamentalmente, qualquer análise dessa natureza requereria que a genotipagem fosse executada com precisão, garantindo a amplificação bem-sucedida do maior número possível de marcadores. A melhor situação seria, naturalmente, aquela em que todos os locos sejam adequadamente amplificados. No entanto, a determinação de um limiar mínimo de locos amplificados demanda estudos de validação forense que, por sua vez, dependeriam de estudos com maior número populacional.

Uma vez identificado um perfil de mistura de marcadores Y-STRs, dever-se-ia ainda confirmar que fosse constituído por, no máximo, dois indivíduos,

representado pela presença de até quatro alelos diferentes no marcador DYS385a|b, e até dois alelos nos demais marcadores.

Na impossibilidade de identificação dos contribuintes dos perfis genéticos obtidos, por falta de suspeitos ou pela não obtenção de coincidências em bancos de perfis genéticos, dar-se-ia início à investigação dos prováveis haplogupos dos contribuintes da mistura, para a qual sugere-se consulta à **Tabela 18** (que registra o percentual das diferenças observadas em cada Y-STR a partir de misturas simuladas envolvendo haplótipos de diferentes haplogrupos) e à **Tabela 19** (que resume as tendências registradas para os marcadores em cada um dos cenários estudados).

Consideramos imprescindível que a interpretação seja realizada à luz de todo o conjunto de marcadores que tenham sido amplificados. Por exemplo, a presença de diferença nos locos DYS390 e DYS392 pode ser erroneamente interpretada como envolvendo misturas de haplogrupos AFR e AMR. Neste caso, entretanto, existe o detalhe de que, nas misturas de haplogrupos AFR e AMR, se espera que sejam três os locos idênticos, como explicado anteriormente.

O primeiro passo consistiria em investigar a existência de diferenças alélicas entre os perfis obtidos. Havendo diferenças, estas devem ser contabilizadas. Esta informação não deveria servir como um critério de conclusão definitiva, sendo apenas útil para complementar a interpretação ao final das análises. A presença de poucas diferenças seriam um indício de que se trate de misturas de indivíduos de mesmo haplogrupo. Caso sejam visualizadas muitas diferenças, poder-se-ia considerar a possibilidade de se estar diante de uma mistura envolvendo indivíduos de haplogrupos distintos.

Na sequência, com base nos achados observados na presente amostra populacional brasileira, a investigação poderia ser primordialmente concentrada naqueles locos identificados como mais informativos, como em situações exemplificadas abaixo:

- DYS393: caso os dois indivíduos apresentassem alelo idêntico para este marcador, considerar-se-ia a possibilidade de serem dois indivíduos de haplogrupos EUR da população urbana brasileira. Aqui caberia a ressalva de que alelos comuns em cada população poderiam apresentar maior probabilidade

de ocorrer com maior frequência, demandando validações forenses para este tipo de análise. Esta interpretação poderia ser reforçada pelo número de diferenças no conjunto de marcadores.

- DYS392: caso os dois indivíduos apresentassem alelo idêntico para este marcador, provavelmente se trataria de mistura entre dois indivíduos de haplogrupos AFR. Nesta situação, poderiam ainda ser iguais os marcadores DYS438, DYS437, DYS448, DYS390 e DYS391. Entende-se que, quanto maior a adequação a esta configuração, incluindo por exemplo a presença de aproximadamente 14 diferenças no conjunto, maior o reforço da tendência de que a mistura envolveria indivíduos de haplogrupos AFR da população brasileira.
- DYS19: caso os dois indivíduos apresentassem alelo idêntico para este marcador, provavelmente se trataria de mistura entre dois indivíduos de haplogrupos AMR. Nesta situação, poderiam ainda ser iguais os marcadores DYS438, DYS437, DYS448, DYS390, DYS392 e DYS391. Esta interpretação poderia ainda ser reforçada no perfil múltiplo que apresentasse poucas diferenças.
- A identificação de alelos iguais para os marcadores DYS438, DYS437 e DYS391, aliada à presença de maior quantidade de diferenças poderia indicar uma mistura entre indivíduos de haplogrupos AFR e AMR. A distinção para com a situação narrada no item anterior se basearia em que, no presente caso, os demais marcadores listados (DYS19, DYS448, DYS390 e DYS392) se apresentaram com alta probabilidade de serem diferentes.
- A inferência dos haplogrupos em misturas envolvendo indivíduos de origens distintas, porém contendo um indivíduo de haplogrupo EUR, demonstrou representar um desafio maior. Trata-se das misturas envolvendo indivíduos de haplogrupos EUR com indivíduos de haplogrupos AFR ou com indivíduos de haplogrupos AMR. Em ambas as situações, não foram detectados marcadores com taxas de similaridade expressivas entre os indivíduos. Assim, concentrou-se em buscar marcadores que tivessem maiores taxas de similaridade em outras populações contrastando com altas taxas de divergência nas misturas envolvendo indivíduos de haplogrupos EUR e uma das outras populações:
    - O marcador DYS390 apresenta, proporcionalmente, altas taxas de discrepância (97,85% e 98,56%, respectivamente) quando a mistura envolve tanto indivíduos de haplogrupos EUR e indivíduos de haplogrupo AFR quanto de

indivíduos de haplogrupos EUR de haplogrupos AMR. Nas misturas envolvendo haplogrupos EUR e haplogrupos AMR, estas taxas de discrepância se reduzem para 52,95%. Logo, se no conjunto de marcadores em que houvesse muitas diferenças e este loco fosse igual, poder-se-ia considerar a possibilidade de que se trate de uma mistura entre indivíduos de haplogrupos EUR e haplogrupos AMR.

- O marcador DYS392 apresenta, proporcionalmente, altas taxas de discrepância (100% e 95,77%, respectivamente) quando a mistura envolve indivíduos de haplogrupo AFR e AMR ou indivíduos de haplogrupos EUR e AMR. Nas misturas envolvendo haplogrupos EUR e AFR, a taxa de discrepância reduz para 57,04%. Logo, se, no conjunto de marcadores em que houvesse muitas diferenças e este loco fosse igual, poder-se-ia considerar a possibilidade de que se tratasse de uma mistura entre haplogrupos EUR e AMR.

O exercício descrito acima teve como propósito exemplificar a possível aplicação do processo de inferência de ABG. No entanto, é importante salientar que as taxas de divergência aqui obtidas, bem como as análises de tendências a partir delas extraídas, foram baseadas em uma amostra brasileira muito pequena, comprometendo sua efetiva representatividade. Dessa maneira, admite-se que os cenários obtidos a partir de outras localidades brasileiras, ou mesmo com outras populações, possa ser diferente.

Além disso, haverá situações em que não seja possível inferir a origem dos contribuintes de uma mistura genética. Falhas de amplificação configuram um dos principais riscos para a aplicação desta abordagem. No entanto, resultados inconclusivos também podem ocorrer em virtude de que as probabilidades foram estimadas a partir de comparações que apresentaram certas limitações, como a discrepância entre os números de haplótipos das diferentes ancestralidades avaliadas aliada ao pequeno número amostral. Esta situação está clara pelo fato de que a diversidade constatada para as populações principais mostrou a influência da discrepância dos tamanhos populacionais, como o fato de que a diversidade do haplogrupo AFR ter sido inferior ao do haplogrupo EUR.

Entende-se que uma real aplicabilidade prática desta abordagem (inferência de haplogrupos em misturas forenses), embora pareça promissora, requer o aprimoramento do estudo e da metodologia, principalmente através de um número amostral mais expressivo e efetivamente representativo dos diversos haplogrupos que compõem a população brasileira urbana e indígena. Além da correção do tamanho populacional, para possibilitar a aplicabilidade da proposição relatada, também seria necessário estabelecer um mecanismo de valoração estatística dos achados. Por fim, este sistema precisaria ser testado com casos reais de misturas genéticas.

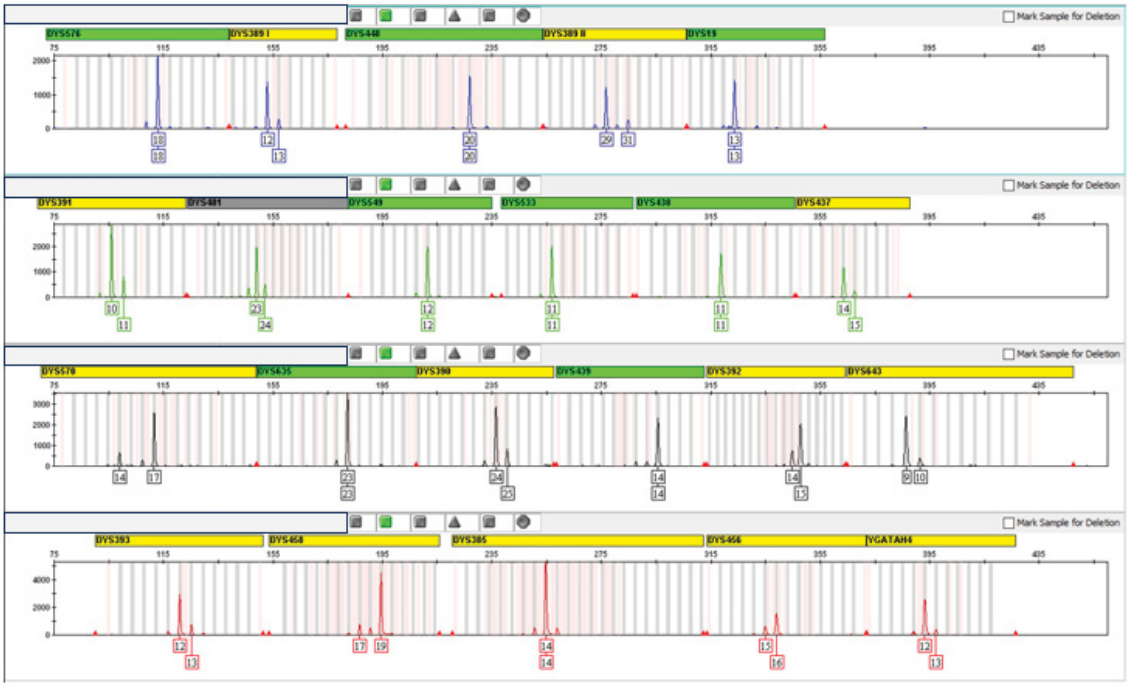
6.6.5 Relato de caso de mistura de Y-STR submetido à análise da origem biogeográfica

A partir das diretrizes descritas, e ainda como uma extensão do exercício proposto anteriormente, foi analisado um perfil de mistura procedente de um caso real da rotina forense (**Figura *16***). Trata-se de uma agressão sexual, cometida em região de reserva indígena e, desconhecendo-se a autoria, seria interessante informar à investigação se os autores pertenceriam a algum haplogrupo diverso aos esperados na população brasileira.

As análises genéticas possibilitaram a obtenção de um eletroferograma dos Y-STRs evidenciando tratar-se de um perfil múltiplo constituído por dois contribuintes. Não foram encaminhados suspeitos para realização de confronto.

Todos os locos foram amplificados com sucesso, revelando um perfil múltiplo que contempla os critérios necessários para ser reconhecido como constituído por duas pessoas distintas. Foram contabilizados 13 marcadores com diferenças e nove marcadores iguais. A análise de cada região igual revelou que os locos DYS19, DYS438 e DYS448 são idênticos, além de DYS576, DYS549, DYS533, DYS635, DYS385 e DYS439.

Figura 16 - Eletroferograma de Y-STR obtido a partir de caso real de mistura de cromossomos Y.



Conforme a análise para inferência de ABG desenvolvida, a presença de alelo idêntico para o marcador DYS19 sugeriria mistura entre dois indivíduos de haplogrupos AMR. Esta interpretação pode ser reforçada pelo fato de que os alelos dos locos DYS438 e DYS448 se apresentaram iguais. E, por fim, reforçariam esta interpretação as poucas diferenças constatadas entre os perfis genéticos obtidos para os dois indivíduos.

Conforme já descrito, nas simulações realizadas para este tipo de mistura (entre haplogrupos AMR), os locos DYS19 e DYS438 apresentaram as maiores taxas de similaridade (acima de 70%). Para o loco DYS448 a taxa de similaridade foi registrada em 59,06% das simulações.

No entanto, constataram-se diferenças em locos que teriam, em teoria, maiores taxas de similaridade para uma mistura de haplogrupos AMR: DYS437 (63,40%), DYS390 (55,56%), DYS392 (52,94%) e DYS391 (50,88%). Também se verificou alelos idênticos para locos para os quais se esperaria maiores taxas de divergência (variando de 52,63% a 81,29%): DYS385, DYS576, DYS549, DYS533, DYS635 e DYS439. Contudo, uma vez que estas taxas se encontram próximas ao valor de 50%, não poderia haver uma expectativa significativa sobre estes marcadores que, neste caso, seriam menos informativos.

A análise genética baseada em simulações, como no presente caso, fornece estimativas aproximadas, indicando a viabilidade da proposta de análise para inferência de haplogrupos. Entretanto, sua aplicação na prática requer estudos mais aprofundados que possam validá-la efetivamente.

7. CONSIDERAÇÕES FINAIS

7.1 Uso do *STRUCTURE* para inferência de ancestralidade pelos Y-STRs

O software *STRUCTURE* pode ser eficientemente utilizado para inferir os haplogrupos EUR, AFR e AMR principais da amostra populacional estudada, apresentando bom desempenho quando comparado a ferramentas conhecidas para esta finalidade.

As três ferramentas testadas, porém, apresentaram erros de atribuição. Tais equívocos em programas dessa natureza podem decorrer, entre outros fatores, do algoritmo específico de cada um. Dado que nosso foco é o estudo da aplicabilidade do *STRUCTURE*, direcionamos nossa análise para os erros identificados por essa ferramenta, buscando compreender as possíveis causas que poderiam contribuir para essas discrepâncias. Estas podem ser devidas a similaridades de haplótipos Y-STRs em haplogrupos distintos e à convergência (WANG *et al.*, 2015) ou ao próprio algoritmo do programa. Não foi possível identificar com precisão uma variável que justificasse as inferências incorretas realizadas pelo *STRUCTURE* e pelos demais *softwares*.

