UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE GENÉTICA E MORFOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA ANIMAL

VALÉRIA MIRANDA AMORIM

Avaliação comparativa de procedimentos para imputação de mitogenomas para inferência de haplogrupos mitocondriais em populações miscigenadas brasileiras (Kalunga e Brasília)

Brasília
2026

Valéria Miranda Amorim

Avaliação comparativa de procedimentos para imputação de mitogenomas para inferência de haplogrupos mitocondriais em populações miscigenadas brasileiras (Kalunga e Brasília)

Dissertação de Mestrado apresentada ao programa de Pós-graduação em Biologia Animal, PPG BioAni, da Universidade de Brasília, como parte dos requisitos necessários à obtenção do título de Mestre em Biologia Animal.

Orientadora: Silviene Fabiana de Oliveira

Brasília

2026

Dedico à minha avó, que aos 95 anos segue sendo fonte de inspiração, força e sabedoria. Ao longo de toda a minha vida, foi quem mais incentivou e acreditou na importância dos meus estudos, mesmo nos momentos em que eu mesma duvidei. Falar sobre herança mitocondrial sem reconhecê-la seria um erro, pois nela se encontram não apenas as origens biológicas que atravessam gerações, mas também o exemplo de perseverança e amor pelo conhecimento que moldaram quem sou.

AGRADECIMENTOS

Lembro que, quando criança, minha mãe me dizia: “Antes de tudo vem o estudo” e, dentre tantas coisas vivenciadas até hoje, esse ensinamento nunca se apagou dentro de mim. Se hoje chego até este momento, é graças a todos vocês. Minha imensa gratidão a todos.

Às minhas gatinhas, Lilica e Vitória.

Nessa, nem toda a gratidão do mundo seria capaz de mensurar o quanto você fez por mim. Você é a maior inspiração da minha vida e espero orgulhá-la assim como você me orgulha. Obrigada por sempre acreditar nos meus sonhos e por tornar possível que eu chegasse até aqui.

À minha avó, quero dizer que você me dá força para viver. Que bom é poder tê-la ao meu lado em todos os momentos, sempre me ajudando, sem medir esforços, para me ver feliz.

Às minhas amigas Dani, Isa e Thay, mesmo nas tantas ausências devido à vida acadêmica, sempre tive a certeza de que poderia contar com vocês. Ainda que não estivéssemos fisicamente juntas, nunca faltaram mensagens, apoio e torcida.

À Sil, que me deu a honra de ser sua orientanda e nunca desacreditou de mim. Há uma frase de Guimarães Rosa que sempre me faz lembrar de você: “*Mestre não é quem sempre ensina, mas quem de repente aprende*”. Você sempre me instigou a buscar o conhecimento, a aproveitar as oportunidades e a não ter medo de ir além — e isso diz muito sobre quem você é. Espero levar comigo tudo o que aprendi com você. Minha eterna gratidão.

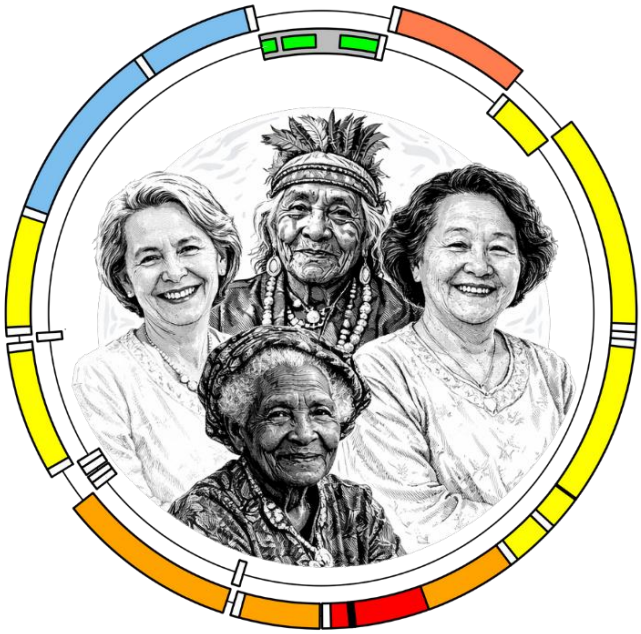
A todos os ex-alunos do Laboratório de Genética Humana, este trabalho só foi possível graças a vocês. Em especial, à Luciana Escher e Sabrina Paiva, responsáveis pelos dados aqui utilizados, o meu muitíssimo obrigada.

Ao Dr. Alexandre C. Pereira, do Laboratório de Genética e Cardiologia Molecular, Instituto do Coração, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brasil, pela realização dos arrays utilizados neste estudo.

Agradeço a todos do laboratório que pude conhecer e com quem convivi — pelas conversas mais aleatórias, pelo aprendizado e pela ajuda. Em especial à Samy, que me acolheu desde o início e foi extremamente carinhosa e paciente comigo. Hoje você é uma grande amiga, e sou profundamente grata por isso.

À CAPES e ao DPG/UnB, pelo suporte que me permitiu seguir na vida acadêmica de forma digna e com dedicação.

À banca examinadora, por aceitarem o convite para compor este momento tão importante da minha trajetória. Agradeço pelas contribuições e pelo aprimoramento deste trabalho.



Vó, como cê conseguiu criar 3 mulheres sozinha?
Na época que mulher não valia nada?
Menina na cidade grande, no susto viúva
E daquela cor que só serve pra ser abusada
Você não costurou só roupa, né
Teve que costurar um mundo
De trauma, abdicação, luta
Pra hoje falar com orgulho
Que essa família não tem vagabundo

Bença, Djonga.

RESUMO

A imputação de mitogenoma é uma técnica cujo objetivo é inferir dados de marcadores ausentes a partir de conjuntos de dados com baixa cobertura, como os derivados de *arrays* de genotipagem, porém para obter resultados informativos essa metodologia requer painéis de referências. Entretanto, existe linhagens, como a indígena, que permanece sub-representada nesses painéis de referências. A população brasileira, altamente miscigenada, abriga uma enorme diversidade genética, dado todo processo histórico e demográfico, sendo valiosa para a genética de populações que utiliza como ferramenta crucial o mitogenoma. O DNA mitocondrial possui herança uniparental, o que permite investigar processos históricos de formação, migração e estruturação populacional. Além disso, a inferência de haplogrupos mitocondriais é amplamente utilizada na caracterização da diversidade e da estrutura genética. Deste modo, o presente estudo teve como objetivo avaliar o desempenho da imputação de mitogenoma na caracterização da diversidade e estrutura genética materna, por meio da comparação entre duas populações brasileiras com histórias demográficas contrastantes. Foram analisados 175 indivíduos, sendo 105 da população de Brasília e 70 da comunidade quilombola Kalunga. A inferência inicial dos haplogrupos foi realizada a partir de um conjunto reduzido de SNPs utilizando o HaploGrep, seguida pela imputação do mitogenoma por meio do *software* Beagle 5.4 e da *pipeline* MitoImpute (Impute2), a fim de aprimorar as atribuições filogenéticas e avaliar o desempenho da imputação sobre qualidade destas. Como resultado, ambas imputações aumentaram o número de sítios informativos e melhoraram a qualidade das inferências dos haplogrupos, com escores médios elevados em ambas as populações. A comparação entre os métodos revelou desempenho semelhante quanto à taxa de atribuição dos haplogrupos, embora o MitoImpute tenha apresentado maior consistência nos escores de qualidade. A análise da composição genética evidenciou predominância de linhagens maternas africanas na comunidade Kalunga, refletindo seu histórico de formação e isolamento relativo, enquanto a população de Brasília apresentou um perfil altamente miscigenado, com maior contribuição europeia, além de componentes africanos e indígenas, e participação minoritária de linhagens asiáticas. As análises de estrutura populacional indicaram diferenciação genética moderada entre as coortes, compatível com suas trajetórias demográficas distintas. Em conjunto, os resultados demonstraram que a imputação pode ser uma eficiente ferramenta para uso em populações miscigenadas e populações mais homogêneas, ainda que a classificação dos haplogrupos ainda seja desafiadora, devido ao baixo número de estudos populacionais que incluem dados brasileiros para classificação no PhyloTree.

Palavras-chave: DNA mitocondrial; mitogenoma; quilombolas; brasileiros; imputação; HaploGrep.

ABSTRACT

Mitogenome imputation is a technique aimed at inferring missing marker data from low-coverage datasets, such as those derived from genotyping arrays; however, to generate informative results, this approach requires appropriate reference panels. Admixed populations, such as the Brazilian population, remain underrepresented in these panels. The Brazilian population, characterized by extensive admixture, harbors remarkable genetic diversity as a consequence of its complex historical and demographic processes, making it particularly valuable for population genetics studies in which the mitogenome constitutes a crucial analytical tool. Mitochondrial DNA exhibits uniparental inheritance and high variability, enabling the investigation of historical processes of population formation, migration, and genetic structuring. In addition, mitochondrial haplogroup inference is widely used to characterize genetic diversity and population structure. Thus, the present study aimed to evaluate the performance of mitogenome imputation in the characterization of maternal genetic diversity and population structure by comparing two Brazilian populations with contrasting demographic histories. A total of 175 individuals were analyzed, comprising 105 from the Brasília population and 70 from the Kalunga quilombola community. Initial haplogroup inference was performed from a reduced set of SNPs using HaploGrep, followed by mitogenome imputation with Beagle 5.4 and the MitoImpute pipeline (Impute2), in order to refine phylogenetic assignments and assess the impact of imputation on inference quality. Both imputation approaches increased the number of informative sites and improved haplogroup inference quality, yielding higher mean quality scores in both populations. Method comparison revealed similar performance in terms of haplogroup assignment rates, although MitoImpute showed greater consistency in quality scores. Analysis of genetic composition revealed a predominance of African maternal lineages in the Kalunga community, reflecting its historical formation and relative isolation, whereas the Brasília population displayed a highly admixed profile, with a greater European contribution alongside African and Indigenous components and a minor proportion of Asian lineages. Population structure analyses indicated moderate genetic differentiation between the cohorts, consistent with their distinct demographic trajectories.Taken together, these results demonstrate that imputation can be an efficient tool for application in both admixed and more homogeneous populations, although haplogroup classification remains challenging due to the limited number of population studies incorporating Brazilian datasets into PhyloTree.

Keywords: mitochondrial DNA; mitogenome; quilombola populations; Brazilians; imputation; HaploGrep.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLA

ABA	Associação Brasileira de Antropologia
AFR	African superpopulation (Superpopulação Africana – 1000 Genomes Project)
AMOVA	Análise de Variância Molecular
AMR	Admixed American superpopulation (Superpopulação Americana – 1000 Genomes Project)
AP	Antes do Presente
BED	Binary PED file
BIM	Binary MAP file
BSB	Brasília
CAAE	Certificado de Apresentação para Apreciação Ética
CEP	Comitê de Ética em Pesquisa
CONEP	Comissão Nacional de Ética em Pesquisa
CRS	Cambridge Reference Sequence
DF	Distrito Federal
DF1	Função Discriminante 1
DNA	Ácido Desoxirribonucleico
DP	Depth of Coverage
EAS	East Asian superpopulation (Superpopulação do Leste Asiático – 1000 Genomes Project)
EMPOP	EDNAP Mitochondrial DNA Population Database

EUR	European superpopulation (Superpopulação Europeia – 1000 Genomes Project)
FAM	Family file
FST	Índice de Fixação
GEN	Genotype file
GST	Coeficiente de Diferenciação Genética
GQ	Genotype Quality
Hd	Diversidade haplotípica
HMM	Hidden Markov Model
HV1	Região Hipervariável 1
HV2	Região Hipervariável 2
HV3	Região Hipervariável 3
IMPUTE2	Software de imputação genotípica
INFO	Escore de informação da imputação
KAL	Kalunga
k	Número médio de diferenças nucleotídicas
LD	Desequilíbrio de Ligação
L-strand	Cadeia leve do DNA mitocondrial
MAC	Minor Allele Count
MAF	Minor Allele Frequency
MITOMAP	Human Mitochondrial Genome Database

MT	Cromossomo mitocondrial
MT-CYB	Gene mitocondrial Cytochrome b
MT-ND	Genes mitocondriais NADH desidrogenase
mtDNA	DNA mitocondrial
mtSNV	Variante de Nucleotídeo Único Mitocondrial
OH	Origem de Replicação da Cadeia Pesada
OL	Origem de Replicação da Cadeia Leve
PCA	Principal Component Analysis
PCo1	Primeira Coordenada Principal
PCo2	Segunda Coordenada Principal
PCoA	Principal Coordinates Analysis
pb	Pares de bases
PERMANOVA	Permutational Multivariate Analysis of Variance
π	Diversidade nucleotídica
rCRS	Revised Cambridge Reference Sequence
R^2	Coeficiente de Determinação
rRNA	RNA ribossomal
SAMPLE	Sample file
SAS	South Asian superpopulation (Superpopulação do Sul Asiático – 1000 Genomes Project)
SNP	Single Nucleotide Polymorphism

SNV	Single Nucleotide Variant
tRNA	RNA transportador
UnB	Universidade de Brasília
VCF	Variant Call Format
ΦST	Índice de Diferenciação Genética baseado em Distância

SUMÁRIO

1 CONTEXTUALIZAÇÃO

Muitas vezes, a elucidação da história de formação das populações pode ser alcançada por meio da compreensão da diversidade genética. A genética de populações, ao analisar a variabilidade genética, permite a compreensão de eventos demográficos passados e da dinâmica populacional, englobando eventos como expansão, migração e fluxo gênico. Entre os marcadores empregados na genética de populações, o DNA mitocondrial (mtDNA), em particular, destaca-se devido à sua herança exclusivamente materna, ausência de recombinação e alta resolução filogenética, tornando-se valioso para investigar linhagens ancestrais e entender a estrutura genética das populações humanas ao longo do tempo e do espaço.

A análise do mtDNA leva à inferência de haplogrupos e à subsequente classificação filogenética. O *software* HaploGrep é uma ferramenta desenvolvida para executar essas tarefas, visto que compara variantes observadas com árvores de referência, como a PhyloTree. No entanto, a precisão na determinação de haplogrupos depende da densidade e distribuição das variantes mitocondriais nos dados analisados. Em muitos estudos populacionais, especialmente os que utilizam *arrays* de genotipagem, a cobertura limitada do mitogenoma resulta em poucos SNPs (*Single Nucleotide Polymorphism*), o que compromete a classificação filogenética e pode gerar escores de confiança baixos ou imprecisos.

Diante disso, a imputação do mitogenoma destaca-se como uma forma para superar as limitações da baixa cobertura de variantes mitocondriais. Essa técnica fundamenta-se na inferência de genótipos ausentes a partir de painéis de referência que incluem mitogenomas inteiros, aumentando significativamente a densidade de marcadores e, consequentemente, a eficácia nas atribuições filogenéticas. Entretanto, o desempenho da imputação está correlacionado à representatividade populacional desses painéis de referência, podendo introduzir vieses quando aplicada a populações miscigenadas ou historicamente sub-representadas.

A população brasileira, por sua vez, é caracterizada por intensa miscigenação, resultado de processos históricos complexos. Possui uma composição genética, composta por diferentes linhagens mitocondriais: europeia, africana, indígena e asiática. Isso confere ao Brasil uma das maiores diversidades genéticas do mundo, tornando-o particularmente relevante para estudos de genética de populações com análise de mtDNA. Além disso, os indivíduos brasileiros

permanecem sub-representados em painéis mitogenômicos globais, o que agrava as limitações para a classificação filogenética dos haplogrupos mitocondriais de linhagem indígena.

As comunidades quilombolas possuem um padrão característico de composição genética, devido a processos históricos marcados por um relativo isolamento geográfico e social. Embora sejam escassos os estudos sobre a diversidade e estrutura mitocondrial em populações quilombolas, já foi descrito que essas comunidades apresentam elevada contribuição de linhagens africanas em comparação com o observado em populações urbanas brasileiras. A comunidade quilombola Kalunga, localizada a aproximadamente 350 km de Brasília, destaca-se por seu tamanho populacional, trajetória histórica e relativa continuidade territorial.