Independentemente do motivo pelos quais ocorreram, previsões incorretas podem induzir a inferências erradas do haplogrupo de um indivíduo. O principal desafio reside na dificuldade de detectar esses erros de inferência, especialmente quando não se dispõe de informações anteriores. Em suma, um erro de atribuição, não sendo detectável, se torna um problema.

Logo, o presente estudo corroborou que a associação de dois programas pode incrementar a segurança do resultado obtido, como já sugerido em publicações anteriores. Na prática, a ausência de compatibilidade entre dois métodos serviria como o maior indicativo da existência de erro de atribuição. Neste caso, associar um terceiro método (no caso, o *STRUCTURE*), segundo os dados aqui apresentados, possibilitaria um resultado conclusivo em, no mínimo, 40% das ocasiões nas quais os outros programas apresentaram resultados discrepantes; em 64,71% dos casos em que um dos outros programas não gerou

resultados; e em 100% dos casos em que nenhum dos outros programas gerou resultados (**Tabela 17**).

A associação entre diferentes métodos para classificar indivíduos em haplogrupos já foi proposta e justifica-se porque cada algoritmo teria um método próprio para efetuar a classificação a partir dos dados referência ou de treinamento. Assim como se observou no presente trabalho, e segundo a literatura na área, os programas não cometem exatamente o mesmo tipo de erro. Logo, uma combinação de resultados obtidos pela combinação dos programas poderia conferir maior robustez à conclusão (EMMEROVA *et al.*, 2017; JORDAMOVIC´ *et al.*, 2021; SCHLECHT *et al.*, 2008). A título de exemplo, o indivíduo de haplogrupo desconhecido foi alocado, pelos três programas, ao mesmo haplogrupo, R1b. Seguindo a lógica desenvolvida, a inferência efetuada somente pelo programa *STRUCTURE* (ou pelos outros programas isoladamente) a este haplogrupo poderia estar certa ou errada. Diante da dúvida, a alocação pode ser confirmada pelos dois outros programas, gerando um resultado conclusivo.

Considerando que *HAPEST* e *NevGen* são duas ferramentas já consolidadas, de fácil acesso, de uso bastante simples e intuitivo, e cujo resultado é enviado instantaneamente, estas poderiam ser consideradas boas ferramentas de primeira opção. O programa *STRUCTURE*, embora mais flexível para se adaptar a diversas demandas e apresentações, é mais complexo e requer acesso a dados das populações de referência e análises mais trabalhosas. Contudo, estudos têm sugerido que o programa seja adequado para determinar o haplogrupo de indivíduos de populações recentemente miscigenadas, considerando as contribuições das possíveis populações parentais (PORRAS-HURTADO *et al.*, 2013). Assim, o *STRUCTURE* mostrou-se apropriado como opção de terceiro programa a funcionar como desempate, nos casos de resultados discrepantes obtidos pelos demais programas. De toda forma, as três ferramentas se mostraram efetivas para inferência de haplogrupo na população miscigenada brasileira.

No entanto, reforça-se a recomendação para que a inferência de haplogrupos através de Y-STRs seja realizada e interpretada com cautela, uma

vez que o método mais eficaz e consolidado permanece sendo a genotipagem de Y-SNPs característicos que permitem a classificação em cada haplogrupo (JANNUZZI *et al.*, 2020; PORRAS-HURTADO *et al.*, 2013; SCHLECHT *et al.*, 2008). Adicionalmente, ao considerar populações brasileiras, que são notavelmente miscigenadas e cujas características fenotípicas não necessariamente refletem o padrão esperado para um haplogrupo europeu, africano ou indígena, essa informação não deve ser tida como suficientemente precisa para determinar a origem biogeográfica com exatidão.

7.2 Comportamento de misturas em termos de diversidade e sua possível aplicação

O conhecimento da diversidade nas diferentes regiões brasileiras, que diferem em virtude de razões histórico-demográficas, pode possibilitar a detecção de agressores que não pertençam a um determinado grupo populacional esperado. Em outras palavras, conhecendo-se o perfil de diversidade da população brasileira, um agressor cujo perfil genético indique um haplogrupo diferente do habitual pode proporcionar um direcionamento investigativo para um grupo populacional específico.

De maneira geral, foi possível constatar que, dentro do mesmo haplogrupo, os marcadores se mostraram menos diferentes do que quando foram combinados indivíduos de haplogrupos diferentes, conforme esperado, uma vez que a diversidade genética dos grupos populacionais é consequência de suas origens geográficas (COURT, 2021).

Misturas envolvendo dois indivíduos de populações AMR apresentaram as menores médias de diferenças (58,68%), uma vez que constituem populações historicamente mais isoladas (PETZL-ERLER; LUZ; SOTOMAIOR, 1993; TSUNETO *et al.*, 2003). É importante notar que nossa amostra da população indígena foi caracterizada por apenas um haplogrupo (Q-M3), que é o mais prevalente entre os indígenas brasileiros. Dado que essas populações também apresentam isolamento entre si, estudos complementares poderiam

explorar como seria o comportamento de misturas semelhantes, incluindo grupos indígenas mais distintos.

Por outro lado, as misturas envolvendo indivíduos de populações AFR apresentaram baixa média de diferenças (61,24%) quando comparada à observada entre populações EUR (66,91%), um resultado oposto ao esperado e possivelmente influenciado pelo reduzido número de haplótipos AFR (n=11) na amostra populacional aqui estudada. Da mesma forma, a diversidade observada entre indivíduos de populações EUR provavelmente se mostrou elevada em virtude de sua predominância (n=148) na amostra.

A partir dos estudos conduzidos com base no padrão de diferenças dos marcadores em cada cenário, acreditamos ser possível investigar se o nível de diversidade pode trazer informações sugestivas da composição dos haplótipos dos contribuintes da mistura.

Propusemos uma série de etapas que poderiam ser seguidas, para investigar os haplogrupos em misturas. No entanto, até o momento, estas etapas limitam-se a fornecer vagas sugestões de inferência, pois os resultados aqui relatados são insuficientes para permitir conclusões efetivas. Entretanto, sugerem ser possível que os Y-STRs apresentem tendências ou padrões de comportamento frente aos diferentes cenários de mistura. Admite-se, no entanto, que tais padrões não sejam absolutos, uma vez que a presente amostra populacional não compreende toda a diversidade haplotípica de um mesmo haplogrupo. Dessa forma, entende-se que a dedução dos haplogrupos dos contribuintes da mistura deveria ser, fundamentalmente, realizada observando-se todo o conjunto de marcadores Y-STR.

A realização concreta da inferência dos haplogrupos em contextos de mistura é uma abordagem preliminar que demanda investigações mais minuciosas, com amostras populacionais mais numerosas, refinamento das técnicas utilizadas e aprimoramento da metodologia. No entanto, essa abordagem poderia trazer um valor adicional à interpretação de misturas, desempenhando um possível papel, cuja viabilidade poderia ser explorada.

8. CONCLUSÕES

Este estudo permitiu as seguintes conclusões:

1) Todas as populações apresentaram ampla diversidade haplotípica de STR, representada por 97,75% de haplótipos diferentes pelo sistema PPY23.
2) A eficiência forense do sistema PPY23 com base nos parâmetros forenses (diversidade haplotípica e capacidade de discriminação) nos indígenas e miscigenados é alta e confirma a potencial aplicação em exames de parentescos e identidade.
3) A performance de atribuição de clados dos indivíduos com base em haplótipos de Y-STRs feita pelo *STRUCTURE* mostrou-se viável, quando comparada com outras ferramentas. Entretanto, é aconselhável que a ferramenta seja utilizada em conjunto com outros preditores, como uma medida adicional de confirmação.
4) A proposta de inferência dos haplogrupos dos contribuintes de misturas genéticas a partir da diversidade de haplogrupos de Y-STRs mostrou-se promissora, mas carece de confirmação em amostras mais numerosas.
5) Para uma investigação mais assertiva e robusta tanto do uso do *STRUCTURE* para inferência de haplogrupos a partir de Y-STRs quanto para análise de misturas genéticas é imprescindível o aumento do número amostral analisado.

REFERÊNCIAS

ALI, Sher; HASNAIN, Seyed Ehtesham. Molecular dissection of the human Y-chromosome. **Gene**, [*s. l.*], v. 283, n. 1–2, p. 1–10, 2002.

ALVES, Hemerson Bertassoni. **Caracterização da variabilidade genética de uma amostra da população do estado do Paraná utilizando marcadores STRs autossômicos e do cromossomo Y**. 2012. 58 f. [*s. l.*], 2012.

ASHRAF A EWIS, JUWON LEE, TOSHIKATSU SHINKA, Yutaka Nakahori. Two Y-chromosome-specific polymorphisms 12f2 and DFFRY in the Japanese population and their relations to other Y-polymorphisms. **The Journal of Medical Investigation**, [*s. l.*], v. 49, p. 11–50, 2002.

ATHEY, T Whit. Haplogroup Prediction from Y-STR Values Using a Bayesian-Allele- Frequency Haplogroup Prediction from Y-STR Values Using a Bayesian-Allele- Frequency Approach. **Journal of Genetic Genealogy**, [*s. l.*], v. 2, n. January 2006, p. 1–5, 2006.

ATHEY, T Whit. Haplogroup Prediction From Y-Str Values Using An Allele Frequency Approach. [*s. l.*], v. 1, n. 1, p. 0–7, 2005.

AUGUSTO, D. G. *et al.* Unsuspected Associations of Variants within the Genes NOTCH4 and STEAP2-AS1 Uncovered by a GWAS in Endemic Pemphigus Foliaceus. **Journal of Investigative Dermatology**, [*s. l.*], v. 141(11), p. 2741–2744, 2021.

AUTON, Adam *et al.* A global reference for human genetic variation. **Nature**, [*s. l.*], v. 526, n. 7571, p. 68–74, 2015.

BALLANTYNE, Kaye N *et al.* Forensic Science International : Genetics A new future of forensic Y-chromosome analysis : Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. **Forensic Science International: Genetics**, [*s. l.*], v. 6, n. 2, p. 208–218, 2012. Disponível em: http://dx.doi.org/10.1016/j.fsigen.2011.04.017.

BALLARD, D J *et al.* The beneficial effect of extending the Y chromosome STR haplotype. [*s. l.*], v. 1288, p. 151–153, 2006.

BARCELOS, Rejane da Silva Sena. Contribuição Genética de Duas Populações

Urbanas da Região Centro-Oeste Brasileira Estimada por Marcadores Uniparentais. 2006. - Universidade de Brasília - UnB, [*s. l.*], 2006.

BATINI, Chiara; JOBLING, Mark A. Detecting past male-mediated expansions using the Y chromosome. **Human Genetics**, [*s. l.*], v. 136, n. 5, p. 547–557, 2017.

BELEZA, Sandra; LOPES, Alexandra; CARRACEDO, Angel. Grouping of Y-STR haplotypes discloses European geographic clines. [*s. l.*], v. 134, p. 172–179, 2003.

BERGER, Burkhard *et al.* Reprint of: High resolution mapping of y haplogroup G in Tyrol (Austria). **Forensic Science International: Genetics**, [*s. l.*], v. 7, n. 6, p. 624–631, 2013.

BERGSTROM, Anders *et al.* Insights into human genetic variation and population history from 929 diverse genomes. **Science**, [*s. l.*], v. 367, n. 6484, 2020.

BORTOLINI, Maria-catira *et al.* Y-Chromosome Evidence for Differing Ancient Demographic Histories in the Americas. [*s. l.*], p. 524–539, 2003.

BRASIL. **Código de Processo Penal. Decreto lei nº 3.689, de 03 de outubro de 1941.** 1941.

BUTLER, John M. **Forensic DNA Typing - Biology, Technology, and Genetics of STR Markers**. 2nd. ed. [*S. l.*]: Elsevier, 2005.

BUTLER, J M *et al.* Short communication DNA Commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis §. [*s. l.*], v. 157, p. 187–197, 2006.

CALLEGARI-JACQUES, Sidia M *et al.* Historical Genetics : Spatiotemporal Analysis of the Formation of the Brazilian Population. [*s. l.*], v. 834, n. May, p. 824–834, 2003.

CANALES SERRANO, Aurora. Forensic DNA phenotyping: A promising tool to aid forensic investigation. Current situation. **Revista Espanola de Medicina Legal**, [*s. l.*], v. 46, n. 4, p. 183–190, 2020.

CANN, H.M. A Human Genome Diversity Cell Line Panel. **Science**, [*s. l.*], v. 296(5566), p. 261–262, 2002.

CAROLINO, Santos *et al.* A systematic literature review on the European , African and Amerindian genetic ancestry components on Brazilian health outcomes. [*s. l.*], n. May, p. 1–11, 2019.

CARVALHO, Mónica *et al.* Analysis of paternal lineages in Brazilian and African populations. **Genetics and Molecular Biology**, [*s. l.*], v. 33, n. 3, p. 422–427, 2010.

CAVALLI-SFORZA, L.L.; PIAZZA, A. Human Genomic Diversity in Europe: A Summary of Recent Research and Prospects for the Future. **Eur J Hum Genet**, [*s. l.*], v. 1, p. 3–18, 1993.

CETKOVIC GENTULA, A.; NEVSKI, M. **NEVGEN Y-DNA Haplogroup Predictor**. [*S. l.: s. n.*], 2015. Disponível em: http://www.nevgen.org/.

CHANG, Christopher C *et al.* Second-generation PLINK : rising to the challenge of larger and richer datasets. [*s. l.*], p. 1–16, 2015.

CHEN, Hao *et al.* Y - LineageTracker : a high - throughput analysis framework for Y - chromosomal next - generation sequencing data. **BMC Bioinformatics**, [*s. l.*], p. 1–15, 2021. Disponível em: https://doi.org/10.1186/s12859-021-04057-z.

CHEVITARESE, Juliana. Determinação da Estrutura Genética das populações Humanas e Interferência dos Fatores Evolutivos que Contribuíram Para a Sua Formação. **Universidade Federal de Minas Gerais**, [*s. l.*], p. 2–99, 2009.