Por outro lado, a população de Brasília passou por um processo de formação mais tardio, por meio de intensos fluxos migratórios internos, resultando em um perfil genético altamente miscigenado, considerado representativo da diversidade genética brasileira. A comparação entre essas duas populações, marcadas por trajetórias demográficas distintas, oferece uma oportunidade valiosa para investigar como diferentes contextos históricos e populacionais moldam a diversidade e a estrutura genética materna.

Diante desse cenário, o presente estudo teve como objetivo avaliar o desempenho da imputação do mitogenoma na caracterização da diversidade e da estrutura genética materna, por meio da comparação entre dados observados e imputados para indivíduos da comunidade Kalunga e da população de Brasília. Inicialmente, a inferência de haplogrupos mitocondriais foi realizada a partir de um conjunto reduzido de SNPs, utilizando o HaploGrep, seguida da imputação do mitogenoma por meio das ferramentas Beagle e IMPUTE2 (*pipeline* MitoImpute). Essas abordagens permitiram avaliar o impacto da imputação na qualidade das inferências filogenéticas, bem como investigar padrões de ancestralidade materna e de estrutura genética populacional por meio de análises estatísticas. Ao integrar uma perspectiva metodológica e populacional, este estudo buscou ampliar o conhecimento sobre a diversidade mitocondrial em populações brasileiras e avaliar a aplicabilidade da imputação do mitogenoma em contextos de miscigenação e de baixa representatividade em painéis de referência globais.

2 REVISÃO TEÓRICA

2.1 DNA mitocondrial humano: características e aplicações

O DNA mitocondrial (mtDNA) foi identificado e isolado primeiramente em 1963, quando Margit Nass e Sylvan Nass, ao estudarem fibras nas mitocôndrias, observaram um comportamento de fixação, estabilização e coloração, que pareciam relacionadas ao DNA (NASS; NASS, 1963). Porém, somente duas décadas depois, em 1981, a sequência completa do mtDNA humano foi publicada, estabelecendo a Sequência de Referência de Cambridge (*Cambridge Reference Sequence* - CRS), marco que a consagrou como o primeiro genoma humano integralmente sequenciado (ANDERSON *et al*., 1981). Esta sequência foi, posteriormente, revisada (rCRS) e permanece como referência para a análise filogenética.

Estruturalmente, o mtDNA humano é uma molécula circular, em fita dupla, de aproximadamente 16.569 pares de bases, embora pequenas variações neste número possam ocorrer devido a inserções ou deleções (HOLLAND; PARSONS, 1999), e fica localizado dentro de organelas celulares chamadas mitocôndrias, que se encontram no citoplasma das células eucarióticas. Convencionada por Anderson *et al*. (1981).

O genoma mitocondrial é compacto, ausente de íntrons, sendo que cerca de 90% da sequência corresponde a genes funcionais. Sua organização é notável pela proximidade e até sobreposição de genes, onde um mesmo nucleotídeo pode atuar como códon de terminação de um gene e de iniciação do seguinte, maximizando a informação codificada (ANDERSON *et al*., 1981; OJALA *et al.*, 1981). Além disso, é dividido em duas cadeias assimétricas: a cadeia pesada (*H-strand*), rica em purinas, e a cadeia leve (*L-strand*), complementar e rica em pirimidinas, cuja diferença na composição de bases define sua densidade molecular (TANAKA; OZAWA, 1994; NONIN-LECOMTE *et al*., 2005).

A região codificadora abriga 37 genes essenciais para o funcionamento das mitocôndrias: 13 codificam subunidades proteicas dos complexos da cadeia respiratória (7 para o Complexo I, 1 para o Complexo III, 3 para o Complexo IV e 2 para o Complexo V), 22 codificam RNAs transportadores (tRNAs) e 2 codificam os RNAs ribossomais (12S e 16S rRNA), fundamentais para a maquinaria de síntese proteica intramitocondrial (ANDERSON *et al*., 1981).

A única região não codificante, denominada Região de Controle ou D-loop (*Displacement Loop*), compreende cerca de 1.122 pb. Esta região abriga os principais elementos

regulatórios, incluindo os promotores de transcrição para ambas as cadeias e as origens de replicação (OH e OL). Por acumular variantes em uma taxa significativamente mais alta, a D-loop contém três segmentos hipervariáveis (HV1, HV2 e HV3), que foram historicamente os primeiros alvos para estudos de diversidade humana e forense, antes dos mitogenomas completos (ANDERSON *et al*., 1981).

Uma das propriedades mais particulares do mtDNA humano, com inúmeras implicações para os estudos populacionais, é o seu padrão de herança estritamente materno. Estabelecido como paradigma após o trabalho seminal de Giles *et al*. (1980), este padrão resulta pelo maior número de cópias de mtDNA no óvulo e da degradação ativa das mitocôndrias paternas presentes no espermatozóide após a fertilização. Consequentemente, as variações no mtDNA não são únicas ao indivíduo, mas são compartilhadas entre todos os descendentes de uma mesma linhagem materna, funcionando como um marcador genético preciso ao longo de gerações (TORRONI *et al*., 2006). Embora casos raros de herança biparental tenham sido documentados, estes constituem exceções que não invalidam a regra, sustentada por mecanismos celulares robustos (LUO *et al*., 2018).

Este modo de herança confere ao mtDNA um valor único para a inferência demográfica. Primeiro, a ausência de recombinação faz com que todo o genoma seja herdado como um único bloco haplóide, simplificando a análise filogenética. Segundo, a sua taxa de mutação é consideravelmente mais alta que a do DNA nuclear, devido em parte à fidelidade mais baixa da polimerase γ mitocondrial e à exposição a espécies reativas de oxigênio no interior da organela. Essa alta taxa gera variabilidade em escalas de tempo evolutivas relevantes para estudos populacionais. Terceiro, o seu tamanho efetivo populacional – que consiste no número de indivíduos que contribuem para a próxima geração – é apenas um quarto do nuclear, tornando-o mais sensível aos efeitos da deriva genética, de gargalos populacionais e de eventos fundadores.

Estas propriedades combinadas transformam o mtDNA em uma ferramenta única para inferir padrões de ancestralidade materna, identificar linhagens populacionais e reconstruir a história evolutiva de diferentes grupos humanos.

2.2 **Haplogrupos mitocondriais e organização filogenética**

O acúmulo de variações nas sequências do mtDNA resultam na variabilidade genética das linhagens maternas, e tem se desenvolvido ao longo do tempo (VAN OVEN; KAYSER,

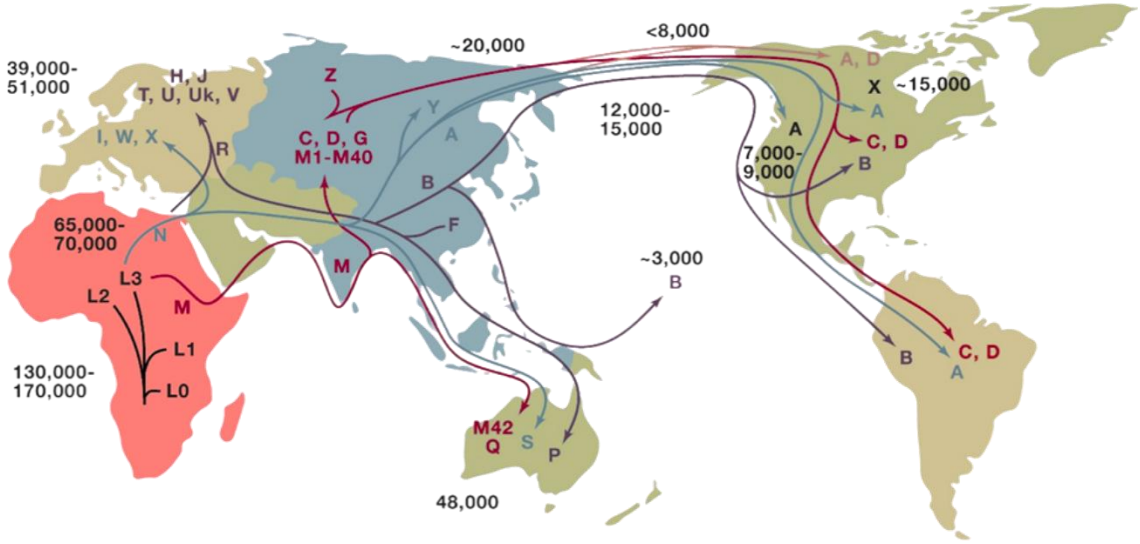
2009). Isso dá origem a marcadores filogeográficos, classificados em haplogrupos. Cada haplogrupo consiste em um conjunto de haplótipos que compartilham características comuns, identificando-os com uma mesma ancestralidade biogeográfica (BUDOWLE *et al*., 2003).

Com intuito de melhorar a análise global das variações do mtDNA, foi construída uma árvore filogenética baseada em dados de sequenciamento completo do genoma mitocondrial disponíveis mundialmente. Esta árvore filogenética inclui haplogrupos previamente publicados, é continuamente atualizada e pode ser consultada no repositório do PhyloTree*mt* (http://www.phylotree.org, VAN OVEN; KAYSER, 2009).

2.2.1 Migrações humanas e os principais haplogrupos mitocondriais

De acordo com os estudos de Vigilant (1991), Cann (1987), Chen (1995) e Watson (1997), o primeiro haplogrupo mitocondrial surge na África, cerca de 40.000–150.000 anos antes do presente (AP), sugerindo que toda a população humana atual descende de um mesmo ancestral. O haplogrupo L dá origem a quatro linhagens: L0, L1, L2 e L3 (130.000–200.000 AP) (CAMPBELL; TISHKOFF, 2008). Enquanto, os dois haplogrupos, M e N, surgiram do haplogrupo africano L3 (65.000–70.000 AP). Conforme foram acontecendo as migrações, o haplogrupo N migrou para a Eurásia e as linhagens do haplogrupo M para a Ásia, dando origem aos haplogrupos A, B, C, D, G e F (TORRONI *et al*., 2006; STEWART; CHINERY, 2015). Na Europa, o haplogrupo N deu origem ao haplogrupo R, que é a raiz dos haplogrupos europeus H, J, T, U e V, que surgiram 39.000–51.000 AP (VAN OVEN; KAYSER, 2008; STEWART; CHINNERY, 2015) e dos haplogrupos S, P e Q são encontrados na Australásia e foram formados 48.000 AP, e os haplogrupos A, B, C e D surgiram menos de 20.000 AP e foram distribuídos pelo Leste Asiático e as Américas. A Figura 2.1 ilustra as migrações através dos continentes e haplogrupos correspondentes.

Figura 1 – Rotas de migrações global e principais haplogrupos mitocondriais humanos.



Fonte: Wallace, 2015.

Nota: Mapa esquemático mostrando a origem africana das linhagens mitocondriais e as rotas migratórias associadas à expansão dos povos para a Eurásia, Oceania e Américas ao longo dos anos. As letras indicam macro-haplogrupos derivados e sua distribuição geográfica aproximada, enquanto as setas representam fluxos migratórios e a cronologia estimada de dispersão em anos antes do presente (AP).

2.2.2 Limitações da inferência de haplogrupos mitocondriais e software HaploGrep

Uma das limitações para inferência de haplogrupos mitocondriais é a ausência de genomas mitocondriais completos ou dados parciais de mtSNV obtidos usando *microarrays*. Isso prejudica a classificação filogenética que depende do reconhecimento de variantes em diferentes ramos e sub-ramos; quando essas não são observadas, torna-se mais difícil distinguir entre subclados próximos ou confirmar um ramo específico. Ferramentas como HaploGrep foram desenvolvidas para automatizar essa comparação entre variantes observadas e a árvore de referência, mas o desempenho cai quando a entrada contém poucos marcadores ou regiões enviesadas do mitogenoma (WEISSENSTEINER *et al*., 2016; SCHÖNHERR *et a*l., 2023).

Sendo assim, conjuntos derivados de *microarrays* de genotipagem ou painéis de baixa cobertura que incluem poucos SNPs mitocondriais — às vezes concentrados em regiões específicas — podem gerar classificações apenas ao nível de macro-haplogrupo (ou a subclados amplos) e aumentar a ambiguidade entre alternativas filogeneticamente próximas. Um efeito comum é o algoritmo identificar um ramo plausível, mas com baixa sustentação, porque faltam variantes essenciais para confirmar ou refutar subclados mais específicos (WEISSENSTEINER *et al.*, 2016).

O HaploGrep utiliza um sistema de pontuação (*quality score*) para refletir o quão bem o conjunto de variantes observado se ajusta às variantes esperadas para um determinado haplogrupo segundo a árvore de referência. Além da atribuição do *"best hit",* a versão atual incorpora recursos de controle de qualidade, destacando variantes ausentes/esperadas, possíveis inconsistências, e outras evidências que afetam a confiança da classificação (WEISSENSTEINER *et al.*, 2016; SCHÖNHERR *et al*., 2023). Para análise metodológica esse escore é crucial para interpretar resultados quando a cobertura é baixa, ou seja, poucos snps genotipados: classificações com escores reduzidos tendem a indicar que o conjunto de variantes é insuficiente para sustentar o haplogrupo determinado, sugerindo cautela na interpretação ou a necessidade de estratégias complementares (como imputação, quando apropriado).

2.3 **Imputação de mitogenoma**

A imputação genômica é uma técnica estatística amplamente utilizada para inferir variantes genéticas não observadas em conjuntos de dados com cobertura horizontal (amplitude) limitada, a partir da informação contida em painéis de referência densamente genotipados ou sequenciados. No contexto da genética de populações, essa metodologia tem sido fundamental para ampliar o poder analítico de dados com baixo nível informativo, permitindo análises comparáveis às obtidas com dados de sequenciamento completo (TRECCANI *et al.*, 2023)**.** Embora originalmente desenvolvida para o genoma nuclear, a imputação tem sido adaptada ao estudo do DNA mitocondrial, com particular relevância em cenários nos quais o número de marcadores mitocondriais observados é reduzido (MCINERNEY *et al*., 2021).

2.3.1 Conceitos gerais de imputação genômica

2.3.1.1 Princípios estatísticos

A imputação genômica baseia-se em modelos probabilísticos que exploram padrões de correlação entre variantes genéticas observadas e não observadas, inferindo genótipos ausentes a partir da similaridade entre alelos do conjunto de estudo e haplótipos presentes em um painel de referência (MARCHINI; HOWIE, 2010). De modo geral, esses modelos utilizam estruturas baseadas em cadeias de *Markov* ocultas (Hidden Markov Models – HMMs), nas quais os

haplótipos observados são representados como mosaicos de haplótipos de referência (LI; STEPHENS, 2003). A probabilidade de cada genótipo imputado é estimada considerando a frequência e a distribuição dos alelos no painel, bem como a consistência do padrão observado ao longo do genoma (BROWNING; BROWNING, 2016).

Esses métodos permitem não apenas a inferência de genótipos ausentes, mas também a atribuição de medidas de incerteza associadas às imputações, o que é essencial para a avaliação da confiabilidade dos resultados em análises subsequentes. Em estudos populacionais, a acurácia da imputação está diretamente relacionada ao tamanho, à diversidade e à representatividade populacional do painel de referência utilizado.

2.3.1.2 Uso de desequilíbrio de ligação

No genoma nuclear, a imputação depende extensivamente do desequilíbrio de ligação (LD), isto é, a associação não aleatória entre variantes localizadas em regiões próximas do genoma (KABISCH, 2017). A eficiência da imputação depende, portanto, da estrutura de LD da população estudada e de sua similaridade com a população representada no painel de referência (BROWNING; BROWNING, 2016).

Embora o conceito de LD seja central na imputação nuclear, sua aplicação ao mtDNA apresenta particularidades importantes, decorrentes das diferenças estruturais e evolutivas entre os dois genomas.

Ao contrário do genoma nuclear, no qual o LD é influenciado por recombinação e segregação independente, no mtDNA as associações entre variantes refletem principalmente a história evolutiva e a estrutura filogenética das linhagens maternas. Consequentemente, a acurácia da imputação mitocondrial está fortemente condicionada ao número de haplogrupos e haplótipos globais.

2.3.1.3 Vantagens e limitações

Entre as principais vantagens da imputação do mitogenoma destaca-se o aumento substancial do número de sítios informativos disponíveis para análise, o que pode resultar em melhorias significativas na classificação filogenética e na inferência de haplogrupos mitocondriais (WEISSENSTEINER *et al*., 2016).

Por outro lado, a imputação mitogenômica apresenta limitações importantes. A principal delas refere-se à sub-representação de populações não europeias em painéis de referência globais, o que pode reduzir a acurácia da imputação em populações miscigenadas ou historicamente isoladas. Além disso, erros sistemáticos podem ocorrer quando variantes raras ou específicas de determinadas linhagens não estão adequadamente representadas no painel, levando a imputações enviesadas ou a classificações filogenéticas incorretas (CROOCK *et al*., 2025).

2.3.2 Métodos utilizados

2.3.2.1 Beagle

O Beagle é um dos *softwares* mais amplamente utilizados para imputação genômica e faseamento de dados genéticos, baseado em modelos probabilísticos eficientes que permitem lidar com grandes conjuntos de dados (BROWNING *et al*., 2018). Embora tenha sido originalmente desenvolvido para imputação de variantes do genoma nuclear, o software Beagle também pode ser aplicado à imputação de variantes mitocondriais quando utilizado com um painel de referência apropriado (DORJI et al., 2024).

Entre as principais vantagens do Beagle destacam-se sua eficiência computacional, flexibilidade na definição de parâmetros e capacidade de lidar com diferentes tamanhos amostrais (BROWNING *et al*., 2018). No contexto do mtDNA, entretanto, seu desempenho depende fortemente da adequação do painel de referência e da densidade de marcadores observados, uma vez que o modelo não foi originalmente otimizado para genomas haplóides não recombinantes.

2.3.2.2 MitoImpute / Impute2

MitoImpute é um fluxo de trabalho automatizado para análise de dados que foi desenvolvido para imputação de genomas mitocondriais usando o *framework* IMPUTE2. Esse pipeline incorpora características específicas do mtDNA, como sua estrutura haplóide e organização filogenética, além de utilizar painéis de referência mitogenômicos completos (MCINERNEY, 2021). Estudos demonstram que essa abordagem pode apresentar desempenho

consistente na imputação de variantes mitocondriais e na melhoria da classificação de haplogrupos (CROOCK, 2025).

2.4 Estatísticas populacionais aplicadas ao mtDNA

A análise estatística de dados mitocondriais é uma etapa central em estudos de genética de populações, pois permite quantificar a diversidade genética, avaliar a estrutura populacional e inferir processos evolutivos e demográficos associados à história das populações. Devido às particularidades do DNA mitocondrial como herança uniparental materna, ausência de recombinação e haploidia, a aplicação e a interpretação das estatísticas populacionais devem considerar pressupostos distintos daqueles empregados para marcadores nucleares biparentais.

Em estudos baseados em mtDNA, as estatísticas populacionais são frequentemente utilizadas para investigar padrões de diferenciação genética entre populações, identificar sinais de estruturação populacional e testar hipóteses relacionadas a eventos demográficos, como expansões populacionais, gargalos e isolamento relativo. Métricas como F_{ST}, AMOVA e testes de neutralidade, incluindo o teste *D* de Tajima (Tajima's *D*), figuram entre as abordagens mais empregadas nesse contexto (HARTL; CLARK, 2010).

A estatística F_{ST} é amplamente utilizada para quantificar o grau de diferenciação genética entre populações, expressando a proporção da variância genética total que pode ser atribuída a diferenças entre grupos. No contexto do mtDNA, o F_{ST} reflete diferenças na distribuição das linhagens maternas entre populações e é particularmente sensível a processos como deriva genética, isolamento geográfico e efeitos fundadores. Valores baixos de F_{ST} indicam populações geneticamente semelhantes em termos de linhagens maternas, enquanto valores mais elevados sugerem maior diferenciação, compatível com histórias demográficas distintas ou com fluxo gênico limitado (WRIGHT, 1951; WRIGHT, 1978).

A Análise de Variância Molecular (AMOVA) constitui uma extensão do conceito de F_{ST}, permitindo a decomposição hierárquica da variância genética em diferentes níveis, como dentro de populações, entre populações de mesmo grupo e entre grupos de populações. Essa abordagem é particularmente útil em estudos mitocondriais, pois possibilita avaliar se a maior parte da variabilidade está concentrada no interior das populações ou se existe estruturação significativa entre grupos definidos a priori, como populações urbanas versus rurais. A AMOVA tem sido amplamente aplicada em estudos com mtDNA para investigar padrões de

estrutura genética associados a fatores históricos, geográficos e socioculturais (EXCOFFIER *et al.*, 1992; EXCOFFIER; LISCHER, 2010).

Além das estatísticas de diferenciação, testes de neutralidade desempenham papel fundamental na interpretação de padrões de variabilidade mitocondrial. O teste *D* de Tajima (Tajima's D) compara duas estimativas da diversidade genética — o número médio de diferenças entre sequências pareadas e o número de sítios segregantes — para detectar desvios em relação ao modelo neutro de evolução. Valores negativos de Tajima's D são geralmente interpretados como indicativos de expansão populacional recente ou seleção purificadora, enquanto valores positivos podem sugerir gargalos populacionais, estruturação populacional ou seleção balanceadora. No contexto do mtDNA, esse teste tem sido utilizado para inferir eventos demográficos associados à história materna das populações humanas (TAJIMA, 1989; ROGERS; HARPENDING, 1992).

A aplicação dessas estatísticas a dados mitocondriais imputados requer atenção adicional, uma vez que a imputação pode influenciar tanto o número de sítios informativos quanto a distribuição das variantes entre indivíduos. Embora o aumento da densidade de marcadores possa melhorar o poder estatístico das análises, é fundamental avaliar se os padrões observados após a imputação são consistentes com aqueles obtidos a partir dos dados originais, evitando interpretações enviesadas decorrentes de artefatos metodológicos. Assim, a combinação de estatísticas populacionais clássicas com a comparação entre dados pré- e pós-imputação representa uma abordagem para investigar a estrutura genética materna e os processos demográficos subjacentes (DE MARINO *et al*., 2022).

2.5 **Populações brasileiras em estudos mitogenômicos**

2.5.1 Histórico Brasileiro

A população brasileira carrega em seu DNA uma imensa diversidade genética como resultado de um processo de miscigenação que se estende por vários séculos. O território que hoje corresponde ao Brasil, no início do século XVI, residiam a aproximadamente 2,5 milhões de indivíduos de grupos indígenas diversos (BETHELL, 1997). Esse panorama foi modificado com o advento das grandes navegações e chegada de colonizadores europeus ao continente.

A presença de europeus no território brasileiro caracterizou-se, nos primeiros séculos, predominantemente por homens portugueses. O segundo maior contingente de migrantes europeus se dá a partir de 1890 pelos italianos, seguidos pelos espanhóis, que migram para o Brasil desde o século XVI, e os alemães, cuja migração se inicia por volta de 1884. De forma menos expressiva, pode-se citar os grupos do Oriente Médio (SANTOS, 2023).

Pesquisas com marcadores uniparentais evidenciam que esse padrão resultou em uma assimetria na contribuição parental, com predominância de linhagens do cromossomo Y de origem europeia, enquanto as linhagens mitocondriais mantiveram maior contribuição indígena e, posteriormente, africana (ALVES-SILVA *et al*., 2000; PENA *et al*., 2011).

Simultaneamente à colonização europeia, o tráfico transatlântico de pessoas escravizadas introduziu no Brasil linhagens provenientes da África Subsaariana, principalmente das regiões do Congo e Angola, além de grupos menores oriundos da Guiné, Gana, Nigéria, Serra Leoa e Moçambique. Do ponto de vista mitocondrial, essa migração forçada foi responsável pela incorporação de linhagens que hoje compõem uma parcela substancial do *pool* genético materno brasileiro (SALAS *et al*., 2004; GONÇALVES *et al*., 2007).

Estudos apontam que houve uma redução de cerca de 90% da população indígena original entre os séculos XVI e XVII, atribuída à violência colonial, epidemias, escravização e deslocamentos forçados (BEDOYA *et al*., 2006; SILVA *et al*., 2022). Apesar desse impacto lastimável, as linhagens indígenas persistiram, especialmente por meio do DNA mitocondrial.

Ao longo dos séculos XIX e XX, novos fluxos ocorreram: houve a imigração europeia tardia e, de forma mais localizada, a imigração asiática - particularmente com a chegada de japoneses a partir de 1908 (NISHIDA, 2017). A maior parte desses imigrantes era composta por camponeses de regiões pobres do norte e sul do Japão, que vieram a trabalho nas prósperas fazendas de café do estado de São Paulo (DAIGO, 2008). Isso resultou, em grande parte, na introdução de linhagens do leste asiático.

A sociedade brasileira possui uma composição genética muito particular, com variações nas proporções de ancestralidade africana, europeia, indígena e asiática observadas entre grandes regiões geográficas, entre estados de uma mesma região e até mesmo dentro de uma mesma cidade, refletindo histórias locais de ocupação, migração interna, urbanização e isolamento geográfico (KEHDY *et al.*, 2015).

2.5.1.1 População de Brasília

A população de Brasília é um modelo valioso para estudos populacionais, pois devido ao seu processo de formação, com indivíduos provenientes de todas as regiões do país, resultou em uma composição genética representativa da população brasileira (SANTOS, 2023).

A cidade de Brasília foi fundada em 1960, após intensos fluxos migratórios internos que aconteceram para sua construção. Pesquisas com a população do Centro-Oeste, região onde Brasília se localiza, evidenciam sua elevada diversidade genética (FREITAS, 2019**;** SANTOS, 2023; JOERIN-LUQUE, 2022). O principal fluxo migratório ocorreu com indivíduos oriundos das regiões Nordeste e Centro-Oeste. Após a consolidação como centro administrativo e político vieram indivíduos do Sul e do Sudeste do país, e uma composição majoritariamente de ancestralidade materna europeia.

Pesquisas com marcadores autossômicos e uniparentais indicam que a população de Brasília apresenta altas proporções de ancestralidade africana, europeia, indígena e asiática, refletindo diretamente a diversidade regional de origem dos seus habitantes. Ademais, o fluxo de migrantes ainda permanece, o que diferencia Brasília de outros regiões do Brasil e contribui para a manutenção dessa diversidade genética e estrutura populacional. (SANTOS, 2023).

2.5.1.2 Comunidade Kalunga

As comunidades quilombolas podem ser entendidas, em termos físicos, como agrupamentos de indivíduos com origem africana. Conforme definição da Associação Brasileira de Antropologia (ABA), o termo “quilombo” abrange todas as comunidades rurais compostas por descendentes de povos escravizados, que vivem em regime de subsistência e cujas manifestações culturais estão profundamente conectadas ao seu passado (PEDROSA, 2006; OLIVEIRA JÚNIOR *et al*., 2000).

Entre as comunidades quilombolas brasileiras, o território Kalunga destaca-se tanto por sua extensão territorial quanto por sua relevância histórica e sociocultural. Reconhecida como o maior quilombo do Brasil, a comunidade Kalunga ocupa uma área aproximada de 253 mil hectares, distribuída em cerca de 56 povoados localizados nos municípios de Cavalcante, Monte Alegre de Goiás e Teresina de Goiás, na região nordeste do estado de Goiás (SILVA, 2016).

Os habitantes dessa comunidade receberam o nome Kalunga ou Calunga — um termo da língua banto, que é falada por diversos povos africanos trazidos durante a diáspora, especialmente aqueles oriundos de Angola, Congo e Moçambique — significando “lugar sagrado” ou “refúgio” (ANAIS..., 2017).

A formação do quilombo Kalunga remonta ao início do século XVIII, período marcado pela expansão da mineração no Planalto Central e pela ocupação do território goiano por bandeirantes de origem luso-brasileira. A partir de 1722, a descoberta de ouro em Goiás intensificou o tráfico interno de pessoas escravizadas, principalmente da costa da África Ocidental, destinadas ao trabalho extenuante nas minas (PALACÍN, 1994; BAIOCCHI, 2006). As condições desumanas enfrentadas pelos escravizados geraram contínuas resistências e fugas para áreas remotas, como vales profundos e serras íngremes; aqueles que conseguiam escapar formavam grupos conhecidos como quilombos.

Nesse contexto, o território Kalunga consolidou-se como espaço estratégico de refúgio, dada sua topografia acidentada e relativa distância dos principais núcleos coloniais. Com o declínio da mineração a partir da segunda metade do século XVIII, especialmente após 1780, muitos dos núcleos coloniais foram abandonados. A retirada dos bandeirantes e mineradores contribuiu para a consolidação definitiva das comunidades quilombolas locais, formadas por ex-escravizados que permaneceram na região e desenvolveram estratégias de subsistência baseadas na agricultura, na coleta, na caça e no uso sustentável dos recursos naturais disponíveis (SOUZA, 2013).

Ao longo do século XIX, a comunidade Kalunga manteve-se relativamente isolada, o que favoreceu a preservação de práticas culturais, religiosas e sociais próprias, mas também impactou a dinâmica demográfica e genética da população. (NUNES *et al*., 2020). Apesar disso, ao longo dos anos essa comunidade construiu relações sociais e econômicas com proprietários rurais da região e/ou com núcleos urbanos regionais próximos (KIMURA *et al*., 2013; GONTIJO *et al*., 2018).

2.6 JUSTIFICATIVA

A população brasileira apresenta uma intensa miscigenação e diversidade demográfica, o que permite constituir um modelo para compreensão genética das populações contemporâneas, apesar de ser ainda sub-representada em estudos com mitogenoma. Somando as comunidades quilombolas permanecem com características de populações menos miscigenadas em comparação a populações urbanas brasileiras, com potencial de portar variantes raras ou ausentes em outras populações. Uma limitação para classificação filogenética na população brasileira está relacionado aos haplogrupos de origem indígenas, como consequência, a investigação da ancestralidade materna e de processos históricos associados à formação, migração e estruturação populacional fica comprometida.

Diante disso, o presente estudo justifica-se, pela necessidade de avaliar o desempenho da imputação do mitogenoma em populações brasileiras com histórias contrastantes. A comparação entre os dados observados e dados imputados, bem como entre diferentes técnicas de imputação, permite não apenas mensurar o ganho metodológico proporcionado pela técnica, mas também avaliar seus impactos na inferência da diversidade e estrutura genética materna.

Adicionalmente, a caracterização das frequências de haplogrupos mitocondriais e a aplicação de análises populacionais – como F_{ST}, AMOVA e testes de neutralidade - contribuem para elucidar os processos históricos que moldaram essas populações. Ao integrar abordagens metodológicas e análises populacionais, este estudo amplia o conhecimento sobre a diversidade mitocondrial brasileira e fornece subsídios relevantes para pesquisas futuras em genética de populações, antropologia genética, saúde e genética forense, além de reforçar a importância da inclusão de populações historicamente sub-representadas em estudos genômicos.

3 OBJETIVOS

Objetivo Geral

Avaliar o desempenho da imputação do mitogenoma na inferência de haplogrupos mitocondriais utilizando, para tanto, dados obtidos para duas populações brasileiras, Kalunga e Brasília, que possuem histórias demográficas e composição genética contrastantes.

Objetivos Específicos

- Inferir haplogrupos mitocondriais com o software HaploGrep3 a partir de um conjunto reduzido de variantes e avaliar a qualidade e a resolução das classificações obtidas antes da imputação;
- Realizar e comparar o desempenho da imputação do mitogenoma utilizando o software Beagle e a *pipeline* MitoImpute, considerando métricas de qualidade e melhoria na atribuição filogenética;
- Comparar as frequências dos haplogrupos mitocondriais entre as duas comunidades e outras populações descritas na literatura;
- Avaliar a estrutura genética e a diferenciação populacional entre as amostras por meio de estatísticas de diversidade e testes populacionais, incluindo F_{ST}, AMOVA e o teste de neutralidade de Tajima (Tajima's D);
- Discutir a relação entre os padrões de diversidade e estrutura genética mitocondrial obtidos e os processos históricos de formação populacional em ambas as amostras.

4 MATERIAIS E MÉTODOS

A coleta de amostras e genotipagem para realização deste presente projeto de pesquisa já havia sido aprovada em seus aspectos éticos pelo Comitê de Ética em Pesquisa CEP/CONEP da Faculdade de Ciências da Saúde da Universidade de Brasília (UnB), de acordo com o Processo CAAE n.º 72917916.3.0000.0030. Foram utilizados dados genotipados disponíveis referentes a 70 indivíduos da comunidade Kalunga e 105 indivíduos residentes no Distrito Federal. As análises dos dados foram realizadas no Laboratório de Genética Humana do Departamento de Genética e Morfologia do Instituto de Ciências Biológicas da UnB.