CHUNG, Yuk-ka; FUNG, Wing K; HU, Yue-qing. Familial database search on two-person mixture. **Computational Statistics and Data Analysis**, [*s. l.*], v. 54, n. 8, p. 2046–2051, 2010. Disponível em: http://dx.doi.org/10.1016/j.csda.2010.03.002.

COURT, Denise Syndercombe. The y chromosome and its use in forensic dna analysis. **Emerging Topics in Life Sciences**, [*s. l.*], v. 5, n. 3, p. 427–441, 2021.

DE KNIJFF, P. Messages through bottlenecks: On the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. **American Journal of Human Genetics**, [*s. l.*], v. 67, n. 5, p. 1055–1061, 2000.

DØRUM, Guro; KAUR, Navreet; GYSI, Mario. Pedigree-based relationship inference from complex DNA mixtures. [*s. l.*], 2017.

EARL, Dent A.; VONHOLDT, Bridgett M. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. **Conservation Genetics Resources**, [*s. l.*], v. 4, n. 2, p. 359–361, 2012.

ELKAMEL, Sarra *et al.* Insights into the Middle Eastern paternal genetic pool in Tunisia: high prevalence of T-M70 haplogroup in an Arab population. **Scientific Reports**, [*s. l.*], v. 11, n. 1, p. 1–12, 2021. Disponível em: https://doi.org/10.1038/s41598-021-95144-x.

EMMEROVA, Barbora *et al.* Forensic Science International : Genetics Supplement Series Comparison of Y-chromosomal haplogroup predictors. **Forensic Science International: Genetics Supplement Series**, [*s. l.*], v. 6, n. September, p. e145–e147, 2017. Disponível em: https://doi.org/10.1016/j.fsigss.2017.09.025.

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. **Molecular Ecology**, [*s. l.*], v. 14, n. 8, p. 2611–2620, 2005.

EXCOFFIER, Laurent; LAVAL, Guillaume; SCHNEIDER, Stefan. Arlequin (version 3.0): An integrated software package for population genetics data analysis. **Evolutionary Bioinformatics**, [*s. l.*], v. 1, p. 117693430500100, 2005.

FEHÉR, T. *et al.* Y-SNP L1034: limited genetic link between Mansi and Hungarian-speaking populations. **Molecular Genetics and Genomics**, [*s. l.*], v. 290, n. 1, p. 377–386, 2015.

FITZPATRICK, Colleen. **The emerging discipline of forensic genetic genealogy**. [*S. l.*]: Elsevier Inc., 2022. *E-book*. Disponível em: https://doi.org/10.1016/B978-0-12-815766-4.00022-4.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA. **Anuário Brasileiro de Segurança Pública**. [*S. l.: s. n.*], 2023. Disponível em: https://forumseguranca.org.br/wp-content/uploads/2023/07/anuario-2023.pdf. .

GENTILE, Fabiano *et al.* Early evaluation of five age-correlated DNA methylation markers in an Italian population sample. **Forensic Science International: Genetics Supplement Series**, [*s. l.*], v. 7, n. 1, p. 424–426, 2019. Disponível em: https://doi.org/10.1016/j.fsigss.2019.10.037.

GRATTAPAGLIA, Dario *et al.* Y-chromosome STR haplotype diversity in Brazilian populations. **Forensic Science International**, [*s. l.*], v. 149, n. 1, p. 99–107, 2005.

GRAVERSEN, Therese; MORTERA, Julia; LAGO, Giampietro. The Yara Gambirasio case: Combining evidence in a complex DNA mixture case. **Forensic Science International: Genetics**, [*s. l.*], v. 40, p. 52–63, 2019. Disponível em: http://dx.doi.org/10.1016/j.fsigen.2018.12.010.

GREEN, Peter J; MORTERA, Julia. Ac ce pt e d us cr ip t. **Forensic Science International: Genetics**, [*s. l.*], 2017. Disponível em: http://dx.doi.org/10.1016/j.fsigen.2017.02.001.

GRUGNI, Viola *et al.* Analysis of the human Y-chromosome haplogroup Q characterizes ancient population movements in Eurasia and the Americas. [*s. l.*], p. 1–14, 2019.

GUSMÃO, L. *et al.* DNA Commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis. **Forensic Science International**, [*s. l.*], v. 157, n. 2–3, p. 187–197, 2006.

GUSMÃO, Leonor *et al.* Revised guidelines for the publication of genetic population data. **Forensic Science International: Genetics**, [*s. l.*], v. 30, p. 160–163, 2017.

HALLAST, Pille *et al.* A Southeast Asian origin for present-day non-African human Y chromosomes. **Human Genetics**, [*s. l.*], v. 140, n. 2, p. 299–307, 2021. Disponível em: https://doi.org/10.1007/s00439-020-02204-9.

HALLAST, Pille *et al.* The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. **Molecular Biology and Evolution**, [*s. l.*], v. 32, n. 3, p. 661–673, 2014.

HAMCZYK, Magda R. *et al.* Biological Versus Chronological Aging: JACC Focus Seminar. **Journal of the American College of Cardiology**, [*s. l.*], v. 75, n. 8, p. 919–930, 2020.

HAMMER, Michael F. *et al.* Extended y chromosome haplotypes resolve multiple

and unique lineages of the Jewish priesthood. **Human Genetics**, [*s. l.*], v. 126, n. 5, p. 707–717, 2009.

HEIDEGGER, A. *et al.* Development and inter-laboratory validation of the VISAGE enhanced tool for age estimation from semen using quantitative DNA methylation analysis. **Forensic Science International: Genetics**, [*s. l.*], v. 56, 2022.

HUANG, Yun Zhi *et al.* Dispersals of the Siberian Y-chromosome haplogroup Q in Eurasia. **Molecular Genetics and Genomics**, [*s. l.*], v. 293, n. 1, p. 107–117, 2018.

IBGE. **Censo Brasileiro de 2022Instituto Brasileiro de Geografia e Estatística**. Rio de Janeiro: [*s. n.*], 2023. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/22827-censo-demografico-2022.html?edicao=37417&t=resultados. .

ILUMÄE, Anne Mai *et al.* Human Y Chromosome Haplogroup N: A Non-trivial Time-Resolved Phylogeography that Cuts across Language Families. **American Journal of Human Genetics**, [*s. l.*], v. 99, n. 1, p. 163–173, 2016.

INSTITUTO DE PESQUISA ECONÔMICA APLICADA - IPEA. [*S. l.*], 2023. Disponível em: https://www.ipea.gov.br/portal. Acesso em: 16 ago. 2023.

INTERNATIONAL SOCIETY OF GENETIC GENEALOGY. Y-DNA HAPLOGROUP TREE 2019, VERSION 15.73 - 11 JULY 2020. [*S. l.*], [*s. d.*]. Disponível em: http://www.isogg.org/tree/. Acesso em: 7 ago. 2023.

JANNUZZI, Juliana *et al.* Male lineages in Brazilian populations and performance of haplogroup prediction tools. **Forensic Science International: Genetics**, [*s. l.*], v. 44, n. August 2019, p. 102163, 2020. Disponível em: https://doi.org/10.1016/j.fsigen.2019.102163.

JOBLING, Mark A. Forensic genetics through the lens of Lewontin: Population structure, ancestry and race. **Philosophical Transactions of the Royal Society B: Biological Sciences**, [*s. l.*], v. 377, n. 1852, 2022.

JOBLING, Mark A. The impact of recent events on human genetic diversity. **Philosophical Transactions of the Royal Society B: Biological Sciences**, [*s.*

l.], v. 367, n. 1590, p. 793–799, 2012.

JOERIN, Iriel A *et al.* Ancestry , diversity , and genetics of health - related traits in African - derived communities (quilombos) from Brazil. **Functional & Integrative Genomics**, [*s. l.*], v. 23, n. 1, p. 1–14, 2023. Disponível em: https://doi.org/10.1007/s10142-023-00999-0.

JOERIN, Iriel A *et al.* Uniparental markers reveal new insights on subcontinental ancestry and sex - biased admixture in Brazil. **Molecular Genetics and Genomics**, [*s. l.*], 2022. Disponível em: https://doi.org/10.1007/s00438-022-01857-7.

JORDAMOVIC´, Naida Babic *et al.* Haplogroup Prediction Using Y-Chromosomal Short Tandem Repeats in the General Population of Bosnia and Herzegovina. [*s. l.*], v. 12, n. June, p. 1–6, 2021.

KARAFET, Tatiana M. *et al.* New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. **Genome Research**, [*s. l.*], v. 18, n. 5, p. 830–838, 2008.

KAUR, Navreet *et al.* Relationship inference based on DNA mixtures. **International Journal of Legal Medicine**, [*s. l.*], 2015. Disponível em: http://dx.doi.org/10.1007/s00414-015-1276-1.

KAYSER, Manfred *et al.* Forensic Science International : Genetics Recent advances in Forensic DNA Phenotyping of appearance , ancestry and age ☆. **Forensic Science International: Genetics**, [*s. l.*], v. 65, n. April, p. 102870, 2023. Disponível em: https://doi.org/10.1016/j.fsigen.2023.102870.

KAYSER, Manfred. Forensic use of Y - chromosome DNA : a general overview. **Human Genetics**, [*s. l.*], v. 136, n. 5, p. 621–635, 2017.

KIDD, Judith R *et al.* Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. [*s. l.*], p. 1–13, 2011.

KIMURA, Motoo. Stepwise mutation model and distribution of allelic frequencies in a. [*s. l.*], v. 75, n. 6, p. 2868–2872, 1978.

KING, Roy J *et al.* ARTICLE A major Y-chromosome haplogroup R1b Holocene

era founder effect in Central and Western Europe. [*s. l.*], n. March 2010, p. 95–101, 2011.

KNIJFF, Peter De. On the Forensic Use of Y-Chromosome Polymorphisms. [*s. l.*], 2022.

KOPELMAN, Naama M *et al.* CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across K. **Mol Ecol Resour**, [*s. l.*], v. 15, n. 5, p. 1179–1191, 2015.

LEITE, Fabio P.N. *et al.* Y-STR analysis in Brazilian and South Amerindian populations. **American Journal of Human Biology**, [*s. l.*], v. 20, n. 3, p. 359–363, 2008.

MEDINA, Laura S.Jurado *et al.* Continental origin for Q haplogroup patrilineages in Argentina and Paraguay. **Human Biology**, [*s. l.*], v. 92, n. 2, p. 63–80, 2020.

MELTON, Phillip. **Genetic history and pre-Columbian Diaspora of Chibchan speaking populations: Molecular genetic evidence**. 2008. [*s. l.*], 2008.

MENDEZ, Fernando L. *et al.* An African American paternal lineage adds an extremely ancient root to the human y chromosome phylogenetic tree. **American Journal of Human Genetics**, [*s. l.*], v. 92, n. 3, p. 454–459, 2013. Disponível em: http://dx.doi.org/10.1016/j.ajhg.2013.02.002.

MEYER, Mikkel *et al.* Forensic Science International : Genetics Identifying the most likely contributors to a Y-STR mixture using the discrete Laplace method. **Forensic Science International: Genetics**, [*s. l.*], v. 15, p. 76–83, 2015. Disponível em: http://dx.doi.org/10.1016/j.fsigen.2014.09.011.

MIZUNO, Natsuko *et al.* A forensic method for the simultaneous analysis of biallelic markers identifying Y chromosome haplogroups inferred as having originated in Asia and the Japanese archipelago. **Forensic Science International: Genetics**, [*s. l.*], v. 4, n. 2, p. 73–79, 2010.

MONTESANTO, Alberto *et al.* A New Robust Epigenetic Model for Forensic Age Prediction. **Journal of Forensic Sciences**, [*s. l.*], v. 65, n. 5, p. 1424–1431, 2020.

MORAN, Colin N *et al.* Y chromosome haplogroups of elite Ethiopian endurance runners. [*s. l.*], p. 492–497, 2004.

MORTERA, Julia. DNA Mixtures in Forensic Investigations : The Statistical State of the Art. [*s. l.*], 2020.

MOUSSA, N M *et al.* Journal of Archaeological Science : Reports Y-chromosomal DNA analyzed for four prehistoric cemeteries from Cis-Baikal , Siberia. **Journal of Archaeological Science: Reports**, [*s. l.*], v. 17, p. 932–942, 2018. Disponível em: http://dx.doi.org/10.1016/j.jasrep.2016.11.003.

MUZZIO, Marina *et al.* Software for Y-haplogroup predictions: A word of caution. **International Journal of Legal Medicine**, [*s. l.*], v. 125, n. 1, p. 143–147, 2011.

NAIDOO, Thijessen *et al.* Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. **Investigative Genetics**, [*s. l.*], v. 1, n. 1, p. 1–11, 2010.

NASCIMENTO, Nicole *et al.* Genetic data and de novo mutation rates in father-son pairs of 23 Y-STR loci in Southern Brazil population. [*s. l.*], p. 389–391, 2014.

NEBEL, Almut *et al.* Haplogroup-specific deviation from the stepwise mutation model at the microsatellite loci DYS388 and. [*s. l.*], p. 22–26, 2001.

NEI, M.; TAJIMA, F. DNA polymorphism detectable by restriction endonucleases. **Genetics1**, [*s. l.*], v. 97, p. 145–163, 1981.

NIEDERSTÄTTER, Harald *et al.* Separate analysis of DYS385a and b versus conventional DYS385 typing : is there forensic relevance ?. [*s. l.*], p. 1–9, 2004.

NÚÑEZ, Carolina *et al.* Y chromosome haplogroup diversity in a Mestizo population of Nicaragua. **Forensic Science International: Genetics**, [*s. l.*], v. 6, n. 6, p. 192–195, 2012.

PALHA, Teresinha *et al.* Disclosing the genetic structure of Brazil through analysis of male lineages with highly discriminating haplotypes. **PLoS ONE**, [*s. l.*], v. 7, n. 7, p. 1–8, 2012.

PENA, S. D.J. *et al.* DNA tests probe the genomic ancestry of Brazilians. **Brazilian Journal of Medical and Biological Research**, [*s. l.*], v. 42, n. 10, p. 870–876, 2009.