O DNA das amostras de sangue foi extraído através de protocolo do método Puregene (QUIAGEN). Posteriormente foi feita a quantificação do DNA por meio de ensaios espectrofotométricos (equipamento *Nanodrop* ND-1000, da *Thermo Fisher Scientific*), para a determinação da concentração. Para avaliar eventual degradação, as amostras foram submetidas a gel de agarose a 1% e imagens obtidas por transiluminação. Na etapa seguinte, diluiu-se as mesmas de forma a ajustar a concentração para a etapa de genotipagem, seguindo o protocolo para *SNParray da Affymetrix/Thermo Fisher Scientific.*

As genotipagens foram realizadas no Laboratório de Genética e Cardiologia Molecular, Instituto do Coração, com o *SNPArray* de alta densidade *AxiomTM Genome-Wide Human Origins* (*Thermo Fisher Scientific*), o qual apresenta aproximadamente 600.000 marcadores do tipo SNPs. Esse *SNPArray* foi desenvolvido para estudos de genética populacional em humanos e não apresenta viés de representação de SNPs como os observados naqueles de interesse clínicos/farmacogenéticos usados nos estudos de associação com doenças (LU *et al*., 2011).

4.1 Controle de qualidade inicial e filtragem dos VCFs

Os dados brutos passaram por uma filtragem de qualidade, onde foram avaliadas métricas como profundidade de leitura (DP), qualidade de genótipo (GQ) e taxa de genótipos ausentes. Em seguida, os dados das populações foram armazenados no software *Axiom Analysis Suite*.

Todos os dados foram baixados em um único arquivo em formato *Variant Call Format* (VCF), com dados de DNA mitocondrial e nuclear. Todas as análises foram conduzidas em ambiente Linux, com controle de versões dos softwares e parâmetros padronizados para garantir reprodutibilidade.

4.2 **Preparação dos dados para imputação**

Para as análises mitocondriais, os VCFs foram restringidos ao cromossomo mitocondrial (chr26/MT), assegurando consistência com a referência utilizada (rCRS). Variantes multialélicas, caracterizadas pela presença de mais de um alelo alternativo em uma mesma posição, foram normalizadas e, quando necessário, decompostas em variantes bialélicas, de modo que cada registro representasse apenas um único alelo alternativo. Esse procedimento é importante para padronizar a representação das variantes e assegurar a compatibilidade dos dados com ferramentas de imputação e classificação de haplogrupos.

4.3 **Inferência dos haplogrupos com software HaploGrep 3**

A inferência dos haplogrupos do DNA mitocondrial foi realizada três vezes após processamentos diferentes da amostra: (1) sem imputação, (2) após imputação com Beagle ou (3) após imputação com MitoImpute. Nas inferências sem imputação e com Beagle foi utilizado o software HaploGrep 3 (3.2.1) (WEISSENSTEINER et al., 2023). O software usa informações da árvore filogenética do DNA mitocondrial (http://phylotree.org/) para classificar as variantes genéticas observadas em haplogrupos, além de utilizar o arquivo de entrada no formato VCF, contendo apenas as variantes do mtDNA, as quais foram alinhadas contra a sequência de referência do DNA mitocondrial humano (*Revised Cambridge Reference Sequence* - rCRS, hg38/build38, PhyloTree 17.2) (VAN OVEN, 2015). Quanto à inferência dos dados imputados com o MitoImpute, utilizou-se o software HaploGrep 3 executado em ambiente Linux (Ubuntu), por meio do comando *classify*, com base na PhyloTree *build* 17.2 (VAN OVEN, 2015), e arquivo de entrada VCF.

4.4 **Imputação com software Beagle**

Os dados foram imputados utilizando o software Beagle, com o objetivo de inferir variantes ausentes e aumentar a cobertura do genoma mitocondrial, empregando como painel de referência os dados do Projeto 1000 Genomas. O painel do 1000 Genomas foi previamente preparado para garantir compatibilidade com os dados das amostras estudadas, incluindo correspondência de coordenadas genômicas e orientação de alelos. O Beagle foi executado com parâmetros padrão recomendados para dados de alta densidade, ajustando-se as opções

relacionadas ao número de *threads* e à memória disponível (BROWNING *et al*., 2018). Os arquivos de saída imputados foram avaliados quanto à qualidade da imputação, considerando métricas como probabilidade posterior dos genótipos e consistência dos haplótipos mitocondriais inferidos. Os genótipos imputados foram filtrados, mantendo-se apenas aqueles com probabilidade ≥ 0,90. Adicionalmente, variantes com baixa qualidade de imputação (INFO < 0,8) ou alta taxa de dados ausentes (> 5%) foram excluídas das análises subsequentes, assegurando maior confiabilidade na inferência dos haplótipos mitocondriais.

4.5 **Imputação do DNA mitocondrial com pipeline MitoImpute**

A imputação foi realizada por meio do pipeline MitoImpute, que combina PLINK, IMPUTE2 e rotinas em R para manipulação específica de variantes mitocondriais. O procedimento seguiu as recomendações metodológicas descritas pelos autores do software (MCINERNEY *et al*., 2021).

Os dados de entrada consistiram em arquivos no formato VCF, contendo variantes mitocondriais provenientes de genotipagem por *SNP array*. Inicialmente, os arquivos VCF foram convertidos para o formato binário do PLINK (BED/BIM/FAM) utilizando o software PLINK v1.9, mantendo a ordem original dos alelos (–keep-allele-order) e preservando a codificação alélica conforme o painel de referência do MitoImpute. Para garantir a compatibilidade com o painel de referência mitocondrial, os SNPs foram filtrados de acordo com a lista de variantes presentes no painel ReferencePanelSNPs_MAF ≥ 0,01, fornecido pelo MitoImpute. Essa etapa assegurou que apenas variantes compartilhadas entre os dados amostrais e o painel de referência fossem mantidas para a imputação.

Quando necessário, as posições e alelos foram ajustados para garantir concordância com o sistema de coordenadas rCRS, evitando inconsistências durante a fase de imputação. Após a padronização dos dados, os arquivos binários do PLINK foram convertidos para o formato Oxford GEN/SAMPLE, exigido pelo algoritmo de imputação. Essa conversão foi realizada com o PLINK, preservando a ordem dos alelos e permitindo cromossomos não autossômicos, uma vez que o mtDNA não segue o padrão diploide nuclear.

Os resultados foram sumarizados em arquivos. gen e .info. Após a imputação, foram mantidas apenas variantes com escore de informação (INFO) ≥ 0,3 e frequência alélica imputada entre 0,1% e 99,9%. A concordância entre genótipos observados e imputados para SNPs em comum foi inspecionada como controle interno de qualidade (MARCHINI et al.,

2007). Os arquivos resultantes foram convertidos novamente para o formato VCF e submetidos a etapas adicionais de filtragem e normalização, incluindo a remoção de variantes de baixa qualidade e a padronização da nomenclatura dos alelos. O conjunto final de dados imputados foi então utilizado para análises subsequentes, incluindo a atribuição de haplogrupos mitocondriais e análises populacionais.

4.6 **Análises dos Haplogrupos**

Para cada grupo amostral, foi analisado o resultado quanto aos haplogrupos inferidos, e as contribuições de cada linhagem materna: indígena, europeia, africana e asiática. Em seguida, foi feita uma busca na literatura a fim de avaliar se havia correspondência com pesquisas anteriores.

O HaploGrep apresentou também um conjunto de métricas de controle de qualidade relacionadas às variantes analisadas no genoma mitocondrial. Entre essas métricas estão o número de variantes de entrada, que corresponde ao total de posições variantes presentes no arquivo analisado, e as incompatibilidades de referência, que indicam discrepâncias entre as variantes observadas e a sequência de referência mitocondrial utilizada na classificação de haplogrupos. A sobreposição de árvores (%) refere-se ao grau de correspondência entre as variantes detectadas e aquelas definidas na árvore filogenética de haplogrupos. O relatório também identifica possíveis viradas de *strand* (inversões de fita), variantes fora de alcance da árvore filogenética, variantes multialélicas (posições com mais de dois alelos), variantes do tipo *indel* (inserções ou deleções), além de variantes filtradas no arquivo VCF e variantes duplicadas. Adicionalmente, são apresentados indicadores de qualidade relacionados à cobertura e consistência dos dados, como amostras com baixa taxa de chamada (*low sample call rate*), variantes monomórficas (sem variação entre as amostras) e variantes com taxa de chamada inferior a 90%, que podem indicar baixa confiabilidade genotípica. Essas métricas auxiliam na avaliação da qualidade dos dados e na identificação de possíveis inconsistências antes da interpretação dos haplogrupos mitocondriais.

Cada haplogrupo mitocondrial foi mapeado a um macro-haplogrupo associado a uma origem geográfica predominante. Em particular, haplogrupos da linhagem L (L0, L1, L2, L3) foram classificados como de origem africana; haplogrupos A, B, C, D e X2a como de ancestralidade indígena/nativo americano; haplogrupos H, V, J, T, U, K, I, W e X (exceto X2a)

como de origem europeia; e o haplogrupo M como asiático. Haplogrupos atípicos ou não passíveis de classificação inequívoca foram agrupados na categoria “Outros”.

4.7 **Populações de referência globais**

Para contextualizar os padrões observados em ambos os resultados, foram utilizados dados públicos de mtDNA de populações do projeto 1000 Genomes Phase 3, representando África, Europa, Américas, Leste Asiático e Sul da Ásia (THE 1000 GENOMES PROJECT CONSORTIUM, 2015).

4.8 **Análises estatísticas da diversidade e estrutura genética**

A diversidade genética mitocondrial foi estimada a partir dos mitogenomas imputados utilizando medidas padrão: número de haplótipos distintos, diversidade haplotípica (Hd), diversidade nucleotídica (p) e número médio de diferenças nucleotídicas entre pares de indivíduos. Cálculos foram realizados em R e/ou softwares especializados, utilizando as matrizes de variantes mitocondriais filtradas.

A diferenciação genética entre populações foi quantificada por meio do índice F_{ST} de Weir & Cockerham (WEIR; COCKERHAM, 1984), calculado a partir das frequências de haplogrupos e, em análises complementares, a partir de variação nucleotídica. Intervalos de confiança foram obtidos por permutação ou *bootstrap*.

Adicionalmente, a estrutura da variabilidade genética foi explorada por meio da Análise de Variância Molecular (AMOVA), que permite particionar a variância genética total em componentes atribuíveis às diferenças entre populações e à variabilidade dentro das populações. Essa abordagem possibilitou avaliar a proporção relativa da variabilidade genética associada a diferentes níveis hierárquicos de organização populacional (EXCOFFIER; SMOUSE; QUATTRO, 1992).

Para a análise multivariada dos dados, foi empregada a Análise de Componentes Principais (PCA), aplicada a matrizes de frequências relativas de haplogrupos mitocondriais por população. A PCA foi utilizada como uma ferramenta exploratória para reduzir a dimensionalidade dos dados e visualizar padrões de agrupamento e diferenciação entre as populações em um espaço bidimensional definido pelos dois primeiros componentes principais.

Essa análise permitiu identificar tendências gerais de composição genética materna e relações entre as populações estudadas.

Os resultados estatísticos foram apresentados na forma de tabelas e figuras, incluindo gráficos de barras, dispersões multivariadas e resumos numéricos, elaborados com o auxílio de pacotes gráficos do R.

5 **RESULTADOS E DISCUSSÃO**

5.1 **Software HaploGrep3**

5.1.1 Parâmetros e qualidade da inferência

A análise do mtDNA foi realizada, num primeiro momento, com o *software* HaploGrep, para os dois grupos amostrais: Brasília e Kalunga. O conjunto de entrada continha 240 variantes. Conforme demonstrado na Tabela 1, nenhuma das amostras foi classificada como "aprovada". Além disso, o grupo amostral de Brasília apresentou 29 avisos e 76 falhas, enquanto todo o grupo amostral de Kalunga foi classificado com falhas, isso significa que o conjunto de dados não possuía variantes determinantes suficientes para inferir os haplogrupos com confiança da classificação, mas o software infere de acordo com as probabilidades. Tendo conhecimento das inúmeras citações e *downloads*, o *software* Haplogrep é uma das principais ferramentas para classificação automatizada de haplogrupos mitocondriais, atuando em diversas áreas da pesquisa (SCHONHERR *et al*., 2023). No entanto, foi uma ferramenta desenvolvida para análise de dados de mtDNA completo. Essa limitação acabou dificultando estudos que utilizam genotipagem de baixa resolução, como os do estudo que são oriundos de marcadores genéticos do *array* de alta densidade de SNPs, *Axiom Human Origins – Thermo Fisher Scientific*.

A sobreposição de árvore filogenética foi de aproximadamente 62%, indicando que mais de um terço das variantes usadas para inferir os haplogrupos estavam faltando nos dados brutos. As variantes fora de alcance, multialélicas, *indel*, filtradas ou duplicadas, não foram identificadas, assim como baixa taxa de amostragem ou taxa de chamadas inferior a 90%. Quanto ao número de variantes monomórficas, Brasília apresentou 140 e Kalunga 166.

Tabela 1 – Parâmetros da inferência de haplogrupos mitocondriais pelo HaploGrep3 nas populações de Brasília e Kalunga.

	Brasília	**Kalunga**
Aprovado	0	0
Avisos	29	0
Falhas	76	70
Arquivos	1	1
Amostras	105	70

Variantes de entrada	240	240
Incompatibilidades de referência	59	60
Sobreposição de árvores (%)	62,14	61,17
Inversões	3	3
Variantes fora do alcance	0	0
Variantes multialélicas	0	0
Variantes *Indel*	0	0
Variantes filtradas VCF	0	0
Variantes duplicadas	0	0
Baixa taxa de amostragem	0	0
Variantes monomórficas	140	166
Taxa de chamadas de variantes < 90%	0	0

Fonte: O autor (2026), a partir dos resultados do HaploGrep.

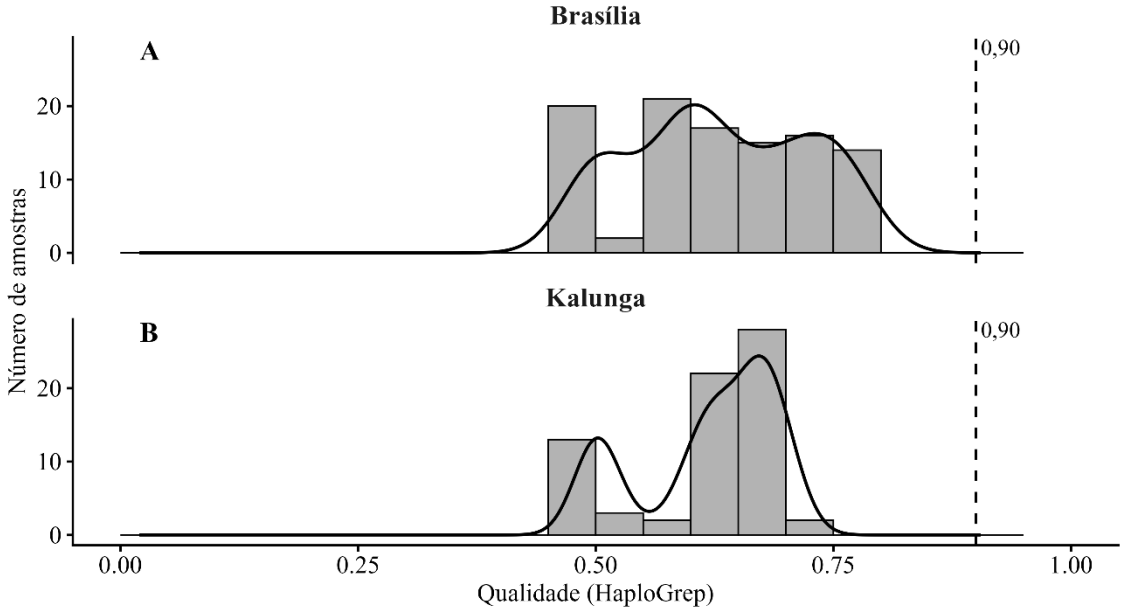
A Tabela 2 traz o escore de qualidade atribuído pelo HaploGrep. Em Brasília, os valores de média, mediana e desvio-padrão foram 0,63, 0,61 e 0,09, respectivamente, enquanto em Kalunga, a distribuição foi menor: média de 0,61, mediana de 0,62 e desvio-padrão de 0,07. Outrossim, em ambos grupos amostrais não foram observadas amostras com escore igual ou superior a 0,90 (Figura 2). Esses resultados estatísticos demonstram que o desempenho é limitado pela baixa densidade de variantes. No contexto brasileiro, Souza *et al*. (2025) obtiveram confiança média de classificação para haplogrupos de origem europeia e africana de ~0,95 e para indígenas de ~0,89. Ou seja, observaram resultados maiores que os analisados. Esses valores indicam baixa confiança na atribuição dos haplogrupos.