PENA, Sérgio D.J.; BORTOLINI, Maria Cátira. Pode a genética definir quem deve

se beneficiar das cotas universitárias e demais ações afirmativas ?. **Estudos Avançados**, [*s. l.*], v. 18, n. 50, p. 31–50, 2004.

PENA, Sergio D J; SANTOS, Fabrício R; TARAZONA-SANTOS, Eduardo. Genetic admixture in Brazil. [*s. l.*], n. October, p. 1–11, 2020.

PEREIRA, Luisa *et al.* African genetic diversity and adaptation inform a precision medicine agenda. **Nature Reviews Genetics**, [*s. l.*], v. 22, n. May, 2021. Disponível em: http://dx.doi.org/10.1038/s41576-020-00306-8.

PEREIRA, Vania *et al.* Evaluation of the Precision of Ancestry Inferences in South American Admixed Populations. [*s. l.*], v. 11, n. August, p. 1–20, 2020.

PETREJČÍKOVÁ, Eva *et al.* Y-SNP analysis versus Y-Haplogroup Predictor in the Slovak population. **Anthropologischer Anzeiger**, [*s. l.*], v. 71, n. 3, p. 275–285, 2014.

PETZL-ERLER, Maria L.; LUZ, Roberto; SOTOMAIOR, Vanessa Santos. The HLA polymorphsm of two distinctive South-American Indian tribes: The Kaingang and the Guarani. **Tissue Antigens**, [*s. l.*], v. 41, n. 5, p. 227–237, 1993.

PHILLIPS, Chris. Forensic genetic analysis of bio-geographical ancestry. **Forensic Science International: Genetics**, [*s. l.*], v. 18, p. 49–65, 2015. Disponível em: http://dx.doi.org/10.1016/j.fsigen.2015.05.012.

PORRAS-HURTADO, Liliana *et al.* An overview of STRUCTURE: Applications, parameter settings, and supporting software. **Frontiers in Genetics**, [*s. l.*], v. 4, n. MAY, p. 1–13, 2013.

POZNIK, G David. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. [*s. l.*], p. 1–5, 2016.

PRITCHARD, Jonathan K.; STEPHENS, Matthew; DONNELLY, Peter. Inference of population structure using multilocus genotype data. **Genetics**, [*s. l.*], v. 155, n. 2, p. 945–959, 2000. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461096/pdf/10835412.pdf.

PURPS, Josephine *et al.* A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. **Forensic Science International: Genetics**, [*s. l.*], v. 12, p. 12–23, 2014.

QUINTANA-MURCI, Âs; KRAUSZ, Csilla; MCELREAVEY, Ken. The human Y chromosome : function , evolution and disease. [*s. l.*], v. 118, 2001.

R CORE TEAM. **R: A language and environment for statistical computing. R Foundation for Statistical Computing**. [*S. l.: s. n.*], 2020. Disponível em: https://www.r-project.org/.

RESQUE, Rafael *et al.* Male lineages in Brazil: Intercontinental admixture and stratification of the European background. **PLoS ONE**, [*s. l.*], v. 11, n. 4, p. 1–17, 2016.

RODRIGUES DE MOURA, Ronald *et al.* Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries. **American Journal of Human Biology**, [*s. l.*], v. 27, n. 5, p. 674–680, 2015.

ROHRLACH, Adam B. *et al.* Using Y-chromosome capture enrichment to resolve haplogroup H2 shows new evidence for a two-path Neolithic expansion to Western Europe. **Scientific Reports**, [*s. l.*], v. 11, n. 1, p. 1–11, 2021. Disponível em: https://doi.org/10.1038/s41598-021-94491-z.

ROOTSI, Siiri *et al.* A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. **European Journal of Human Genetics**, [*s. l.*], v. 15, n. 2, p. 204–211, 2007.

ROOTSI, Siiri *et al.* Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. **American Journal of Human Genetics**, [*s. l.*], v. 75, n. 1, p. 128–137, 2004.

ROSENBERG, Noah A *et al.* Clines , Clusters , and the Effect of Study Design on the Inference of Human Population Structure. [*s. l.*], v. 1, n. 6, 2005.

SAHAKYAN, Hovhannes *et al.* Origin and diffusion of human Y. **Scientific Reports**, [*s. l.*], p. 1–14, 2021. Disponível em: https://doi.org/10.1038/s41598-021-85883-2.

SALZANO, Francisco Mauro; SANS, Mónica. Interethnic admixture and the evolution of Latin American populations. [*s. l.*], v. 1, p. 151–170, 2014.

SCHLECHT, Joseph *et al.* Machine-learning approaches for classifying haplogroup from Y chromosome STR data. **PLoS Computational Biology**, [*s.*

l.], v. 4, n. 6, 2008.

SCHNEIDER, Peter M.; PRAINSACK, Barbara; KAYSER, Manfred. The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry. **Deutsches Arzteblatt International**, [*s. l.*], v. 116, n. 51–52, p. 873–880, 2019.

SCOZZARI, Rosaria *et al.* Molecular Dissection of the Basal Clades in the Human Y Chromosome Phylogenetic Tree. **PLoS ONE**, [*s. l.*], v. 7, n. 11, 2012.

SHABALALA, Sthabile; GUAI, Meenu; OKPEKU, Moses. Y - STR Kits and Y - STR Diversity in the South African Population :. **Journal of Forensic Science and Medicine**, [*s. l.*], v. 8, p. 104–113, 2022.

SHARMA, Dhruw; SK, Shukla. Importance of Y- STR profiling in sexual assault cases with mixed DNA profile. [*s. l.*], v. 3, n. 1, p. 42–45, 2018.

SOLÉ-MORATA, Neus *et al.* Whole Y-chromosome sequences reveal an extremely recent origin of the most common North African paternal lineage E-M183 (M81). [*s. l.*], v. 1, n. May, p. 1–11, 2017.

SOUZA, Aracele Maria De *et al.* A systematic scoping review of the genetic ancestry of the Brazilian population. **Genetics and Molecular Biology**, [*s. l.*], v. 508, p. 495–508, 2019.

TARAZONA-SANTOS, Eduardo *et al.* Genetic Differentiation in South Amerindians Is Related to Environmental and Cultural Diversity : Evidence from the Y Chromosome. **American Journal of Human Genetics**, [*s. l.*], v. 68, p. 1485–1496, 2001.

THOMAS, Mark G *et al.* Y Chromosomes Traveling South : The Cohen Modal Haplotype and the Origins of the Lemba — the “ Black Jews of Southern Africa ”. [*s. l.*], p. 674–686, 2000.

TROMBETTA, Beniamino *et al.* A new topology of the human Y chromosome haplogroup E1b1 (E-P2) revealed through the use of newly characterized binary polymorphisms. **PLoS ONE**, [*s. l.*], v. 6, n. 1, p. 6–9, 2011.

TROMBETTA, Beniamino *et al.* Phylogeographic Refinement and Large Scale Genotyping of Human Y Chromosome Haplogroup E Provide New Insights into the Dispersal of Early Pastoralists in the African Continent. **Genome Biology**

and Evolution, [*s. l.*], v. 7, n. 7, p. 1940–1950, 2015.

TSUNETO, L. T. *et al.* HLA class II diversity in seven Amerindian populations. Clues about the origins of the Aché. **Tissue Antigens**, [*s. l.*], v. 62, n. 6, p. 512–526, 2003.

UNDERHILL, Peter A. *et al.* The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. **European Journal of Human Genetics**, [*s. l.*], v. 23, n. 1, p. 124–131, 2015.

UNDERHILL, P.A. *et al.* The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. **Ann. Hum. Genet.**, [*s. l.*], v. 65, p. 43–62, 2001.

UNDERHILL, Peter A.; KIVISILD, Toomas. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. **Annual Review of Genetics**, [*s. l.*], v. 41, p. 539–564, 2007.

VAN OVEN, Mannis *et al.* Seeing the wood for the trees: A minimal reference phylogeny for the human Y chromosome. **Human Mutation**, [*s. l.*], v. 35, n. 2, p. 187–191, 2014.

WANG, Chuan Chao *et al.* Convergence of Y Chromosome STR Haplotypes from Different SNP Haplogroups Compromises Accuracy of Haplogroup Prediction. **Journal of Genetics and Genomics**, [*s. l.*], v. 42, n. 7, p. 403–407, 2015. Disponível em: http://dx.doi.org/10.1016/j.jgg.2015.03.008.

WANG, Sijia *et al.* Genetic Variation and Population Structure in Native Americans. [*s. l.*], v. 3, n. 11, 2007.

WATAHIKI, Haruhiko *et al.* Polymorphisms and microvariant sequences in the Japanese population for 25 Y-STR markers and their relationships to Y-chromosome haplogroups. **Forensic Science International: Genetics**, [*s. l.*], 2019. Disponível em: http://dx.doi.org/10.1016/j.fsigen.2019.03.004.

WILSON, James F *et al.* Genetic evidence for different male and female roles during cultural transitions in the British Isles. [*s. l.*], v. 98, n. 9, 2001.

WOŹNIAK, Anna *et al.* Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones. **Aging**, [*s.*

l.], v. 13, n. 5, p. 6459–6484, 2021.

XU, Hongyang *et al.* Inferring population structure and demographic history using Y-STR data from worldwide populations. **Molecular Genetics and Genomics**, [*s. l.*], v. 290, n. 1, p. 141–150, 2015.

YANG, Ning Ning *et al.* Contrasting Patterns of Nuclear and mtDNA Diversity in Native American Populations. **Annals of Human Genetics**, [*s. l.*], v. 74, n. 6, p. 525–538, 2010.

YANG, Yaran *et al.* Haplotypic polymorphisms and mutation rate estimates of 22 Y-chromosome STRs in the Northern Chinese Han father – son pairs. [*s. l.*], n. April, p. 6–11, 2018.

ZHONG, Hua *et al.* Extended y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route. **Molecular Biology and Evolution**, [*s. l.*], v. 28, n. 1, p. 717–727, 2011.

APÊNDICES

APÊNDICE 1 - Haplótipos encontrados para a população miscigenada brasileira classificada como europeia (n = 148).

HAPLÓTIPO	DYS19	DYS385a	DYS385B	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS481	DYS533	DYS549	DYS570	DYS576	DYS635	DYS643	YGATAH4
H1	14	11	14	13	29	23	10	13	13	15	12	12	18	15	18	22	12	12	17	18	23	10	11
H2	13	13	14	13	29	24	9	11	13	14	10	11	20	16	18	27	11	11	23	19	21	12	12
H3	14	14	19	13	30	22	10	11	12	14	10	11	19	15	17.2	25	11	12	17	16	20	9	11
H4	15	13	17	12	28	24	10	11	12	16	9	11	19	14	16	23	12	11	18	16	22	9	11
H5	14	13	18	13	29	23	11	11	12	14	10	11	20	12	18.2	25	12	12	18	19	21	9	12
H6	13	10	14	12	28	24	10	13	13	14	12	12	18	15	17	22	12	13	19	17	23	10	11
H7	14	11	14	14	30	24	11	13	13	15	12	13	19	16	18	22	12	12	18	18	24	10	12
H8	13	16	16	12	29	25	10	11	12	14	10	12	21	15	19	24	10	12	19	18	21	12	11
H9	13	13	14	14	30	24	9	11	13	14	10	10	20	15	18	29	11	11	22	18	21	12	12
H10	15	11	13	13	29	24	11	13	13	15	12	12	19	16	17	22	13	13	17	18	23	11	12
H11	NA	15	15	13	29	23	11	NA	14	NA	10	NA	20	17	15	26	15	12	20	18	20	NA	NA
H12	12	12	14	13	29	24	11	13	13	15	12	11	20	15	16	23	12	13	17	17	23	9	11
H13	13	13	14	14	30	24	9	11	13	14	10	10	20	17	19	27	11	11	23	18	22	12	12
H14	15	13	17	14	31	25	9	11	12	14	9	13	21	16	14	22	12	12	18	16	21	NA	112
H15	13	15	18	13	30	24	10	11	13	14	10	12	20	17	15	23	12	12	22	17	21	11	11
H16	14	11	14	13	29	24	11	13	14	15	13	11	19	17	18	23	14	11	18	19	24	10	12
H17	16	11	14	12	29	25	11	11	13	14	11	10	20	15	16	23	12	12	19	18	23	10	12
H18	14	12	14	13	29	24	11	13	13	15	12	13	19	15	18	22	11	12	17	17	23	11	11
H19	14	12	14	12	29	24	11	13	13	15	10	13	19	15	16	22	12	13	17	18	23	10	12
H20	15	13	15	12	28	21	10	11	15	15	10	11	23	18	16	21	10	12	18	16	21	12	11
H21	15	13	15	12	28	22	10	11	13	16	10	11	21	15	15	22	9	14	17	17	21	11	12
H22	14	12	14	14	30	24	10	13	13	15	12	12	19	15	18	23	13	13	19	16	23	10	12
H23	14	11	14	13	29	24	11	13	13	15	12	12	19	16	16	22	12	12	17	18	23	10	12
H24	16	15	17	12	29	25	11	11	13	15	10	12	21	15	16	23	11	12	17	19	21	13	10
H25	14	15	18	13	31	24	10	11	12	14	10	12	20	15	17.2	24	11	14	16	18	21	9	11
H26	NA	11	14	13	28	24	11	12	13	NA	NA	11	17	NA	17	22	13	13	18	18	23	9	NA
H27	14	11	14	13	29	23	11	13	13	15	12	12	19	15	17	23	11	12	18	17	23	11	14
H28	14	11	13	13	29	24	11	13	14	15	12	12	19	16	17	22	12	13	17	20	23	10	12
H29	14	12	14	13	29	24	10	13	13	14	12	12	18	15	17	22	12	14	17	21	23	10	12
H30	14	14	14	12	27	23	10	11	13	16	10	11	20	14	17	26	11	11	20	16	22	12	11
H31	14	11	14	13	29	23	10	13	13	15	12	12	19	15	16	22	12	13	17	17	23	10	12
H32	13	13	14	14	30	25	10	11	13	14	NA	11	20	16	18	26	11	NA	22	20	22	NA	13
H33	14	11	14	13	29	23	10	13	14	15	12	12	19	15	18	22	13	14	18	19	23	NA	12
H34	15	11	15	13	30	24	10	13	14	15	11	12	20	15	17	22	12	13	19	18	24	10	12
H35	13	13	14	14	30	24	9	11	13	14	10	10	20	15	17	27	11	11	25	18	21	12	12
H36	14	13	15	14	31	23	10	11	12	14	10	11	20	15	18.2	24	11	12	17	19	20	12	11
H37	16	12	13	13	29	23	10	13	13	14	11	13	18	15	14	24	12	13	17	18	24	10	11
H38	14	11	14	13	29	25	10	13	13	15	12	12	18	15	17	23	10	13	18	17	23	10	12
H39	14	11	14	13	29	24	11	13	13	15	12	13	19	15	19	23	12	11	17	16	23	10	12
H40	14	11	14	13	30	23	10	14	13	15	12	12	19	16	17	22	12	13	19	18	23	10	12
H41	15	9	13	14	30	24	11	13	12	15	12	12	19	15	18	23	12	11	19	18	23	10	12
H42	16	15	15	14	30	23	10	12	14	15	10	12	17	18	16	27	13	12	19	17	20	13	11
H43	14	11	15	13	29	24	11	13	13	15	12	12	18	16	18	22	13	12	19	17	23	10	11
H44	13	16	18	13	31	24	10	11	12	14	10	12	21	17	15	21	12	12	22	19	23	12	12
H45	14	11	14	13	29	23	11	13	13	15	12	12	19	16	18	21	12	12	17	17	23	10	12
H46	13	14	14	13	29	25	9	11	13	14	10	10	20	16	18	27	11	11	22	17	21	12	12
H47	15	11	14	13	29	25	10	13	13	15	12	12	19	15	17	22	12	12	17	20	23	10	11
H48	15	12	15	13	29	23	10	11	13	14	10	12	18	14	18	23	13	12	16	17	21	12	11
H49	14	14	14	12	29	23	10	11	13	16	10	11	20	14	15	24	11	12	22	17	22	12	11
H50	13	14	18	13	30	24	10	11	13	14	10	12	20	17	16	22	12	12	19	18	23	11	12
H51	13	16	18	13	30	24	10	11	13	14	10	12	21	16	16	22	12	12	21	19	23	12	12
H52	16	14	14	13	31	22	10	11	14	16	10	11	21	16	16	21	9	12	17	14	20	11	12
H53	14	11	16	13	29	24	11	13	13	15	12	12	19	15	18	22	12	13	15	17	23	10	12
H54	15	11	15	13	30	25	11	13	13	14	12	12	18	17	16	22	12	12	18	18	23	10	11
H55	14	11	14	14	31	24	11	13	13	15	12	11	19	16	17	24	12	12	17	17	23	10	12
H56	13	16	19	13	30	24	10	11	12	14	10	11	21	17	15	22	12	12	22	18	23	12	12
H57	14	11	14	13	29	25	10	14	13	15	12	12	19	15	17	22	12	12	17	16	23	12	13
H58	15	15	17	13	31	23	10	11	14	14	11	12	20	16	15	22	11	11	20	15	22	12	11
H59	14	13	15	12	28	23	10	11	13	16	10	11	20	14	15	25	11	12	22	16	21	12	11
H60	14	13	17	13	30	23	10	11	12	14	10	11	21	15	18.2	24	12	12	17	17	20	9	11
H61	14	11	14	13	29	23	11	13	13	15	12	12	19	15	17	23	11	12	18	18	23	11	12
H62	14	11	14	13	29	24	10	13	13	15	12	12	19	16	16	23	14	13	19	17	23	10	12
H63	13	11	15	13	30	25	11	13	13	15	12	13	19	16	16	22	12	12	17	18	23	10	12
H64	14	11	13	13	29	24	11	14	13	15	12	13	19	15	17	22	12	14	17	20	24	10	11
H65	16	15	15	13	29	23	10	13	15	15	10	11	20	17	16	26	12	12	20	17	20	13	11

HAPLÓTIPO	DYS19	DYS385a	DYS385B	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS481	DYS533	DYS549	DYS570	DYS576	DYS635	DYS643	YGATAH4
H66	14	14	18	12	28	24	10	11	12	14	9	11	21	16	17	22	11	10	17	18	22	9	11
H67	14	11	14	15	31	24	11	13	13	15	12	12	19	16	17	21	13	11	18	17	23	10	12
H68	14	11	14	15	31	24	11	13	13	15	12	12	20	15	16	22	12	13	19	18	23	10	12
H69	14	10	14	13	29	24	11	13	12	15	12	12	19	15	16	22	12	11	18	16	23	10	13
H70	14	11	14	13	29	24	11	13	13	15	12	12	19	17	17	22	12	14	17	18	23	10	11
H71	16	11	15	13	31	24	11	11	13	14	11	11	21	16	16	22	12	13	20	19	24	11	12
H72	14	11	14	13	29	24	11	13	13	15	12	12	19	15	17	22	12	12	19	19	23	10	13
H73	13	13	14	14	30	24	9	12	13	14	10	10	20	16	18	27	11	11	22	18	21	12	12
H74	14	12	15	13	29	25	11	12	13	15	13	12	18	15	16	22	12	13	16	18	23	10	12
H75	14	12	13	14	30	24	10	14	13	15	12	12	19	16	16	22	12	13	17	18	23	10	12
H76	14	11	14	13	29	24	11	13	13	15	12	12	19	15	18	22	13	12	17	18	23	10	12
H77	15	14	14	13	29	24	9	11	13	14	10	10	20	18	18	28	11	11	22	19	21	13	11
H78	13	14	14	14	30	24	9	11	13	14	10	10	20	16	18	27	11	11	23	18	21	12	12
H79	14	13	13	13	29	24	11	13	13	15	11	11	19	16	18	22	12	13	18	18	23	10	13
H80	14	11	13	13	29	24	11	13	12	14	12	12	19	15	16	22	13	12	17	18	23	10	12
H81	14	11	14	13	30	23	11	13	13	15	12	12	19	16	19	21	12	12	17	17	23	10	13
H82	14	11	14	13	29	25	10	14	13	15	12	12	19	15	18	22	13	13	19	19	23	10	12
H83	15	14	15	12	30	23	11	11	14	16	10	12	21	15	15	23	10	13	18	16	23	11	12
H84	14	14	122	14	29	24	11	14	12	15	12	13	19	15	18	22	12	12	17	18	24	10	13
H85	14	11	14	13	29	24	11	13	13	15	12	12	19	17	15	23	13	13	18	17	23	11	12
H86	17	12	17	13	30	24	11	11	13	14	10	11	20	17	17	25	12	11	20	17	21	12	11
H87	15	15	15	13	30	24	10	12	15	15	10	11	20	14	16	24	10	12	18	19	19	11	11
H88	NA	12	14	12	?	24	10	13	13	NA	12	12	18	15	17	22	12	13	18	19	23	NA	12
H89	14	13	17	13	29	23	11	11	12	14	10	11	20	13	18.2	25	12	12	18	18	21	9	12
H90	NA	12	15	13	29	25	11	NA	13	NA	13	NA	18	15	16	22	12	12	16	19	23	NA	12
H91	14	11	14	13	29	24	10	13	13	14	12	12	18	16	18	23	12	12	18	20	23	10	11
H92	14	13	13	13	30	24	10	13	13	15	12	11	19	15	17	22	12	13	18	20	24	11	12
H93	14	11	14	14	30	24	11	13	14	15	12	13	19	15	17	22	12	13	18	18	23	10	12
H94	14	11	15	13	28	24	11	13	14	15	12	11	19	16	18	23	12	13	16	18	23	11	12
H95	15	11	14	14	30	24	11	13	13	15	12	12	19	16	17	22	13	13	16	19	23	10	12
H96	14	14	14	12	29	23	10	11	13	16	10	11	20	14	15	24	11	12	22	17	22	12	11
H97	13	13	14	13	29	24	9	11	13	14	10	11	19	16	18	27	12	12	21	18	22	12	12
H98	14	11	14	13	29	23	11	13	13	15	12	12	19	16	19	22	12	14	18	19	24	10	12
H99	13	13	14	14	30	24	9	11	13	14	10	10	20	16	18	27	11	11	22	19	22	12	11
H100	15	11	13	13	29	24	11	13	13	14	12	11	18	17	18	22	11	13	17	19	23	9	12
H101	14	13	15	12	28	23	10	11	13	16	10	11	20	14	15	25	11	12	21	16	21	12	11
H102	13	16	18	13	31	23	10	11	13	14	10	11	20	15	15	21	12	14	21	17	22	12	11
H103	14	11	14	12	28	23	10	12	14	15	12	11	19	15	18	21	12	11	16	21	23	10	12
H104	14	11	14	14	30	24	10	13	12	NA	12	12	19	16	15	21	12	13	20	18	23	10	14
H105	13	14	16	13	32	23	10	11	12	14	9	11	20	16	19	23	11	11	14	15	22	NA	11
H106	NA	17	18	12	NA	24	10	NA	NA	14	10	12	20	16	15	22	12	12	20	17	21	NA	12
H107	13	16	18	13	32	26	10	11	13	14	10	13	20	15	17	25	10	13	18	18	21	13	11
H108	13	15	17	12	30	24	10	11	13	14	10	12	19	16	16	23	13	12	20	19	22	13	12
H109	NA	11	14	13	29	24	10	NA	13	NA	NA	13	19	15	17	23	12	13	17	18	23	NA	12
H110	15	13	13	14	32	22	12	11	13	16	10	13	21	13	19	21	11	12	18	18	21	11	11
H111	17	19	20	13	31	24	10	11	13	14	10	11	20	18	15	24	12	13	21	17	20	12	12
H112	16	14	16	13	29	23	10	13	13	14	9	11	19	15	18	22	12	13	19	16	22	10	11
H113	14	11	14	13	29	24	11	13	13	15	12	11	19	16	17	22	12	12	16	18	24	10	12
H114	14	11	14	12	28	24	11	13	13	15	10	12	19	16	17	22	11	12	20	18	23	10	12
H115	14	13	15	12	28	23	10	11	13	16	10	11	20	15	15	25	11	12	19	14	21	12	11
H116	14	13	17	12	30	23	11	11	12	14	10	11	20	16	17.2	27	11	12	17	18	20	10	11
H117	14	11	14	13	29	23	10	13	13	15	12	11	18	15	16	22	12	13	16	19	23	10	12
H118	15	11	14	13	29	24	11	13	13	15	12	12	19	15	17	23	13	13	17	17	23	9	12
H119	16	16	17	12	28	24	10	11	12	16	9	12	19	13	16	23	12	13	20	17	21	9	11
H120	14	11	13	12	28	25	11	13	13	15	12	12	19	15	17	23	12	14	18	18	24	10	11
H121	14	11	14	13	28	23	10	13	13	15	12	11	19	16	17	23	12	12	17	16	23	10	12
H122	14	11	15	14	30	24	11	14	13	15	12	12	19	16	?	21	13	12	16	17	23	11	12
H123	14	12	14	13	30	24	11	13	13	15	12	12	19	16	19	22	12	13	16	19	23	10	12
H124	14	13	17	13	28	24	10	11	13	15	9	13	21	15	16	22	11	15	16	16	22	10	12
H125	14	12	15	14	30	24	9	11	13	14	10	10	19	15	18	28	12	11	22	19	22	12	12
H126	14	13	13	13	29	25	10	11	12	15	9	11	21	15	20	24	12	12	15	19	22	11	11
H127	14	12	12	12	29	24	11	14	13	15	12	14	19	15	14	22	11	13	18	18	23	10	11
H128	14	11	14	13	28	24	11	13	13	15	12	12	19	15	17	22	12	14	17	18	23	10	12
H129	13	13	14	13	29	24	9	11	13	14	10	11	20	16	18	28	11	11	22	19	21	12	12
H130	15	13	14	12	30	23	10	11	15	16	10	11	20	17	16	22	9	12	18	14	21	10	12

HAPLÓTIPO	DYS19	DYS385a	DYS385B	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS481	DYS533	DYS549	DYS570	DYS576	DYS635	DYS643	YGATAH4
H131	13	13	14	13	29	24	10	11	13	14	10	10	20	16	18	29	11	11	23	18	21	12	12
H132	NA	15	16	13	30	23	10	12	15	NA	10	11	20	NA	15	29	12	13	17	17	21	12	NA
H133	14	11	14	13	28	24	10	13	13	15	12	11	19	15	16	22	13	13	19	18	23	10	12
H134	14	12	18	14	30	25	10	11	12	14	10	11	21	14	21.2	26	12	11	19	17	23	9	11
H135	14	11	14	14	NA	24	11	NA	13	NA	12	12	19	16	18	22	12	13	17	17	23	10	NA
H136	15	10	14	12	28	24	11	13	13	NA	NA	13	19	NA	18	23	12	14	17	19	23	NA	NA
H137	14	13	14	13	29	24	9	11	13	NA	NA	10	20	NA	17	31	11	11	22	19	21	NA	11
H138	NA	12	14	12	26	24	11	NA	13	NA	12	10	19	16	16	22	11	13	17	18	24	NA	12
H139	14	11	14	13	29	25	10	13	14	16	12	13	19	15	16	21	14	13	18	16	23	11	11.3
H140	13	16	16	13	30	24	10	11	13	14	10	11	NA	NA	16	21	12	12	21	17	22	12	NA
H141	15	11	14	14	29	24	11	13	13	15	12	11	17	16	16	22	12	13	17	21	23	10	NA
H142	16	14	15	12	28	22	10	11	14	NA	11	11	21	16	16	23	9	13	18	17	20	11	13
H143	14	13	16	13	30	25	10	11	13	15	10	11	20	16	16	25	12	12	16	18	21	9	10
H144	14	11	15	13	29	24	11	13	12	15	12	12	19	15	16	23	12	13	16	18	23	10	11
H145	14	11	14	13	29	24	11	13	13	15	12	12	18	15	18	22	12	12	18	19	23	10	13
H146	14	12	14	13	30	24	11	13	13	15	12	13	21	16	18	22	12	12	18	18	23	11	13

APÊNDICE 2 - Haplótipos encontrados para a população miscigenada brasileira classificada como africana (n=11).