Tabela 2 – Estatística descritiva do escore de qualidade da inferência de haplogrupos mitocondriais (HaploGrep) nas populações analisadas.

Escore de qualidade	Brasília	Kalunga
Número de amostras	105	70
Média	0,63	0,61
Mediana	0,61	0,62
Desvio-padrão	0,09	0,07
Mínimo	0,50	0,50
Máximo	0,78	0,70
Proporção de amostras com qualidade ≥ 0,90	0	0

Fonte: O autor (2026).

Figura 2 – Distribuição do escore de qualidade da atribuição de haplogrupos mitocondriais (HaploGrep) em duas populações.

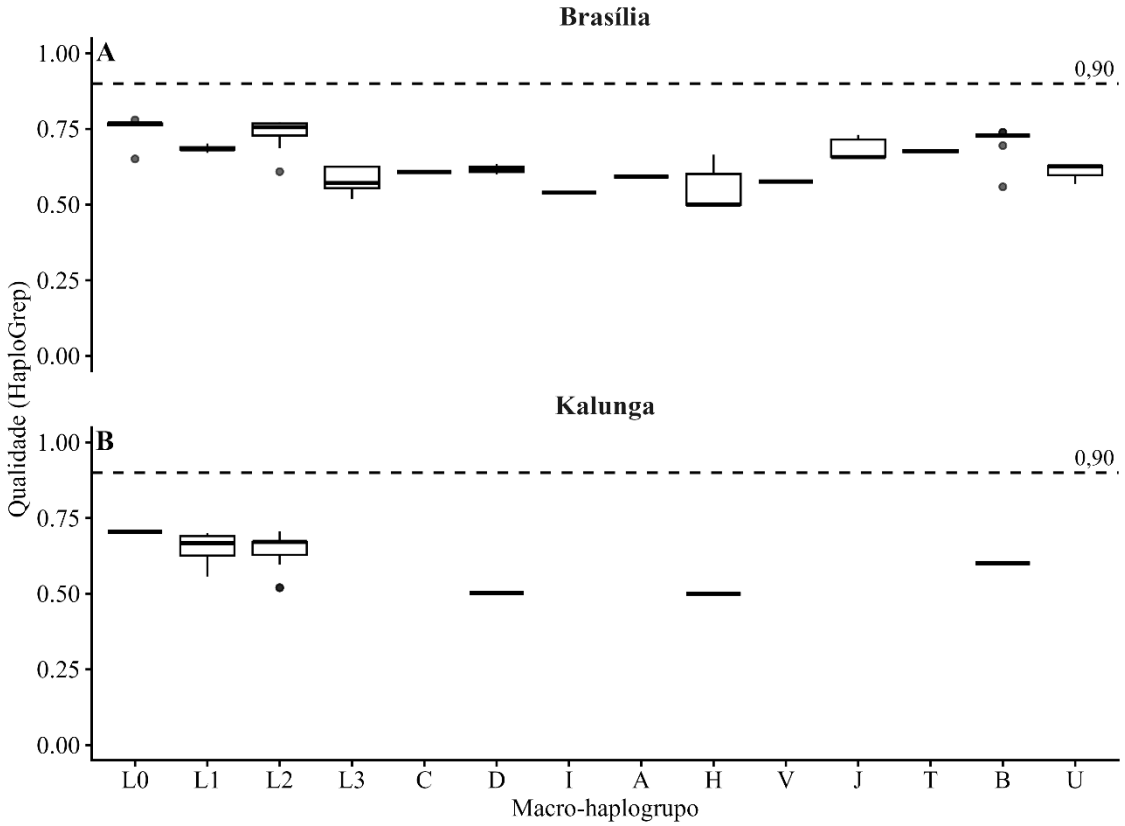


Fonte: O autor (2026), a partir dos resultados do HaploGrep.

Nota: (A) Brasília; (B) Kalunga. As barras representam a frequência das amostras por intervalo de qualidade e a linha contínua indica a densidade estimada. A linha tracejada marca o limiar 0,90.

A qualidade dos macro-haplogrupos inferidos mostrou variação discreta entre categorias. No entanto, não houve aproximação ao limiar de 0,90. Os macro-haplogrupos europeus e africanos apresentaram medianas superiores em relação a alguns macro-haplogrupos indígenas, em ambas as amostras (Figura 3). Como esperado, visto que Souza *et al*. (2025) mostrou em sua análise que a baixa representação de indígenas brasileiros em bases como *gnomAD* e *EMPOP* leva a possíveis erros de atribuição e à ausência desses macro-haplogrupos.

Figura 3 – Distribuição do escore de qualidade da atribuição de haplogrupos mitocondriais (HaploGrep) segundo os macro-haplogrupos identificados nas populações analisadas.



Fonte: O autor (2026), a partir dos resultados do HaploGrep.

Nota: (A) Brasília; (B) Kalunga. As caixas representam a mediana e o intervalo interquartil, enquanto os pontos indicam valores discrepantes. A linha tracejada horizontal marca o limiar de qualidade 0,90.

5.1.2 Inferência dos haplogrupos e macro-haplogrupos mitocondriais

De acordo com os haplogrupos inferidos para a população de Brasília, o haplogrupo H2a2a1 (n=21) foi o mais frequente, seguido por A16 (n=11) e B2 (n=10). Todavia, na população Kalunga, observou-se maior frequência do haplogrupo L2b2a (n=15) e, em seguida, H2a2a1(n=13) e L1b (n=7), conforme pode ser observado Figura 4.

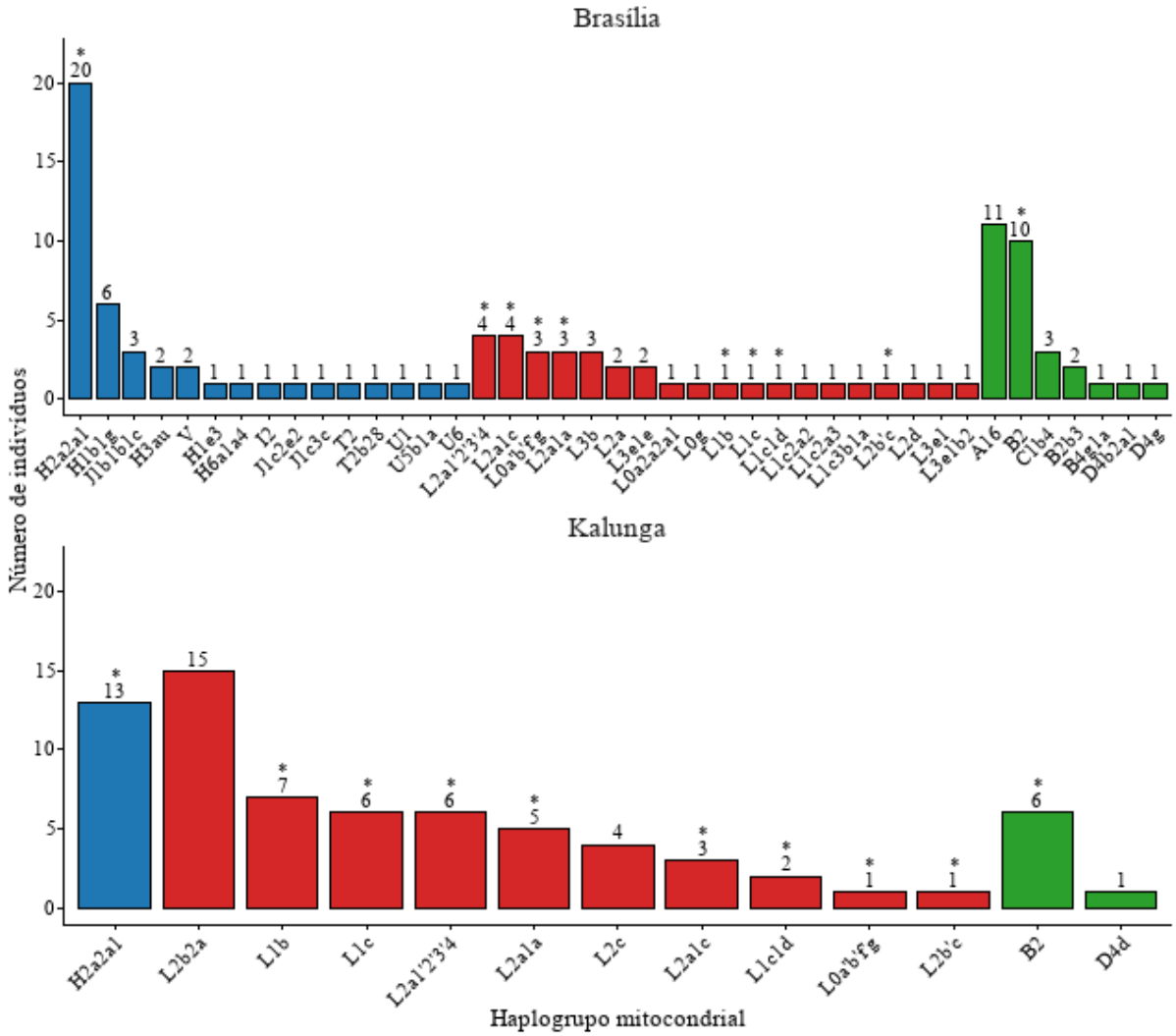
O H2a2a1 é classificado em grandes bancos de dados - como *MITOMAP* e *gnomAD* - como europeu ocidental (LARICCHIA *et al*., 2022). Em ambas as amostras o haplogrupo H2a2a1 mostrou elevada frequência (1° posição em Brasília e 2º posição em Kalunga). Algo que surpreende nesses resultados dos haplogrupos de Brasília é o número de indivíduos atribuídos ao haplogrupo A16, pertence ao macro-haplogrupo A, cuja origem está associada a populações da Ásia Oriental e do norte da Ásia, sendo observado em grupos da Sibéria e da Ásia Central (VAN OVEN; KAYSER, 2009). Esse haplogrupo não foi descrito nominalmente em amostras brasileiras publicadas. O haplogrupo B2 é comum em linhagens indígenas

brasileiras e possui elevada frequência em povos Tupis, Urubu-Kaapor e Jêan (GONÇALVES, 2020).

Os macro-haplogrupos A, B, C, D frequentemente são atribuídos a populações indígenas (AVILA *et al*., 2019). Evidências apontam que o macro-haplogrupo B tenha atingido o continente americano através de uma rota alternativa contornando a Sibéria. Esta hipótese para o padrão de migração do macro-haplogrupo B se deve ao fato de ser o único encontrado em nativos americanos que está ausente em populações do norte da Sibéria, sendo extremamente raro no Norte do continente americano e a diversidade de sequência deste haplogrupo ser mais baixa em relação aos haplogrupos A, C e D (MISHMAR *et al*., 2003; WALLACE, 2005).

O haplogrupo L2b2a é um ramo de L2b que tem origem na África Ocidental, e pode ter estado envolvido na expansão Bantu (SILVA *et al*., 2015). O haplogrupo L1b apresenta uma distribuição geográfica comum em afro-americanos, causada pelo tráfico transatlântico de escravos (SALAS *et al*., 2002).

Figura 4 - Distribuição dos haplogrupos mitocondriais nas populações de Brasília e Kalunga.



Fonte: O autor (2026), a partir dos resultados do HaploGrep.

Nota: As barras representam o número de indivíduos atribuídos a cada haplogrupo mitocondrial, organizados separadamente por população. As cores indicam a ancestralidade materna: azul (europeia), vermelho (africana) e verde (indígena). Os valores numéricos acima das barras correspondem ao número de indivíduos por haplogrupo, e o asterisco (*) indica haplogrupos compartilhados entre as duas populações.

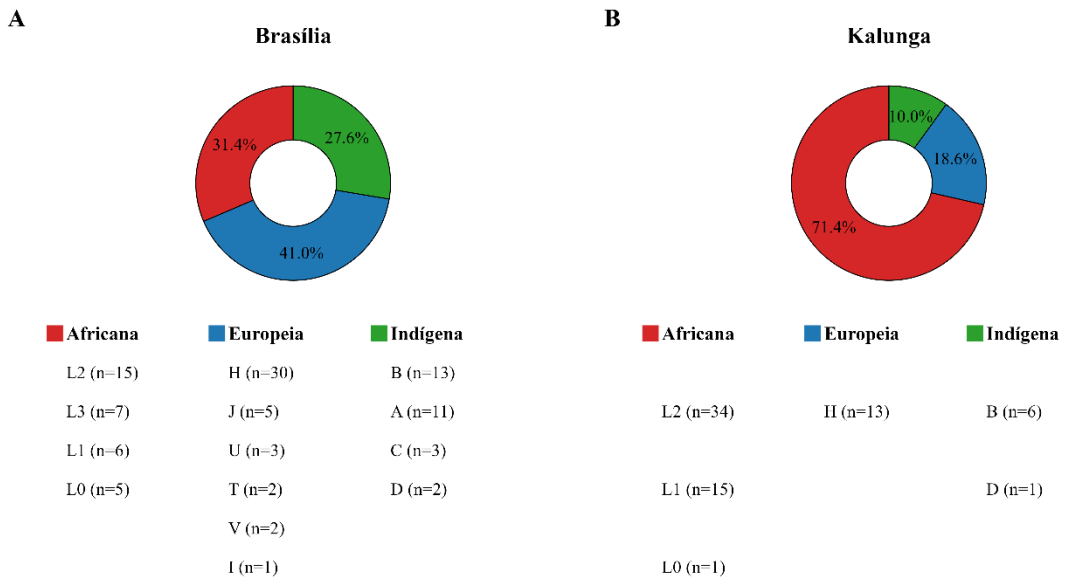
A Figura 5 mostra a distribuição dos macro-haplogrupos em Brasília (A) e Kalunga (B). Em Brasília o macro-haplogrupo H (n=30) está presente no maior número de indivíduos, seguido pelos macro-haplogrupos L2 (n=15) e B (n=13). A ancestralidade materna da população foi estimada em 41,0% europeia, 27,6% indígena e 31,4% africana. Esse resultado é contrastante com os de Freitas *et al*., (2019), que observou como principal macrohaplogrupos de origem africana L (49,02%), seguidos pelos haplogrupos nativos americanos típicos A, B, C e D (33,33%) e os 17,65% restantes distribuídos entre outros haplogrupos em uma amostra da população de Brasília obtida na Fundação Hemocentro da cidade. Outro estudo com a população de Brasília, Santos (2023), observou 37,5% dos haplogrupos de origem africana,

31,7% europeia, 28,8% nativo americana e 1,9% do leste asiático. Joerin-Luque *et al*., (2022) observou como resultados no Centro-Oeste: 24% de europeus, 31% de africanos e 44% de nativo americanos. Porém como não é retratado nesse estudo da região Centro-Oeste, qual população é analisada, ainda que na mesma região geográfica, a diferença de população da amostra pode ter ocasionado as divergências nos resultados.

Em Kalunga, os macro-haplogrupos foram concentrados em L (n=49), com L1 e L2, seguido pelo macro-haplogrupo H (n=13). Consoante com a classificação filogenética do *gnomAD*, a partir dos macro-haplogrupos, a ancestralidade materna da comunidade Kalunga é majoritariamente africana. De acordo com as frequências populacionais do *gnomAD*, os macro-haplogrupos L0, L1 e L2 constituem os macro-haplogrupos mais antigos e frequentes na população africana; além disso, representam aproximadamente 76% de todas as linhagens subsaarianas. O macro-haplogrupo L2 foi inferido para 34 indivíduos, sendo o mais frequente na amostra. O macro-haplogrupo H, foi observado em 13 indivíduos. Ambos os macro-haplogrupos B e D apresentam altas frequências populacionais em indígenas, porém B foi observado em seis indivíduos e o D foi observado em apenas um indivíduo.

A comparação entre os macro-haplogrupos das duas coortes analisadas evidenciou diferenças. Na população de Brasília observou-se predominância de linhagens europeias, em contraste com a população Kalunga, que apresenta maior proporção de linhagens africanas.

Figura 5 - Distribuição dos macro-haplogrupos e ancestralidades mitocondriais nas populações de Brasília e Kalunga.



Fonte: O autor (2026), a partir dos resultados do HaploGrep.

Nota: A (Brasília) e B (Kalunga) mostram as frequências relativas das principais origens (Africana, Europeia e Indígena) e seus respectivos macro-haplogrupos.

Em conjunto, os resultados obtidos refletem a história demográfica, formação e composição genética diferenciada entre as duas populações analisadas. Em Brasília, observa-se maior proporção de linhagens europeias, seguidas por componentes africanos e indígenas. Esse padrão é compatível com a estrutura genética marcada por intensa migração interna durante o processo de construção da capital federal, a partir da década de 1960. A migração para Brasília envolveu contingentes populacionais de diferentes regiões do país, especialmente do Sudeste e Nordeste (SANTOS, 2023).

Em contraste, a população Kalunga apresenta predominância marcante de linhagens africanas, padrão consistente com sua origem histórica como comunidade quilombola. Estudos genéticos em populações quilombolas brasileiras têm evidenciado elevada frequência de haplogrupos L (L0–L3), característicos do continente africano, confirmando a manutenção da herança materna africana nessas comunidades (RIBEIRO-DOS-SANTOS *et al*., 2002; GONÇALVES *et al*., 2008). Há predominância de linhagens L2 e L3, frequentemente associadas às regiões da África Ocidental (FERREIRA *et al*., 2006).

A presença de linhagens indígenas em ambas as populações também é historicamente coerente. Haplogrupos como A, B, C e D são amplamente reconhecidos como característicos das populações nativo-americanas e resultam dos eventos de povoamento inicial das Américas via Beríngia durante o Pleistoceno final (ACHILLI *et al*., 2008; TAMM *et al*., 2007).

5.2 **Imputação com Beagle**

5.2.1 Eficiência da imputação

A Tabela 3 apresenta os parâmetros utilizados para a imputação de dados com o software Beagle 5.4. O painel de referência empregado para imputação é do 1000 Genomes Project, que contém 2.534 amostras e 3.617 marcadores mitocondriais. As amostras do estudo antes da imputação após normalização continham 175 amostras (Brasília, 105; Kalunga, 70) e apenas 59 variantes. A ferramenta Beagle abrange praticamente todo o genoma mitocondrial, com tamanho efetivo populacional estimado igual 4.620.899 e taxa de erro assumida igual $1{,}3 \times 10^{-3}$, parâmetros que indicam boa adequação do modelo probabilístico ao conjunto de dados analisado.

Tabela 3 - Parâmetros utilizados para a imputação mitocondrial com o software Beagle 5.4.

Parâmetro	Valor
Versão do Beagle	Beagle 5.4 (19Apr22)
Amostras de referência - 1000G	2.534
Amostras do estudo	175
Marcadores de referência (mtDNA) -1000G	3.617
Marcadores do estudo (mtDNA)	59
Tamanho da janela	1.000 pb
Sobreposição	200 pb
Ne estimado	4.620.899
Erro estimado	$1{,}3 \times 10^{-3}$
Tempo total de execução	3 segundos

Fonte: O autor (2026), a partir da execução do Beagle 5.4.

5.2.2 Qualidade da inferência dos haplogrupos após imputação

Os dados imputados com o software Beagle apresentaram aumento na qualidade da inferência de haplogrupos, avaliada pelo HaploGrep3, em todas as amostras analisadas. A média do escore de qualidade aumentou de 0,63 para 0,79 na amostra de Brasília; além disso, houve redução do desvio padrão, indicando maior precisão na atribuição de haplogrupos. Da mesma forma, na amostra de Kalunga, a média aumentou de 0,62 para 0,80. A Tabela 4 apresenta esses resultados e demonstra a contribuição da imputação para os valores de qualidade.

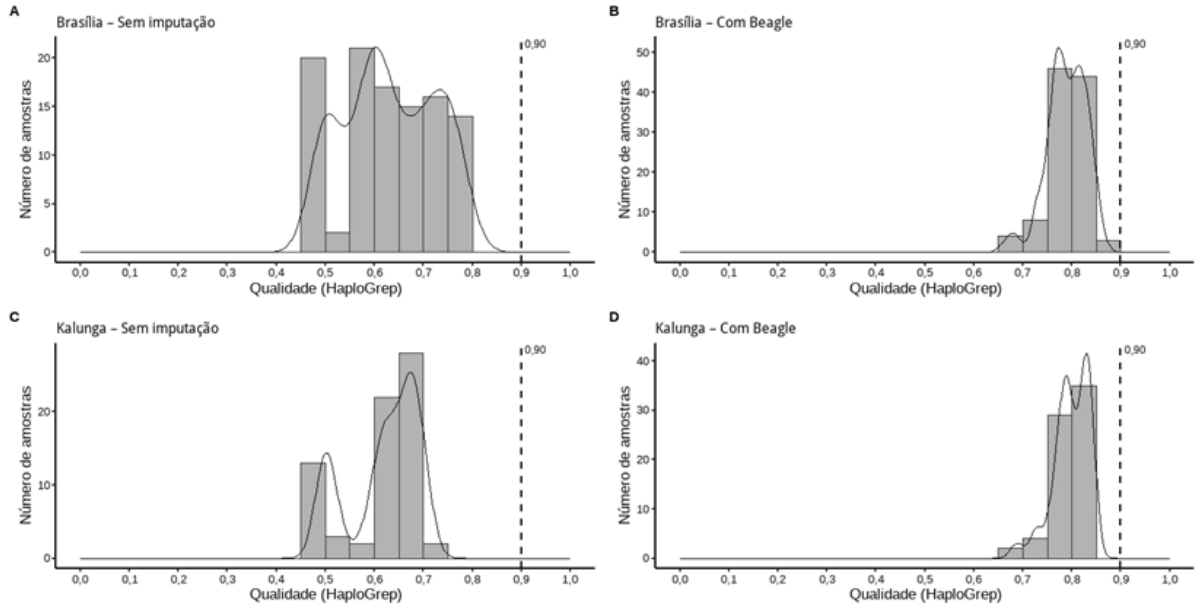
Tabela 4 - Comparação das estatísticas do escore de qualidade (HaploGrep) antes e após imputação com Beagle, nas amostras de Brasília e Kalunga.

População	Método	Número de amostras	Média	Mediana	Desvio-padrão
Brasília	Beagle	105	0,78	0,78	0,03
Brasília	Antes	105	0,63	0,61	0,09
Kalunga	Beagle	70	0,79	0,80	0,03
Kalunga	Antes	70	0,61	0,62	0,07

Fonte: O autor (2026).

Após a imputação, observou-se um deslocamento das concentrações de escores para faixas próximas ao referencial para atribuições com alta confiabilidade, especialmente em Brasília, ainda que sem atingir o limiar de 0,90, conforme a Figura 6. Comparando-se os dados sem imputação e com imputação pelo Beagle, nota-se aumento das medianas e estreitamento das distribuições após a imputação em ambas as populações, conforme o gráfico de violino (Figura 7).

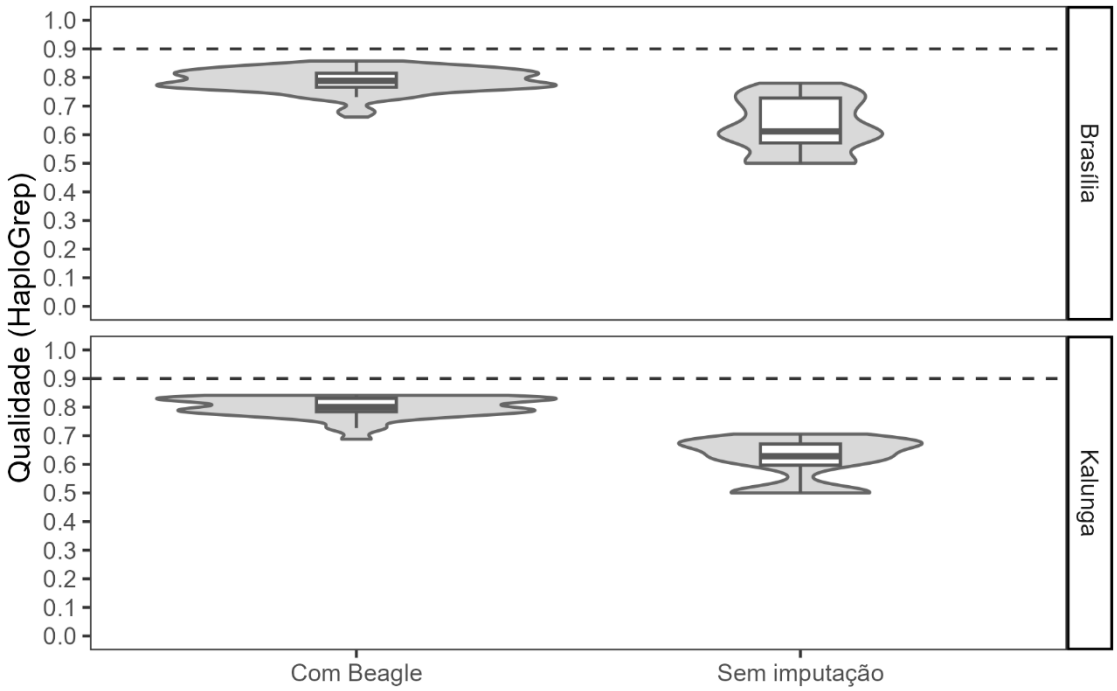
Figura 6 – Distribuição dos escores de qualidade (HaploGrep) nas amostras de Brasília e Kalunga, antes e após imputação com Beagle.



Fonte: O autor (2026), a partir dos resultados do HaploGrep.

Nota: Histogramas representam o número de amostras por faixa de qualidade, com curvas de densidade sobrepostas. A linha tracejada indica o limiar de qualidade 0,90.

Figura 7 – Comparação global do escore de qualidade (HaploGrep) antes e após imputação com Beagle em Brasília e Kalunga.



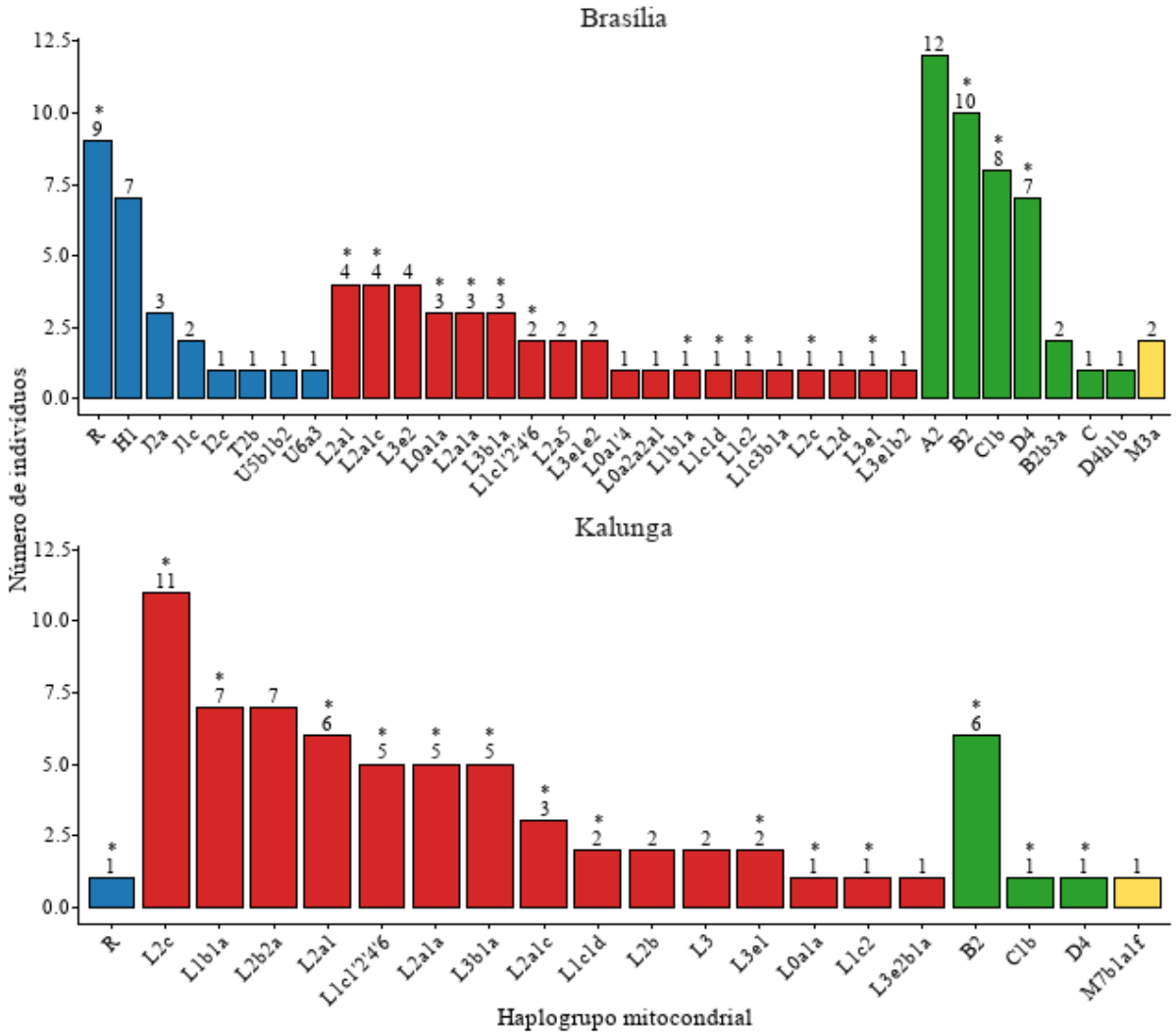
Fonte: O autor (2026), a partir dos resultados do HaploGrep.

Nota: Os violinos representam a distribuição; as caixas indicam mediana e intervalo interquartil; a linha tracejada marca o limiar de 0,90.

5.2.3 Inferência dos haplogrupos e macro-haplogrupos mitocondriais

A inferência dos haplogrupos e macro-haplogrupos também foi realizada pelo HaploGrep3 após a imputação. A Figura 8 dispõe todos os haplogrupos mitocondriais, em Brasília (n=105) e Kalunga (n=70), como observado houveram muitas mudanças. Em Brasília, o haplogrupo A2 (n=12) foi o mais frequente, seguido por B2 (n=10) e R (n=9). Todavia, em Kalunga, observou-se maior frequência do haplogrupo L2c (n=11) e, em seguida, L1b1a e L2b2a (n=7).

Figura 8 - Distribuição dos haplogrupos mitocondriais nas populações de Brasília (BSB) e Kalunga (KAL), após imputação com beagle.



Fonte: O autor (2026), a partir dos resultados do HaploGrep.

Nota: As barras representam o número de indivíduos atribuídos a cada haplogrupo mitocondrial, organizados separadamente por população. As cores indicam a ancestralidade materna: azul (europeia), vermelho (africana) e verde (indígena). Os valores numéricos acima das barras correspondem ao número de indivíduos por haplogrupo, e o asterisco (*) indica haplogrupos compartilhados entre as duas populações.

A partir da Tabela 5, é possível observar que em Brasília houve redução do número total de haplogrupos, de 41 para 35, assim como no número de macro-haplogrupos, de 20 para 18. Essa aparente redução não indica perda de diversidade, mas sim refinamento da classificação: haplogrupos amplos foram redistribuídos em clados filogeneticamente mais coerentes. A substituição de H2 (20 amostras antes da imputação) por haplogrupos dentro de R e H1, por exemplo, sugere correção de classificações supergeneralizadas decorrentes da baixa cobertura inicial.

No grupo amostral de Kalunga, o resultado foi o contrário, com aumento do número de haplogrupos de 13 para 21 e macro-haplogrupos de 6 para 9. Isso pode ser resultado da detecção

de macro-haplogrupos africanos previamente não inferidos, como L3, que passou de ausente para 14,3% após imputação.