HAPLÓTIPO	DYS19	DYS385a	DYS385B	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS481	DYS533	DYS549	DYS570	DYS576	DYS635	DYS643	YGATAH4
H147	17	15	19	14	31	21	10	11	15	14	11	12	21	14	16	24	11	17	18	21	13	11	11
H148	14	14	20	12	28	25	11	11	13	14	11	11	19	15	18	25	12	20	15	23	11	11	11
H149	15	14	15	12	28	22	11	11	12	16	9	12	NA	12	17	26	13	14	16	20	10	11	11
H150	16	14	16	12	29	21	10	11	14	14	11	13	21	15	16	25	12	20	17	21	12	13	13
H151	16	16	18	13	31	21	10	11	15	14	11	12	21	16	16	26	11	20	17	21	13	11	11
H152	17	16	17	12	29	21	10	11	13	14	11	13	21	15	18	29	12	21	17	22	14	12	12
H153	15	16	17	13	31	21	11	11	13	14	11	11	21	15	16	28	11	20	15	22	14	13	13
H154	NA	16	18	13	30	21	10	11	15	14	11	11	21	15	16	24	11	16	17	21	13	11	11
H155	NA	15	20	14	31	21	10	NA	13	NA	11	13	21	NA	16	27	11	20	14	21	0	12	12
H156	17	17	20	14	31	21	10	11	15	14	11	12	21	16	17	25	11	17	16	21	11	11	11
H157	15	15	18	13	30	21	10	11	14	14	11	12	21	15	16	28	11	17	14	22	12	12	12

APÊNDICE 3 - Haplótipos encontrados para a população miscigenada brasileira classificada como ameríndia (n=19).

HAPLÓTIPO	DYS19	DYS385a	DYS385B	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS481	DYS533	DYS549	DYS570	DYS576	DYS635	DYS643	YGATAH4
H158	13	14	17	12	30	24	10	14	13	14	11	13	20	16	17	28	13	12	17	20	22	10	12
H159	13	14	16	14	30	23	11	14	13	14	10	11	20	15	18	24	NA	12	18	19	23	10	12
H160	12	14	20	14	30	25	10	15	13	14	11	12	20	15	17	24	12	12	17	20	23	10	11
H161	13	14	16	14	31	24	11	14	11	15	11	14	19	15	15	24	12	13	17	18	22	12	11
H162	13	14	16	14	31	24	11	NA	11	NA	11	15	19	NA	15	24	12	13	17	18	22	11	10
H163	16	14	17	13	30	24	11	14	14	14	11	12	20	15	20	24	11	13	18	19	22	10	11
H164	13	14	14	12	29	24	10	15	12	14	11	13	20	16	19	23	11	12	17	18	23	9	12
H165	13	14	16	13	30	24	10	14	13	14	11	11	20	16	19	24	12	12	16	20	22	10	11
H166	13	14	14	12	29	24	10	15	12	14	11	12	20	16	19	23	11	12	17	18	23	9	11
H167	13	13	14	12	29	24	10	15	12	14	11	12	20	16	19	23	12	12	17	19	23	9	11
H168	13	15	18	13	29	24	10	14	13	14	11	11	19	15	16	24	11	13	16	15	22	10	10
H169	13	14	14	11	28	24	10	15	12	14	11	13	20	16	18	23	12	12	17	19	23	9	12
H170	13	14	14	13	31	25	11	14	13	15	11	12	20	15	17	24	11	13	12	17	23	10	12
H171	13	13	17	13	30	23	10	14	14	14	10	11	19	15	16	27	12	13	18	15	22	11	12
H172	13	14	14	12	29	24	10	15	12	14	11	13	20	16	19	23	12	13	17	18	23	9	12
H173	14	13	14	13	29	24	10	14	13	14	11	12	20	15	15	25	10	12	17	19	23	10	11
H174	13	14	14	13	31	25	11	14	13	15	11	12	20	16	18	24	11	13	15	17	23	10	13

Legenda: NA: falha de amplificação.

APÊNDICE 4 - Coeficientes de probabilidade de inferência estimados pelos três softwares para os 22 indivíduos que apresentaram erros de atribuição pelo STRUCTURE considerando o conjunto de 15 corridas de K=20.

INDIVÍDUO	HAPLOGRUPO (CONFORME Y-SNPs)	PROGRAMA	CORRIDA	ATRIBUIÇÃO A HAPLOGRUPO / CLADO	PROBABILIDADE DA INFERÊNCIA	ERRO	INCERTEZA
GRC 134	Q1b1a1a M3	STRUCTURE	1	L,RT	56,40%	6,67%	13,33%
			2	Q	79,90%		
			3	E,N,Q	65,60%		
			4	N,Q,RT	66,50%		
			5	N,Q,RT	65,80%		
			6	Q	88,40%		
			7	E,N,Q,RT	59,50%		
			8	E,N,Q	80,70%		
			9	Q	77,50%		
			10	INCERTEZA	<50%		
			11	INCERTEZA	<50%		
			12	N,Q,RT	66,20%		
			13	E,N,Q	63,50%		
			14	D,E,N,Q	59,60%		
			15	E,N,Q,S	63,40%		
		HAPEST	-	Q	99,90%	-	-
		NEVGEN	-	INCERTEZA	<50%	-	-
GRC 135	Q1b1a1a M3	STRUCTURE	1	L,RT	56,40%	6,67%	13,33%
			2	Q	79,90%		
			3	E,N,Q	65,60%		
			4	N,Q,RT	66,50%		
			5	N,Q,RT	65,80%		
			6	Q	88,40%		
			7	E,N,Q,RT	59,50%		
			8	E,N,Q	80,70%		
			9	Q	77,50%		
			10	INCERTEZA	<50%		
			11	INCERTEZA	<50%		
			12	N,Q,RT	66,20%		
			13	E,N,Q	63,50%		
			14	D,E,N,Q	59,60%		
			15	E,N,Q,S	63,40%		
		HAPEST	-	Q	99,90%	-	-
		NEVGEN	-	INCERTEZA	<50%	-	-
P294NAS	Q1b1a1a M3	STRUCTURE	1	C,D,E,F,Q	89,00%	26,67%	0,00%
			2	Q	70,70%		
			3	E,N,Q	72,90%		
			4	B,C,D,F	54,20%		
			5	B,C,D,F	51,80%		
			6	Q	51,30%		
			7	B,C,D,F	54,10%		
			8	E,N,Q	74,50%		
			9	Q	51,10%		
			10	B,C,D,E,F,Q	80,10%		
			11	B,C,D,E,N,Q	80,70%		
			12	B,C,D,F	52,90%		
			13	E,N,Q	69,30%		
			14	D,E,N,Q	69,30%		
			15	E,N,Q	70,10%		
		HAPEST	-	Q	99,90%	-	-
		NEVGEN	-	Q	89,95%	-	-

INDIVÍDUO	HAPLOGRUPO (CONFORME Y-SNPs)	PROGRAMA	CORRIDA	ATRIBUIÇÃO A HAPLOGRUPO / CLADO	PROBABILIDADE DA INFERÊNCIA	ERRO	INCERTEZA
P360JIS	Q1b1a1a M3	STRUCTURE	1	C,D,E,F,Q	98,50%	6,67%	33,33%
			2	Q	72,40%		
			3	E,N,Q	84,30%		
			4	INCERTEZA	<50%		
			5	INCERTEZA	<50%		
			6	INCERTEZA	<50%		
			7	INCERTEZA	<50%		
			8	E,N,Q	81,60%		
			9	B,C,D,E	50,80%		
			10	B,C,D,E,F,Q	97,50%		
			11	B,C,D,E,N,Q	96,00%		
			12	INCERTEZA	<50%		
			13	E,N,Q	84,60%		
			14	D,E,N,Q	89,90%		
			15	E,N,Q,S	85,50%		
		HAPEST	-	Q	99,60%	-	-
		NEVGEN	-	Q	85,95%	-	-
P298MNC	Q1b1a1a M3	STRUCTURE	1	RB	76,50%	60,00%	6,67%
			2	Q	76,90%		
			3	E,N,Q	74,30%		
			4	N,Q,RT	82,70%		
			5	N,Q,RT	51,10%		
			6	RB	54,40%		
			7	E,N,Q,RT	54,50%		
			8	RB	58,50%		
			9	RB	54,10%		
			10	RB	74,00%		
			11	RB	69,00%		
			12	INCERTEZA	<50%		
			13	RB	60,50%		
			14	RB	63,00%		
			15	RB	60,00%		
		HAPEST	-	Q	100,00%	-	-
		NEVGEN	-	Q	0,25%	-	-
CP457LCJ	E1b1b1a1 M78	STRUCTURE	1	C,D,E,F,Q	66,60%	13,33%	26,67%
			2	E	69,60%		
			3	INCERTEZA	<50%		
			4	E,H,I,J,S,T	63,60%		
			5	E,G,H,I,J,S,T	60,60%		
			6	B,C,D,E	93,00%		
			7	INCERTEZA	<50%		
			8	INCERTEZA	<50%		
			9	B,C,D,E	91,00%		
			10	B,C,D,E,F,Q	82,10%		
			11	B,C,D,E,N,Q	55,10%		
			12	E,H,I,J,S,T	55,00%		
			13	B,C,D,F	51,40%		
			14	INCERTEZA	<50%		
			15	B,C,D,F	51,30%		
		HAPEST	-	E1b1b	100,00%	-	-
		NEVGEN	-	INCERTEZA	<50%	-	-

INDIVÍDUO	HAPLOGRUPO (CONFORME Y-SNPs)	PROGRAMA	CORRIDA	ATRIBUIÇÃO A HAPLOGRUPO / CLADO	PROBABILIDADE DA INFERÊNCIA	ERRO	INCERTEZA
CP446TEO	E1b1b1b2a1 M123	STRUCTURE	1	C,D,E,F	80,90%	66,67%	6,67%
			2	B,C,D,F	59,20%		
			3	B,C,D,F	65,40%		
			4	B,C,D,F,	55,00%		
			5	B,C,D,F	51,10%		
			6	B,C,D,E	75,00%		
			7	B,C,D,F	53,40%		
			8	B,C,D,F	66,90%		
			9	B,C,D,E	82,90%		
			10	B,C,D,E,F,Q	84,60%		
			11	C,F,J	64,40%		
			12	B,C,D,F	58,30%		
			13	B,C,D,F	66,90%		
			14	INCERTEZA	<50%		
			15	B,C,D,F	66,20%		
		HAPEST	-	E1b1b	100,00%	-	-
		NEVGEN	-	E1b1b	95,64%	-	-
CP439JCO	J2a1a L26	STRUCTURE	1	INCERTEZA	<50%	13,33%	26,67%
			2	E,H,J,T	64,70%		
			3	H,J,T,	71,60%		
			4	N	71,60%		
			5	E,G,H,I,J,S,T	76,20%		
			6	B,C,D,E	79,40%		
			7	H,I,J,S,T	54,50%		
			8	INCERTEZA	<50%		
			9	H,J	54,30%		
			10	INCERTEZA	<50%		
			11	INCERTEZA	<50%		
			12	E,H,I,J,S,T	82,80%		
			13	H,J,T	70,20%		
			14	H,J,T	69,20%		
			15	H,J,T	70,20%		
		HAPEST	-	J1	99,40%	ERRO	-
		NEVGEN	-	J2a1	70,83%	-	-
P364INS	I2a1b1 M223	STRUCTURE	1	INCERTEZA	<50%	6,67%	93,33%
			2	INCERTEZA	<50%		
			3	INCERTEZA	<50%		
			4	INCERTEZA	<50%		
			5	INCERTEZA	<50%		
			6	INCERTEZA	<50%		
			7	INCERTEZA	<50%		
			8	INCERTEZA	<50%		
			9	INCERTEZA	<50%		
			10	B,C,D,E,F,Q	53,70%		
			11	INCERTEZA	<50%		
			12	INCERTEZA	<50%		
			13	INCERTEZA	<50%		
			14	INCERTEZA	<50%		
			15	INCERTEZA	<50%		
		HAPEST	-	I2a (xI2a1)	87,50%	-	-
		NEVGEN	-	I2a2a	73,54%	ERRO	-

INDIVÍDUO	HAPLOGRUPO (CONFORME Y-SNPs)	PROGRAMA	CORRIDA	ATRIBUIÇÃO A HAPLOGRUPO / CLADO	PROBABILIDADE DA INFERÊNCIA	ERRO	INCERTEZA
C611AKN	E1b1b1b2a1 M123	STRUCTURE	1	C,D,E,F,Q	62,80%	33,33%	20,00%
			2	B,C,D,F	51,00%		
			3	B,C,D,F,(1A)	68,50%		
			4	E,H,I,J,S,T	57,40%		
			5	E,G,H,I,J,S,T	57,50%		
			6	B,C,D,E	88,50%		
			7	INCERTEZA	<50%		
			8	B,C,D,F,(1A)	66,60%		
			9	B,C,D,E	88,10%		
			10	B,C,D,E,F,Q	63,00%		
			11	INCERTEZA	<50%		
			12	E,H,I,J,S,T	63,90%		
			13	B,C,D,F	68,80%		
			14	INCERTEZA	<50%		
			15	B,C,D,F	67,20%		
		HAPEST	-	E1b1b	96,80%	-	-
		NEVGEN	-	E1b1b	79,09%	-	-
CP331TRR	I2a1b1 M223	STRUCTURE	1	C,D,E,F,Q	60,70%	40,00%	46,67%
			2	E	71,60%		
			3	INCERTEZA	<50%		
			4	E,H,I,J,S,T	52,60%		
			5	E,G,H,I,J,S,T	50,40%		
			6	B,C,D,E	99,10%		
			7	INCERTEZA	<50%		
			8	INCERTEZA	<50%		
			9	B,C,D,E	98,90%		
			10	B,C,D,E,F,Q	75,10%		
			11	B,C,D,E,N,Q	70,00%		
			12	INCERTEZA	<50%		
			13	INCERTEZA	<50%		
			14	INCERTEZA	<50%		
			15	INCERTEZA	<50%		
		HAPEST	-	I2a (xI2a1)	0,997	ERRO	-
		NEVGEN	-	I2a1	95,55%		-
P324DPS	I1 M253	STRUCTURE	1	B,H,I,J	86,10%	6,67%	0,00%
			2	E,G,I	51,50%		
			3	E,G,I	79,00%		
			4	E,H,I,J,S,T	83,10%		
			5	E,G,H,I,J,S,T	79,90%		
			6	B,C,D,E	52,40%		
			7	H,I,J,S,T	55,30%		
			8	H,I,J,L,	65,00%		
			9	B,G,I	63,40%		
			10	H,I,J,RT	84,70%		
			11	G,I	77,50%		
			12	E,H,I,J,S,T	79,00%		
			13	E,G,I	79,60%		
			14	G,I	80,00%		
			15	E,G,I	78,50%		
		HAPEST	-	I1	100%	-	-
		NEVGEN	-	INCERTEZA	<50%	-	-