Tabela 5 – Distribuição dos macro-haplogrupos em Brasília e Kalunga, antes (sem imputação) e após imputação (Beagle).

Brasília

Macro-haplogrupo	Número (Beagle)	Número (sem imputação)	Porcentagem (%) com Beagle	Porcentagem (%) sem imputação
L2	15	15	14,3	14,3
A2	12	0	11,4	0,0
B2	12	12	11,4	11,4
L3	11	7	10,5	6,7
R	9	0	8,6	0,0
C1	8	3	7,6	2,9
D4	8	2	7,6	1,9
H1	7	7	6,7	6,7
L1	6	6	5,7	5,7
L0	5	5	4,8	4,8
J2	3	0	2,9	0,0
J1	2	5	1,9	4,8
M3	2	0	1,9	0,0
C	1	0	1,0	0,0
I2	1	1	1,0	1,0
T2	1	2	1,0	1,9
U5	1	1	1,0	1,0
U6	1	1	1,0	1,0
A16	0	11	0,0	10,5
B4	0	1	0,0	1,0
H2	0	20	0,0	19,0
H3	0	2	0,0	1,9
H6	0	1	0,0	1,0
U1	0	1	0,0	1,0
V	0	2	0,0	1,9

Kalunga

Macro-haplogrupo	Número (Beagle)	Número (sem imputação)	Porcentagem (%) com Beagle	Porcentagem (%) sem imputação
L2	34	34	48,6	48,6
L1	15	15	21,4	21,4
L3	10	0	14,3	0,0
B2	6	6	8,6	8,6
C1	1	0	1,4	0,0
D4	1	1	1,4	1,4
L0	1	1	1,4	1,4
M7	1	0	1,4	0,0
R	1	0	1,4	0,0
H2	0	13	0,0	18,6

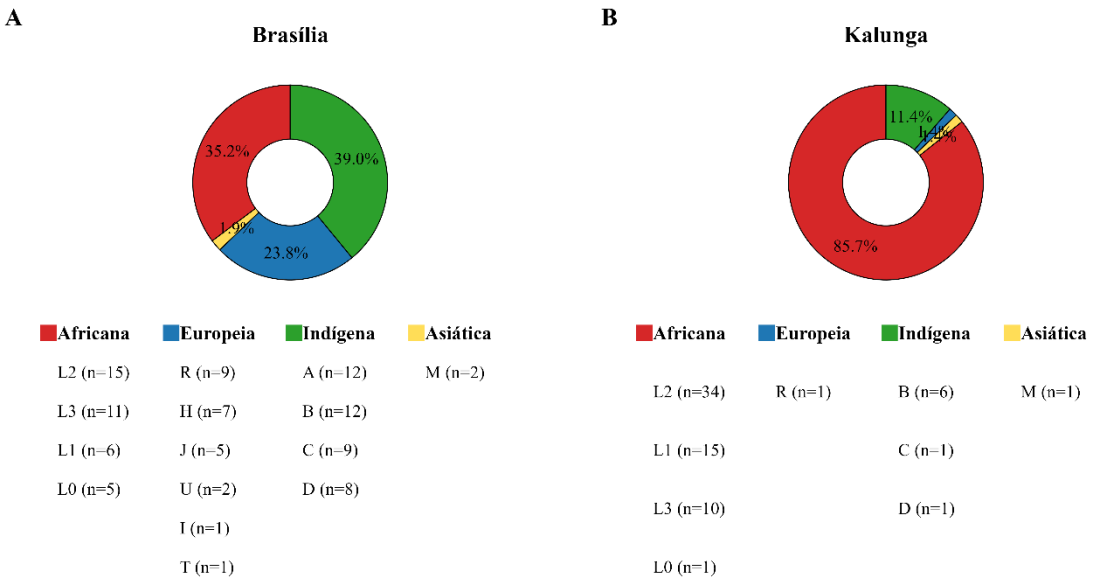
Fonte: O autor (2026).

Na população de Brasília, verifica-se grande mudança no percentual da ancestralidade materna, inferida pelos macro-haplogrupos mitocondriais, 39% indígena, 35,2% africana, 23,8% europeia e 1,9% asiática (Figura 9.A). Esse resultado está associado principalmente à reclassificação de A16 (antes 10,5%) para A2 (11,4%). A literatura reconhece A2, como um dos principais fundadores maternos das populações nativo-americanas (ACHILLI *et al*., 2008; TAMM *et al*., 2007). Outros macro-haplogrupos, como R, J2, M3 e C1, que estavam ausentes nos dados sem imputação, foram inferidos e reduziu a frequência em haplogrupos H.

Na população Kalunga, manteve-se predominância africana (85,7%) (Figura 9.B). A manutenção de L2 (48,6%) e L1 (21,4%) está em concordância com estudos de comunidades afrodescendentes brasileiras (RIBEIRO-DOS-SANTOS *et al*., 2002; GONÇALVES *et al*., 2008). Entretanto, a emergência de macro-haplogrupos M, C e R após imputação, em baixa frequência, sugere possível sobreimputação, isso ocorre quando o algoritmo de imputação infere variantes ou haplótipos que não estavam suficientemente suportados pelos dados originais das amostras, geralmente devido ao número reduzido de marcadores observados antes da imputação ou à forte dependência do painel de referência. As amostras de Brasília apresentaram redução da contribuição africana e aumento de linhagens europeias e indígenas,

além da inferência de linhagens asiáticas. Outrossim, nas amostras de Kalunga, houve redução da contribuição europeia observada anteriormente, sem imputação.

Figura 9 - Distribuição dos macro-haplogrupos e das ancestralidades mitocondriais nas populações de Brasília e Kalunga, após imputação com Beagle.



Fonte: O autor (2026), a partir dos resultados do HaploGrep.

Nota: A (Brasília) e B (Kalunga) mostram as frequências relativas das principais origens (Africana, Europeia e Indígena) e seus respectivos macro-haplogrupos após imputação com Beagle.

Os resultados obtidos após a imputação são consistentes com o princípio de que maior densidade de SNPs aumenta a acurácia da atribuição filogenética em *software* como o HaploGrep3, cuja classificação depende da correspondência com a árvore filogenética do PhyloTree (WEISSENSTEINER *et al.*, 2016; VAN OVEN, 2015). Porém, o Beagle é uma técnica que foi desenvolvida para genoma nuclear.

Diferentemente do genoma nuclear, o mtDNA não sofre recombinação, e sua estrutura filogenética é estritamente hierárquica.

Além disso, a inferência de macro-haplogrupos como M3 e M7 — associados à Ásia — deve ser interpretada com cautela nas amostras. O macro-haplogrupo M é amplamente distribuído na Ásia e raro nas Américas, exceto por subclados específicos (ACHILLI *et al.*, 2008). A presença de um indivíduo com macro-haplogrupo M na comunidade Kalunga resulta de algum ruído de imputação, ou efeito do painel de referência (THE 1000 GENOMES PROJECT CONSORTIUM, 2015) ou erro de classificação em regiões de baixa cobertura original.

5.3 Imputação com MitoImpute (IMPUTE2)

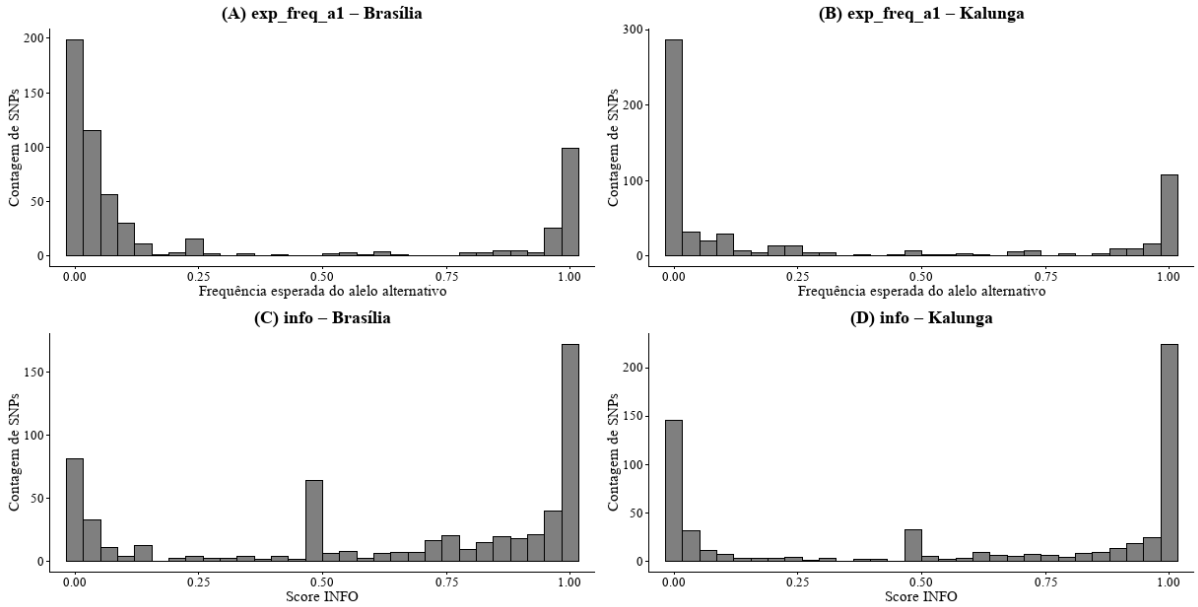
5.3.1 Eficiência da imputação

A imputação do mitogenoma utilizando o software IMPUTE2, a partir de dados de genotipagem, resultaram em um aumento na resolução. Inicialmente eram 234 SNPs genotipados nos dois grupos amostrais, que passou a incluir 591 posições com o painel imputado, um aumento de aproximadamente 2,5 vezes mais sítios informativos. Após imputação, foram observados 304 sítios segregantes em Brasília e 219 em Kalunga, ou seja, maior número de variantes polimórficas na amostra de Brasília.

A distribuição das frequências esperadas do alelo alternativo e dos escores de informação para os sítios imputados nas duas populações foi avaliada, apresentando um padrão bimodal característico, com acúmulo de SNPs com frequências próximas a zero e a um (Figura 10, painéis A e B). O padrão similar em ambos grupos amostrais indica consistência na imputação entre populações contrastantes.

Em relação a confiabilidade estatística dos genótipos imputados, os escores INFO apresentaram uma distribuição concentrada em valores elevados em ambas populações (Figura 10, painéis C e D). Observou-se um conjunto menor com valores intermediários ou baixos de INFO, compatível com regiões menos informativas ou com menor representação no painel de referência utilizado.

Figura 10 - Distribuição das métricas de qualidade da imputação mitogenômica nas populações de Brasília e Kalunga.



Fonte: O autor (2026).

Nota: (A) Distribuição das frequências esperadas do alelo alternativo (exp_freq_a1) na população de Brasília; (B) Distribuição das frequências esperadas do alelo alternativo (exp_freq_a1) na população Kalunga; (C) Distribuição do escore de informação (INFO) dos SNPs mitocondriais imputados na população de Brasília; (D) Distribuição do escore de informação (INFO) dos SNPs mitocondriais imputados na população Kalunga.

5.3.2 Qualidade da inferência dos haplogrupos após imputação

A avaliação da qualidade da inferência dos haplogrupos após imputação com MitoImpute, pelo escore de classificação do HaploGrep3, foi elevada. Em Brasília, o escore médio é igual a 0,83 e mediana de 0,85, enquanto em Kalunga o valor médio é igual a 0,89 e mediana de 0,91 (Tabela 6).

Tabela 6 - Estatísticas descritivas do escore de qualidade da atribuição de haplogrupos (HaploGrep) para as populações de Brasília e Kalunga.

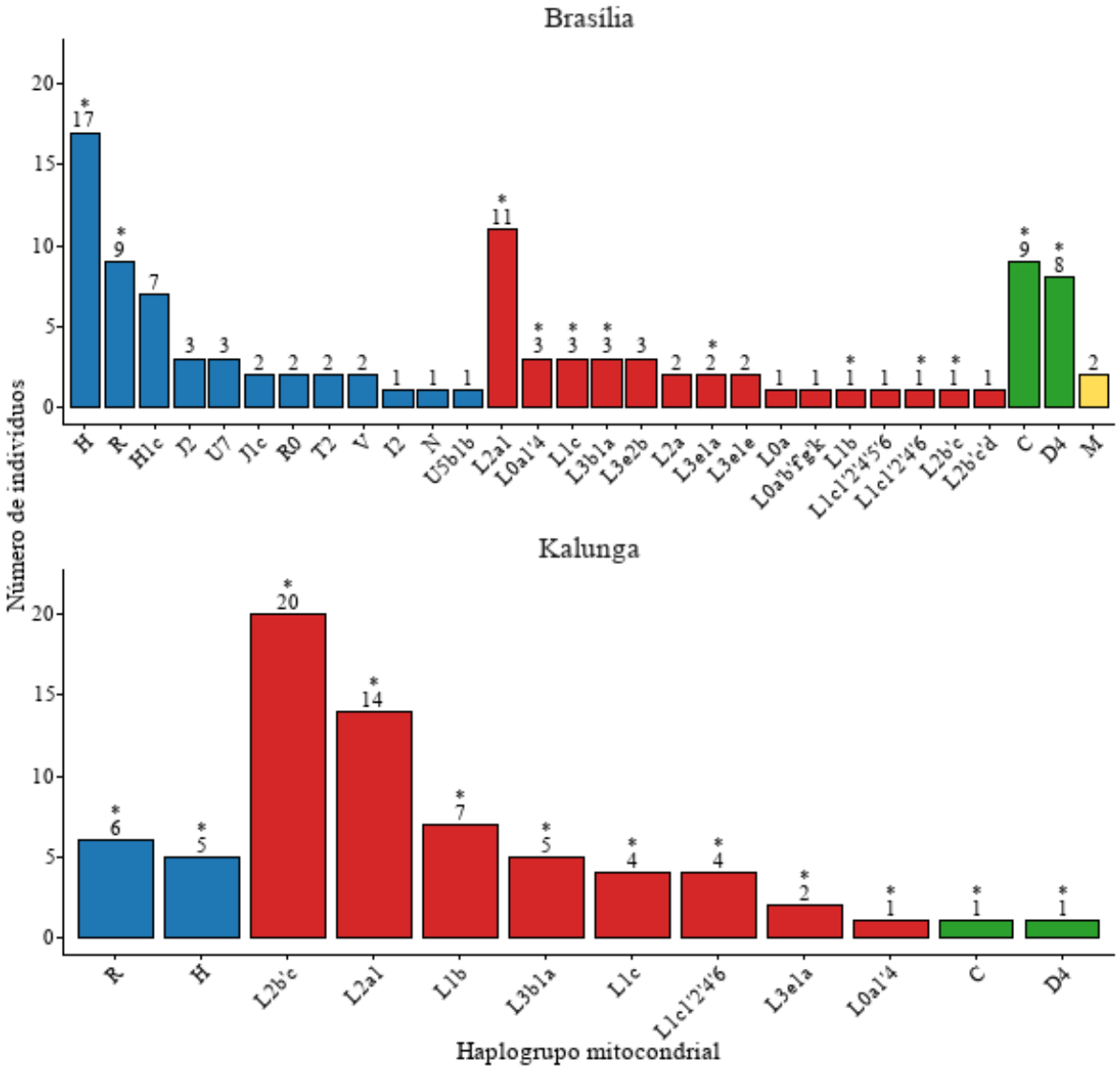
Escore de qualidade	Brasília	Kalunga
Número de amostras	105	70
Média	0,81	0,85
Mediana	0,82	0,87
Mínimo	0,69	0,74
Máximo	0,92	0,92
Proporção de amostras com qualidade ≥ 0,90	0,10	0,22

Fonte: O autor (2026), a partir dos resultados do HaploGrep 3.