INDIVÍDUO	HAPLOGRUPO (CONFORME Y-SNPs)	PROGRAMA	CORRIDA	ATRIBUIÇÃO A HAPLOGRUPO / CLADO	PROBABILIDADE DA INFERÊNCIA	ERRO	INCERTEZA
CP456CLO	G-M201	STRUCTURE	1	B,H,I,J	90,60%	46,67%	0,00%
			2	E,G,I	68,30%		
			3	E,G,I	81,10%		
			4	E,H,I,J,S,T	92,20%		
			5	E,G,H,I,J,S,T	90,90%		
			6	H,I,J,L,S	85,00%		
			7	H,I,J,S,T	87,50%		
			8	H,I,J,L,	93,20%		
			9	B,G,I	66,80%		
			10	H,I,J,RT	83,40%		
			11	G,I	73,80%		
			12	E,H,I,J,S,T	89,80%		
			13	E,G,I	80,40%		
			14	G,I	77,60%		
			15	E,G,I	79,70%		
		HAPEST	-	G2a	100%	-	-
		NEVGEN	-	INCERTEZA	<50%	-	-
P389DVC	I2a1b1 M223	STRUCTURE	1	C,D,E,F,Q	71,10%	73,33%	0,00%
			2	B,C,D,F	64,80%		
			3	B,C,D,F	83,20%		
			4	E,H,I,J,S,T	61,30%		
			5	E,G,H,I,J,S,T	64,30%		
			6	H,I,J,L,S	62,10%		
			7	B,C,D,F	51,10%		
			8	B,C,D,F	77,20%		
			9	B,C,D,E	52,40%		
			10	B,C,D,E,F,Q	91,10%		
			11	C,F,J	63,20%		
			12	E,H,I,J,S,T	66,60%		
			13	B,C,D,F	82,20%		
			14	B,C,F	56,00%		
			15	B,C,D,F	82,90%		
		HAPEST	-	I2a (xI2a1)	99,90%	ERRO	-
		NEVGEN	-	I2a1	95,77%	-	-
P203JCA	R1b1a1b1a L51	STRUCTURE	1	C,D,E,F,Q	71,10%	100%	0,00%
			2	B,C,D,F	64,80%		
			3	B,C,D,F	83,20%		
			4	E,H,I,J,S,T	61,30%		
			5	E,G,H,I,J,S,T	64,30%		
			6	B,C,D,E	61,10%		
			7	B,C,D,F	51,10%		
			8	B,C,D,F	77,20%		
			9	B,C,D,E	52,40%		
			10	B,C,D,E,F,Q	91,10%		
			11	C,F,J	63,20%		
			12	E,H,I,J,S,T	66,60%		
			13	B,C,D,F	82,20%		
			14	B,C,F	56,00%		
			15	B,C,D,F	81,40%		
		HAPEST	-	J1	98%	ERRO	-
		NEVGEN	-	INCERTEZA	<50%	-	-

INDIVÍDUO	HAPLOGRUPO (CONFORME Y-SNPs)	PROGRAMA	CORRIDA	ATRIBUIÇÃO A HAPLOGRUPO / CLADO	PROBABILIDADE DA INFERÊNCIA	ERRO	INCERTEZA
P147CAS	J2a1a1a2b2 M67	STRUCTURE	1	INCERTEZA	<50%	13,33%	60,00%
			2	INCERTEZA	<50%		
			3	INCERTEZA	<50%		
			4	E,H,I,J,S,T	53,40%		
			5	E,G,H,I,J,S,T	54,70%		
			6	B,C,D,E	74,20%		
			7	INCERTEZA	<50%		
			8	INCERTEZA	<50%		
			9	B,C,D,E	61,90%		
			10	INCERTEZA	<50%		
			11	C,F,J	59,40%		
			12	E,H,I,J,S,T	60,60%		
			13	INCERTEZA	<50%		
			14	INCERTEZA	<50%		
			15	INCERTEZA	<50%		
		HAPEST	-	J1	100%	ERRO	-
		NEVGEN	-	INCERTEZA	<50%	-	-
P179ASC	R1a1a1 M417	STRUCTURE	1	RB	99,70%	100,00%	0,00%
			2	RB	99,40%		
			3	RB	99,10%		
			4	RB	99,10%		
			5	RB	99,20%		
			6	RB	99,60%		
			7	RB	99,30%		
			8	RB	99,40%		
			9	RB	99,50%		
			10	RB	99,70%		
			11	RB	99,30%		
			12	RB	99,30%		
			13	RB	99,30%		
			14	RB	99,30%		
			15	RB	99,30%		
		HAPEST	-	R1b	100%	ERRO	-
		NEVGEN	-	R1b	99,92%	ERRO	-
CP335JPS	I2a1b1 M223	STRUCTURE	1	B,H,I,J	73,80%	46,67%	40,00%
			2	M,N,O,S	52,70%		
			3	INCERTEZA	<50%		
			4	E,H,I,J,S,T	52,60%		
			5	M,N,O	55,20%		
			6	M,N,O,T	72,30%		
			7	M,O	54,10%		
			8	M,O,S,T	61,80%		
			9	M,N,O,RT,S	52,50%		
			10	M,O,S,T	55,50%		
			11	INCERTEZA	<50%		
			12	INCERTEZA	<50%		
			13	INCERTEZA	<50%		
			14	INCERTEZA	<50%		
			15	INCERTEZA	<50%		
		HAPEST	-	I2b1	99,70%	ERRO	-
		NEVGEN	-	I2a2a	63,76%	ERRO	-

INDIVÍDUO	HAPLOGRUPO (CONFORME Y-SNPs)	PROGRAMA	CORRIDA	ATRIBUIÇÃO A HAPLOGRUPO / CLADO	PROBABILIDADE DA INFERÊNCIA	ERRO	INCERTEZA
CP445LPS	E1b1b1a1b1a V13	STRUCTURE	1	C,D,E,F,Q	86,10%	6,67%	6,67%
			2	INCERTEZA	<50%		
			3	E,N,Q	71,30%		
			4	E,H,I,J,S,T	93,00%		
			5	E,G,H,I,J,S,T	86,30%		
			6	B,C,D,E	60,70%		
			7	H,I,J,S,T	56,50%		
			8	E,N,Q	51,80%		
			9	B,C,D,E	81,50%		
			10	B,C,D,E,F,Q	91,90%		
			11	B,C,D,E,N,Q	92,60%		
			12	E,H,I,J,S,T	97,30%		
			13	E,N,Q	75,60%		
			14	D,E,N,Q	90,20%		
			15	E,N,Q,S	75,30%		
		HAPEST	-	E1b1b	100%	-	-
		NEVGEN	-	E1b1b	85,85%	-	-
CP458DPD	T1a M70	STRUCTURE	1	B,H,I,J	71,30%	20,00%	20,00%
			2	INCERTEZA	<50%		
			3	H,J,T	58,40%		
			4	E,H,I,J,S,T	94,50%		
			5	E,G,H,I,J,S,T	94,00%		
			6	H,I,J,L,S	78,40%		
			7	H,I,J,S,T	85,20%		
			8	H,I,J,L,	81,20%		
			9	INCERTEZA	<50%		
			10	H,I,J,RT	70,30%		
			11	INCERTEZA	<50%		
			12	E,H,I,J,S,T	96,50%		
			13	H,J,T	57,60%		
			14	H,J,T	53,20%		
			15	H,J,T	59,80%		
		HAPEST	-	E1b1b	100%	ERRO	-
		NEVGEN	-	T	99,74%	-	-
KRC064	Q1b1a1a M3	STRUCTURE	1	C,D,E,F,Q	89,30%	13,33%	26,67%
			2	Q	77,40%		
			3	E,N,Q	91,00%		
			4	INCERTEZA	<50%		
			5	INCERTEZA	<50%		
			6	B,C,D,E	53,40%		
			7	INCERTEZA	<50%		
			8	E,N,Q	89,70%		
			9	B,C,D,E	55,20%		
			10	B,C,D,E,F,Q	79,70%		
			11	B,C,D,E,N,Q	87,00%		
			12	INCERTEZA	<50%		
			13	E,N,Q	92,80%		
			14	D,E,N,Q	94,20%		
			15	E,N,Q,S	93,20%		
		HAPEST	-	Q	99,80%	-	-
		NEVGEN	-	Q	93,78%	-	-

INDIVÍDUO	HAPLOGRUPO (CONFORME Y-SNPs)	PROGRAMA	CORRIDA	ATRIBUIÇÃO A HAPLOGRUPO / CLADO	PROBABILIDADE DA INFERÊNCIA	ERRO	INCERTEZA
C6490PS	E1a2a1b E-L133.1	STRUCTURE	1	B,H,I,J	81,40%	80,00%	6,67%
			2	E,G,I	60,70%		
			3	H,J,T,	50,80%		
			4	E,H,I,J,S,T	60,80%		
			5	E,G,H,I,J,S,T	58,80%		
			6	H,I,J,L,S	57,20%		
			7	H,I,J,S,T	56,00%		
			8	H,I,J,L,	56,10%		
			9	B,G,I	60,40%		
			10	H,I,J,RT	59,50%		
			11	H,J,RT	54,60%		
			12	E,H,I,J,S,T	64,60%		
			13	INCERTEZA	<50%		
			14	H,J,T	51,80%		
			15	H,J,T	50,40%		
		HAPEST	-	L	70,60%	ERRO	-
		NEVGEN	-	J2a2	<50%	-	INCERTEZA

Legenda: Cor cinza: incertezas de atribuição, com valores de probabilidade inferiores a 50%. Cor amarela: erros de atribuição quando comparados ao haplogrupo definido através dos Y-SNPs (Joerin et al. (2022). RB: Haplogrupos R1b e seus subhaplogrupos. RA: haplogrupo R1a e seus subhaplogrupos.

APÊNDICE 5 - Resultados da inferência de haplogrupos realizada pelos três programas *STRUCTURE, HAPEST e NevGen*.

ID	HAPLOGRUPO Y-SNP (Joerin-Luque et al, 2022)	HAPEST	PROBABILIDADE HAPEST (%)	NEVGEN	Probabilidade NEVGEN (%)	STRUCTURE (K=20)	PROBABILIDADE STRUCTURE (%)
C649OPS	**E1a2a1b**	L	70,6	INCERTEZA	<50	ERRO	
C689EGB	**E1b1a1a1**	E1b1a	100	E1b1a	88,46	E	99,7
315DGA	**E1b1a1a1a1c1a1a**	E1b1a	100	E1b1a	98,32	E	100
CP352CCS	**E1b1a1a1a1c1a1a**	E1b1a	100	E1b1a	99,76	E	99,9
P102LAC	**E1b1a1a1a1c1a1a**	E1b1a	100	E1b1a	100	E	100
P290ORC	**E1b1a1a1a1c1a1a**	E1b1a	100	E1b1a	99,11	E	99,9
CP465NGP	**E1b1a1a1a2a1**	E1b1a	100	E1b1a	92,68	E	91,8
GRC159	**E1b1a1a1a2a1**	E1b1a	100	E1b1a	100	E	100
P126FGA	**E1b1a1a1a2a1**	E1b1a	100	E1b1aE1b1a	96,91	E	99,5
PK209BJS	**E1b1a1a1a2a1**	E1b1a	100	E1b1b	100	E	100
CP457LCJ	**E1b1b1a1**	E1b1b	100	INCERTEZA	<50	ERRO	
P216WIC	**E1b1b1a1**	I2b1	92,1	I2a2a	84,93	INCERTEZA	<50
CP416LUN	**E1b1b1a1b1a**	E1b1b	100	E1b1b	96,82	E,N,Q	71
CP445LPS	**E1b1b1a1b1a**	E1b1b	100	E1b1b	98,85	E,N,Q	75,3
P186LPM	**E1b1b1a1b1a**	E1b1b	100	E1b1b	99,47	E,N,Q	82,8
P238JCL	**E1b1b1a1b1a**	E1b1b	100	E1b1b	99,85	E,N,Q	50,4
P385BXI	**E1b1b1a1b1a**	E1b1b	100	E1b1b	97,69	E	52,7
P516PDL	**E1b1b1a1b1a**	E1b1b	100	E1b1b	99,53	E	96,6
P519CPB	**E1b1b1a1b1a**	E1b1b	100	E1b1b	99,85	E	79,3
PK238LAS	**E1b1b1a1b1a**	E1b1b	100	E1b1b	97,23	INCERTEZA	<50
CP449VVF	**E1b1b1a1b1a V13**	E1b1b	100	E1b1b	95,13	E	99,6
C617DAC	**E1b1b1b1a1**	E1b1b	99,3	E1b1b	100	E	99,7
C635AAT	**E1b1b1b1a1**	E1b1b	99,6	E1b1b	100	E	90,8
C703RNS	**E1b1b1b1a1**	E1b1b	100	E1b1b	100	E	99,6
CK520PCM	**E1b1b1b1a1**	E1b1b	88,4	E1b1b	100	E	81,7
CP302JVS	**E1b1b1b1a1**	E1b1b	100	E1b1b	100	E	100
CP368LFS	**E1b1b1b1a1**	E1b1b	99,8	E1b1b	100	E	63,8
CP382GFA	**E1b1b1b1a1**	E1b1b	100	E1b1b	100	E	99,9
CP491RJF	**E1b1b1b1a1**	E1b1b	96,3	E1b1b	100	E	75,7
P109IFX	**E1b1b1b1a1**	E1b1b	100	E1b1b	100	E	98,3
P122FPS	**E1b1b1b1a1**	E1b1b	99,7	E1b1b	100	E	99,7
P167HRS	**E1b1b1b1a1**	E1b1b	59,2	E1b1b	99,98	E	85,4
P225SOL	**E1b1b1b1a1**	E1b1b	100	E1b1b	100	E	100
P316VRS	**E1b1b1b1a1**	E1b1b	100	E1b1b	100	E	99,9
P330WLI	**E1b1b1b1a1**	E1b1b	99,7	E1b1b	100	E	95,2
P387YLU	**E1b1b1b1a1**	E1b1b	100	E1b1b	100	E	99,6
C611AKN	**E1b1b1b2a1**	E1b1b	96,8	E1b1b	79,09	ERRO	
CP446TEO	**E1b1b1b2a1**	E1b1b	100	E1b1b	95,64	ERRO	
PK256GMS	**E1b1b1b2a1**	E1b1b	86,3	INCERTEZA	<50	E,N,Q	67,1
C648PQM	**E2b1**	Q	98,7	E2	100	E,N,Q	97,2
C629GAV	**G**	G2a	100	G2a2b2a1c	86,81	G,I	100
CP315WMA	**G-M201**	G2a	100	G2a2b2a1c	50,62	G,I	99,9
CP456CLO	**G-M201**	G2a	100	INCERTEZA	<50	G,I	79,7
P121OMA	**G-M201**	G2a	100	INCERTEZA	<50	G,I	99,1
P196PBA	**G-M201**	G2a	100	G2a2	99,98	G,I	99
P262OBL	**G-M201**	G2a	100	G2a2b1	99,99	E,G,I	88,2
P523JAR	**G-M201**	G2a	100	G2a2	75,56	G,I	99,6
CP413FFS	**I1**	I1	100	I1	50,79	G,I	88,1
CP461BEM	**I1**	I1	100	INCERTEZA	<50	G,I	76,8
P324DPS	**I1**	I1	100	INCERTEZA	<50	G,I	78,5
PK265MLP	**I1 M253**	I1	100	I1	55,1	G,I	89,8
CP365JNU	**I1a1b1**	I1	100	INCERTEZA	<50	G,I	91,6
P504ABA	**I1a1b1**	I1	100	INCERTEZA	<50	G,I	91,7
C676WBS	**I2a1b1**	I2b1	100	I2a2a	77,95	INCERTEZA	<50
CP331TRR	**I2a1b1**	I2a (xI2a1)	99,7	I2a1	95,55	INCERTEZA	<50

ID	HAPLOGRUPO Y-SNP (Joerin-Luque et al, 2022)	HAPEST	PROBABILIDADE HAPEST (%)	NEVGEN	Probabilidade NEVGEN (%)	STRUCTURE (K=20)	PROBABILIDADE STRUCTURE (%)
CP335JPS	**I2a1b1**	I2b1	99,7	I2a2a	63,76	INCERTEZA	<50
P364INS	**I2a1b1**	I2a (xI2a1)	87,5	I2a2a	73,54	INCERTEZA	<50
P389DVC	**I2a1b1**	I2a (xI2a1)	99,9	I2a1	95,77	ERRO	
P123JNO	**I2a1b1 M223**	I2b1	100	I2a2a	77,39	INCERTEZA	<50
C646AGO	**I2a1b2a**	I2a (xI2a1)	91,9	I2a2b	77,21	INCERTEZA	<50
P283DGO	**I2a1b2a**	I2a (xI2a1)	91,9	I2a2b	77,21	INCERTEZA	<50
CP439JCO	**J2a1a**	J1	99,4	J2a1	70,83	H,J,T	70,2
311RVS	**J2a1a1a2b2**	J1	100	J1a	58,36	J	76,2
637CJB	**J2a1a1a2b2**	J1?	100	INCERTEZA	<50	J	92
C679IVS	**J2a1a1a2b2**	J1	95,4	J2a1	62,91	H,J,T	98,4
CP347NRA	**J2a1a1a2b2**	J1	100	J1a	70,89	J	94
CP462IVA	**J2a1a1a2b2**	J1	100	INCERTEZA	<50	J	81,2
P062FF	**J2a1a1a2b2**	J1	99,9	J2a1	100	H,J,T	65
P147CAS	**J2a1a1a2b2**	J1	100	INCERTEZA	<50	INCERTEZA	<50
P235VJS	**J2a1a1a2b2**	E1b1b	85	J2a1	62,48	H,J,T	88,1
P287EBE	**J2a1a1a2b2**	J1	100	INCERTEZA	<50	J	82,8
P339OSS	**J2a1a1a2b2**	J1	100	INCERTEZA	<50	J	99,7
PK269CJF	**J2a1a1a2b2**	I1	100	INCERTEZA	<50	J	99,4
CP490UMR	**J2a1a1a2b2 M67**	INCERTEZA	<50	INCERTEZA	<50	H,J,T	96,9
369GAC	**J2b2a**	J2b	100	J2b2a	91,16	H,J,T	96
CP470CMB	**J2b2a**	J2b	100	J2b2a	91,31	H,J,T	73,8
PK251DDF	**L1**	L	100	L1a	100	L	91,7
C647JFM	**Q1b1a1a**	Q	97	Q	62,54	E,N,Q	99,3
C656ITS	**Q1b1a1a**	Q	99,8	INCERTEZA	<50	E,N,Q	95,2
CP466MMR	**Q1b1a1a**	Q	100	INCERTEZA	<50	E,N,Q	53
GRC134	**Q1b1a1a**	Q	99,9	INCERTEZA	<50	E,N,Q	63,4
GRC145	**Q1b1a1a**	Q	99,9	INCERTEZA	<50	E,N,Q	63,4
GRC150	**Q1b1a1a**	Q	96,2	INCERTEZA	<50	E,N,Q	73,4
KIV016	**Q1b1a1a**	Q	97,7	Q	93,78	E,N,Q	67,8
KRC064	**Q1b1a1a**	Q	99,8	Q	93,78	E,N,Q	93,2
KRC065	**Q1b1a1a**	Q	100	Q	97,49	E,N,Q	96,3
KRC070	**Q1b1a1a**	Q	100	Q	97,49	E,N,Q	89,2
P127ECO	**Q1b1a1a**	Q	99,8	Q	93.46	E,N,Q	90
P280MME	**Q1b1a1a**	Q	99,5	Q	74,17	E,N,Q	87,2
P288JSV	**Q1b1a1a**	Q	100	Q	99,24	E,N,Q	99,8
P294NAS	**Q1b1a1a**	Q	99,9	Q	85,95	E,N,Q	70,1
P298MNC	**Q1b1a1a**	Q	100	INCERTEZA	<50	ERRO	
P309APE	**Q1b1a1a**	Q	100	INCERTEZA	<50	E,N,Q	98,4
P360JIS	**Q1b1a1a**	Q	99,6	Q	85,95	E,N,Q	85,5
PK260SRO	**Q1b1a1a**	Q	100	INCERTEZA	<50	E,N,Q	91,7
PK262JBO	**Q1b1a1a**	Q	100	INCERTEZA	<50	E,N,Q	69,9
C692LHF	**R1a1a1**	R1a	99,9	R1a	99,98	RA	88,6
P179ASC	**R1a1a1**	R1b	100	R1b	99,92	R1b	100
P243ERS	**R1a1a1**	R1a	100	R1a	100	RA	100
P348NPS	**R1b1a1b**	INCERTEZA	<50	INCERTEZA	<50	RB	82
C687KOA	**R1b1a1b1**	R1b	100	R1b	100	RB	100
CP322LSO	**R1b1a1b1**	R1b	100	R1b	99,31	RB	99,6
CP431ICV	**R1b1a1b1**	R1b	100	R1b	99,84	RB	95,4
P355MRB	**R1b1a1b1**	R1b	100	R1b	99,84	RB	99,8
P203JCA	**R1b1a1b1a**	J1	98	INCERTEZA	<50	ERRO	
C601VRD	**R1b1a1b1a1**	R1b	100	R1b	99.94	RB	92,9
C607LFS	**R1b1a1b1a1**	R1b	100	R1b	100	RB	99,9
C640RLO	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
C659MBF	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
C675WRB	**R1b1a1b1a1**	R1b	100	R1b	100	RB	99,9
C681VAS	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
C702FRE	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
CK500HLJ	**R1b1a1b1a1**	R1b	100	R1b	99.96	RB	99,9
CK507RSC	**R1b1a1b1a1**	R1b	100	R1b	100	RB	99,7
CK510DBG	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
CP303FSG	**R1b1a1b1a1**	R1b	100	R1b	99,28	RB	99,7
CP311MON	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
CP313SFS	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
CP330ETA	**R1b1a1b1a1**	R1b	100	R1b	100	RB	99,1
CP342ECS	**R1b1a1b1a1**	R1b	100	R1b	99,99	RB	99,8

ID	HAPLOGRUPO Y-SNP (Joerin-Luque et al, 2022)	HAPEST	PROBABILIDADE HAPEST (%)	NEVGEN	Probabilidade NEVGEN (%)	STRUCTURE (K=20)	PROBABILIDADE STRUCTURE (%)
CP348JCP	**R1b1a1b1a1**	R1b	100	R1b	98,41	RB	94,2
CP353BMR	**R1b1a1b1a1**	R1b	100	R1b	100	RB	90,7
CP355GVI	**R1b1a1b1a1**	R1b	100	R1b	98,55	RB	99,8
CP357DPN	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
CP412CRR	**R1b1a1b1a1**	R1b	100	R1b	100	RB	88,4
CP420AJO	**R1b1a1b1a1**	R1b	100	R1b	99,69	RB	87,9
CP454PES	**R1b1a1b1a1**	R1b	100	R1b	99,96	RB	94,5
CP459LLF	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
CP460UGO	**R1b1a1b1a1**	R1b	100	R1b	99,79	RB	90,8
CP463AUP	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
CP472AFM	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
CP478DMP	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
CP485JDL	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
P065JJS	**R1b1a1b1a1**	R1b	100	R1b	94,89	RB	96,5
P066RCS	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
P130JGR	**R1b1a1b1a1**	R1b	100	R1b	100	RB	99,9
P150PFN	**R1b1a1b1a1**	R1b	100	R1b	99,96	RB	98
P171ELI	**R1b1a1b1a1**	R1b	99,9	R1b	99,31	RB	69,8
P213JPS	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
P214EJC	**R1b1a1b1a1**	R1b	100	R1b	99,94	RB	99,9
P223MRO	**R1b1a1b1a1**	R1b	100	R1b	99,97	RB	99,8
P252MMW	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
P257ESP	**R1b1a1b1a1**	R1b	100	R1b	99,85	RB	100
P274AGN	**R1b1a1b1a1**	R1b	100	R1b	100	RB	99,7
P275APA	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
P305FPF	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
P322JCT	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
P327TMA	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
P328ESQ	**R1b1a1b1a1**	R1b	99,7	R1b	99,62	RB	95
P352UDS	**R1b1a1b1a1**	R1b	100	R1b	100	RB	99,4
P353JME	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
P365LHS	**R1b1a1b1a1**	R1b	99,7	R1b	100	RB	97,3
P386JGY	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
P388MCY	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
PK208MSM	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
PK212GZE	**R1b1a1b1a1**	R1b	100	R1b	100	RB	96,7
PK232DCM	**R1b1a1b1a1**	R1b	100	R1b	100	RB	100
P212COG	**R1b1a1b1a1a1**	R1b	100	R1b	100	RB	100
P311ROB	**R1b1a1b1a1a1**	R1b	100	R1b	100	RB	100
P354GGS	**R1b1a1b1a1a1**	R1b	100	R1b	100	RB	99,2
PK248CGC	**R1b1a1b1a1a1**	R1b	99,8	R1b	100	RB	97,3
CP477JFS	**R1b1a1b1a1a1c2b1**	R1b	100	R1b	100	RB	99,9
CP370RBR	**R1b1a1b1a1a2b**	R1b	100	R1b	100	RB	100
P192NRR_2	**R1b1a1b1a1a2b**	R1b	100	R1b	99,96	RB	99,8
P297MRO	**R1b1a1b1a1a2b**	R1b	100	R1b	99,6	RB	62,5
CP314PMW	**R1b1a1b1a1a2b1a1**	R1b	100	R1b	100	RB	98,2
P326AAA	**R1b1a1b1a1a2b1a1**	R1b	100	R1b	100	RB	100
C618JPS	**R1b1a1b1a1a2c1**	R1b	100	R1b	100	RB	100
C657FMM	**R1b1a1b1a1a2c1**	R1b	100	R1b	100	RB	99,9
C686JMS	**R1b1a1b1a1a2c1**	R1b	100	R1b	100	RB	100
C688SBA	**R1b1a1b1a1a2c1**	R1b	100	R1b	100	RB	100
CP356FFB	**R1b1a1b1a1a2c1**	R1b	100	R1b	100	RB	100
CP358BRM	**R1b1a1b1a1a2c1**	R1b	100	R1b	100	RB	100
CP467EDS	**R1b1a1b1a1a2c1**	R1b	100	R1b	100	RB	100
P148CPS	**R1b1a1b1a1a2c1**	R1b	100	R1b	100	RB	100
P241SRM	**R1b1a1b1a1a2c1**	R1b	100	R1b	99,88	RB	92,8
P302JDD	**R1b1a1b1a1a2c1**	R1b	100	R1b	100	RB	99,7
PK270RPS	**R1b1a1b1a1a2c1**	R1b	100	R1b	100	RB	99,8
CP458DPD	**T1a**	E1b1b	100	T	99,74	H,J,T	59,8
640UCX	**UNDEFINED**	R1b	100	R1b	100	RB	100