5.3.3 Inferência dos haplogrupos e macro-haplogrupos mitocondriais

Os haplogrupos e macro-haplogrupos mitocondriais foram inferidos a partir dos mitogenomas imputados utilizando o software HaploGrep 3, executado em ambiente Linux (Ubuntu). A imputação aumentou a resolução filogenética das amostras, possibilitando a atribuição de alguns haplogrupos mais específicos, porém outros foram reclassificados apenas em macro-haplogrupos. Outra particularidade foi que todos os haplogrupos da amostra de Kalunga foram também observados em Brasília. Como se pode notar na Figura 11, a população de Brasília (n = 105), teve como haplogrupos mais frequentes após a imputação H, L2a, R e C. Ou seja, observa-se uma maior participação de linhagens europeias na população, seguida da africana. Por outro lado, a imputação revelou forte predominância de linhagens africanas na população Kalunga (n = 70), com destaque para L2b'c, L2a1 e L1b.

Figura 11 – Distribuição dos haplogrupos mitocondriais nas populações de Brasília e Kalunga, após imputação com MitoImpute.



Fonte: O autor (2026), a partir dos resultados do HaploGrep.

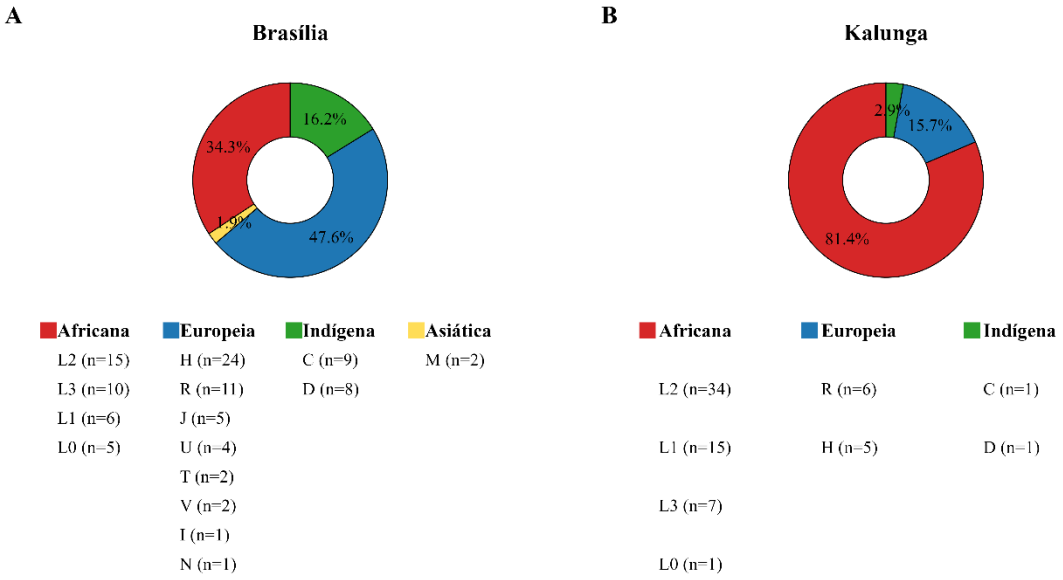
Nota: As cores indicam a ancestralidade materna: azul (europeia), vermelho (africana) e verde (indígena). O asterisco (*) indica haplogrupos compartilhados entre as duas populações.

A Figura 12.A, mostra que Brasília apresenta uma diversidade intra-populacional, com ancestralidade materna europeia (47,6%), africana (34,3%) e indígena (16,2%), além de contribuição minoritariamente asiática (1,9%). Esse conjunto confirma o perfil tetra-híbrido, como também apontado por Escher *et al.* (2022), característico da população brasileira, nas quais linhagens maternas europeias e africanas persistem em proporções relevantes, acompanhados por linhagens indígenas e asiáticas. Em contraste, temos a diversidade observada na comunidade Kalunga, que permanece com elevada porcentagem de ancestralidade materna africana (81,4%), sendo compatível com a história demográfica de comunidades

quilombolas, formadas majoritariamente por descendentes de africanos escravizados (FERREIRA *et al*., 2006). Porém em 2002, Abê-Sandes (2002) observou em comunidades remanescentes de quilombos no estado da Bahia, Barra e São Gonçalo ausência da contribuição europeia e indígena (Figura 12.B).

A comparação entre os resultados evidenciam que o MitoImpute proporcionou maior resolução filogenética sem distorcer o perfil demográfico das populações analisadas. Em Brasília, a atribuição de ancestralidade materna europeia aumentou e houve reclassificação de dois indivíduos para asiáticos. Em Kalunga, a predominância africana permaneceu evidente, porém com uma notória participação de indivíduos com ancestralidade europeia.

Figura 12. Distribuição dos macro-haplogrupos e ancestralidades mitocondriais nas populações de Brasília e Kalunga, após imputação com Mitoimpute.



Fonte: O autor (2026), a partir dos resultados do HaploGrep.

Nota: A (Brasília) e B (Kalunga) mostram as frequências relativas das principais origens e seus respectivos macro-haplogrupos, após imputação com IMPUTE2.

A pipeline MitoImpute demonstrou ser metodologicamente adequada para dados mitocondriais, permitindo refinamento na atribuição de haplogrupos e melhor caracterização da ancestralidade materna. Entretanto, determinadas classificações de haplogrupos foram generalizadas para o macro-haplogrupo com o objetivo de melhorar a qualidade de inferência.

Do ponto de vista histórico-demográfico, os resultados reforçam dois padrões distintos. Brasília é caracterizada por intensa miscigenação e fluxo migratório recente, refletindo a formação urbana e a integração de diferentes matrizes populacionais brasileiras. Já a

comunidade Kalunga preserva uma estrutura genética mais homogênea e predominantemente africana, coerente com seu contexto histórico de formação e relativo isolamento.

5.4 Comparação entre metodologias: Sem imputação × Beagle × MitoImpute

5.4.1 Qualidade da inferência dos haplogrupos

A qualidade da inferência de haplogrupos mitocondriais teve valores fornecidos pelo HaploGrep, comparando cada metodologia: sem imputação e as técnicas de imputação Beagle e MitoImpute nas populações de Brasília e Kalunga. A Tabela 7 e Figura 13 apresenta esses dados.

Tabela 7- Estatísticas descritivas do índice de qualidade da inferência de haplogrupos mitocondriais nas populações de Brasília e Kalunga, segundo método.

População	Método	Número de indivíduos	Média	Mediana	Desvio-padrão	Mínimo	Máximo
Brasília	Beagle	105	0,78	0,78	0,03	0,66	0,85
Brasília	MitoImpute	105	0,81	0,82	0,06	0,69	0,92
Brasília	Sem imputação	105	0,63	0,61	0,09	0,50	0,78
Kalunga	Beagle	70	0,79	0,80	0,03	0,68	0,84
Kalunga	MitoImpute	70	0,85	0,87	0,05	0,74	0,92
Kalunga	Sem imputação	70	0,61	0,62	0,07	0,50	0,70

Fonte: O autor (2026), a partir dos resultados do HaploGrep.

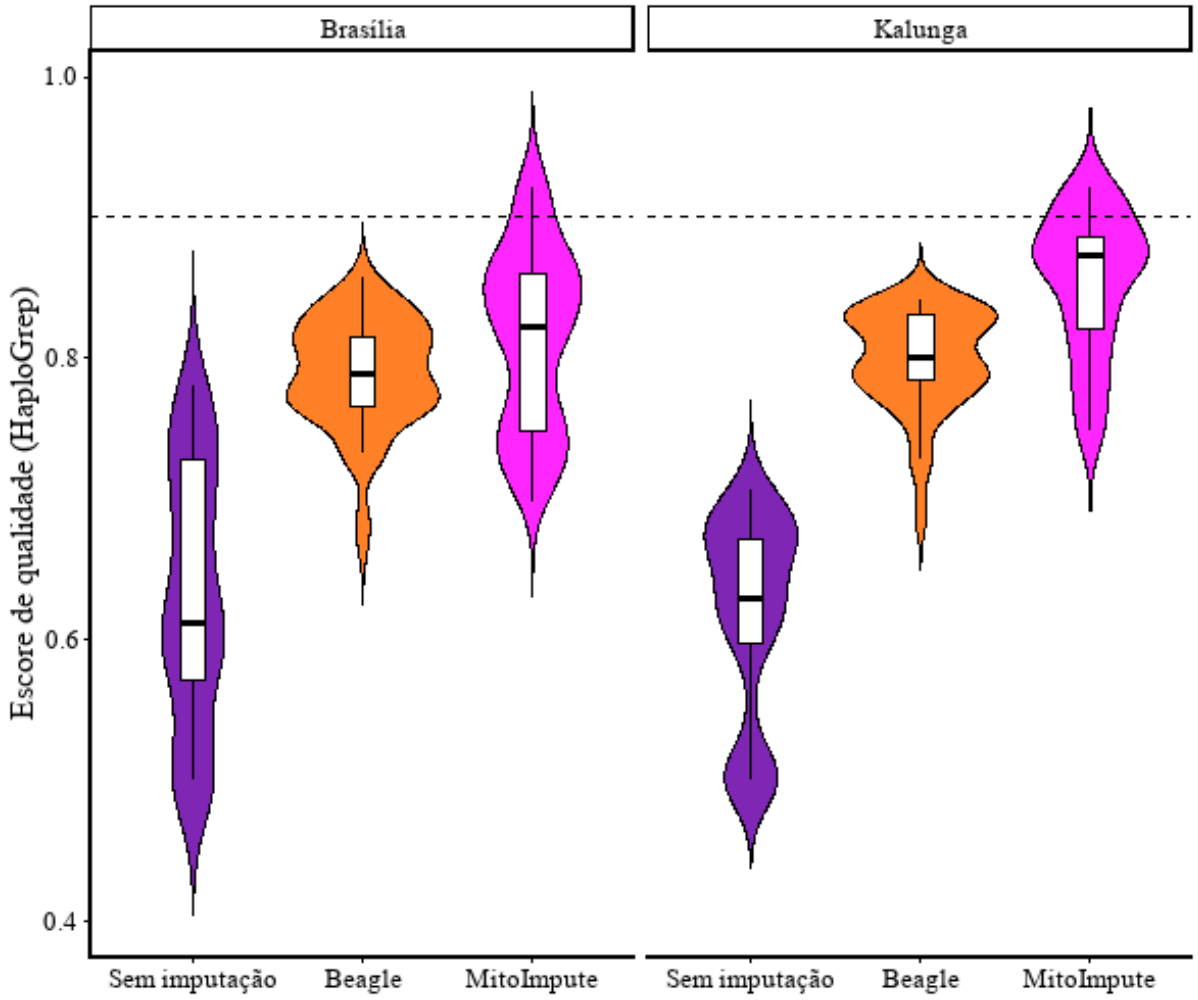
Nota: A tabela apresenta medidas em cada método analítico (Beagle, MitoImpute e sem imputação).

Houve um aumento nos escores médios de qualidade após imputação em ambas populações. Em Brasília, a média foi de 0,631 sem imputação, aumentando para 0,788 com Beagle e 0,814 com MitoImpute. Em Kalunga, os valores foram 0,618 (sem imputação), 0,798 (Beagle) e 0,853 (MitoImpute). O mesmo foi observado nas medianas, indicando melhoria na confiabilidade da classificação em haplogrupos. Além disso, apenas o MitoImpute apresentou

amostras com escore superior a 0,90 (> 90), tanto em Brasília (10,47%) quanto em Kalunga (22,85%), enquanto os demais métodos não atingiram esse limiar.

A distribuição dos escores (Figura 13) expõe que há maior dispersão e presença de valores baixos nos dados sem imputação. Uma distribuição e elevação dos escores nos dados com imputação Beagle, enquanto o MitoImpute apresentou deslocamento das distribuições para valores elevados, aproximando-se do limiar de alta confiabilidade (linha tracejada em 0,90). Em Kalunga, a diferença entre métodos foi ainda mais evidente, com o MitoImpute apresentando distribuição mais concentrada em valores superiores a 0,85.

Figura 13 – Distribuição dos escores de qualidade da inferência de haplogrupos mitocondriais para as populações de Brasília e Kalunga.



Fonte: O autor (2026), a partir dos resultados do HaploGrep.

Nota: Considerando três condições analíticas: sem imputação, imputação pelo Beagle e imputação pelo MitoImpute. Os violinos representam a densidade dos escores por método, enquanto as caixas internas indicam a mediana e o intervalo interquartil. A linha tracejada horizontal marca o limiar de qualidade ≥ 0,90, adotado como critério para classificações de alta confiança.

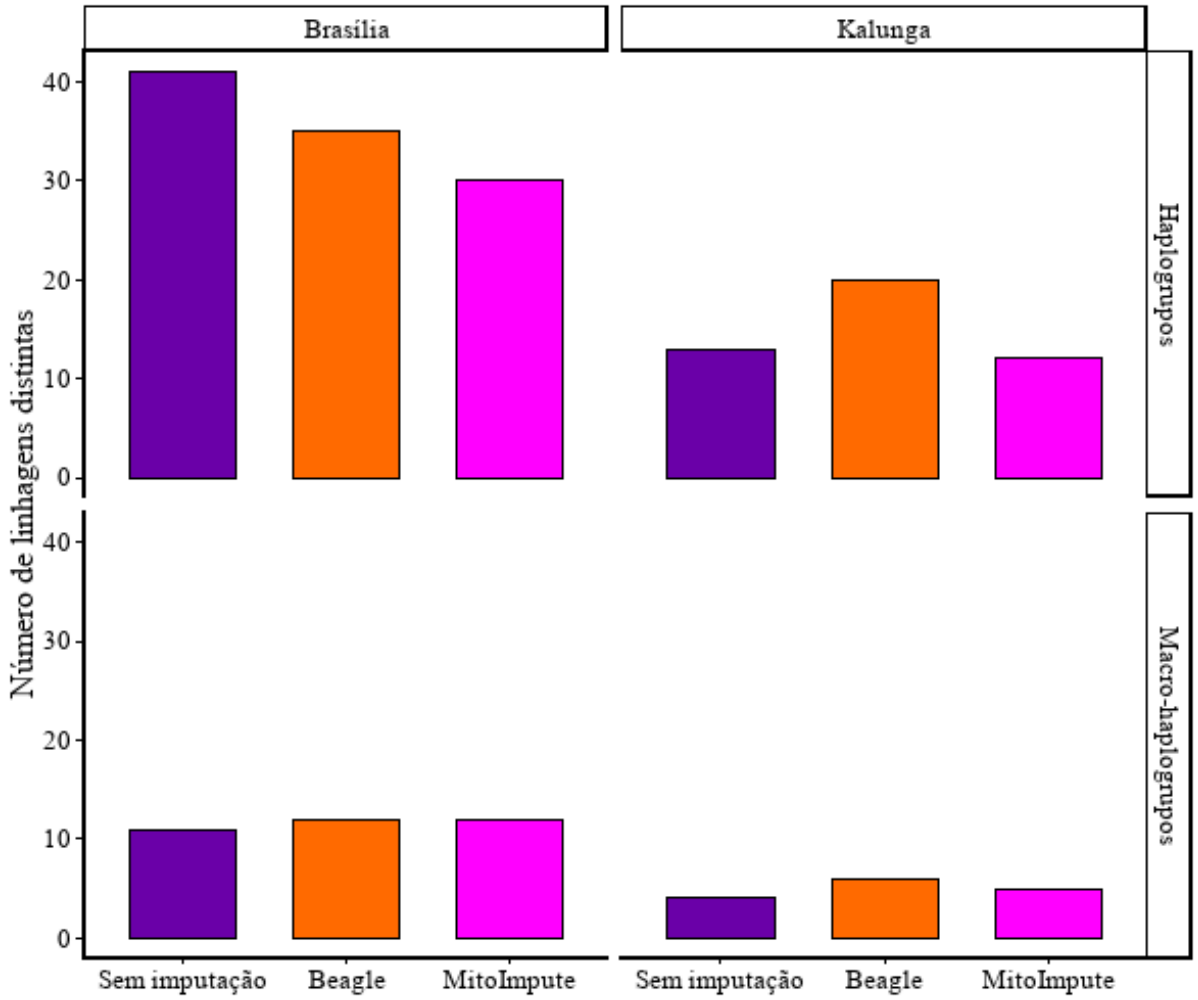
Esses resultados correspondem ao impacto direto da densidade de marcadores sobre a acurácia da classificação filogenética mitocondrial. A inferência de haplogrupos depende da presença de variantes diagnósticas específicas ao longo do genoma mitocondrial, conforme estabelecido na filogenia global descrita no PhyloTree (VAN OVEN; KAYSER, 2009). Quando o conjunto de SNPs é limitado, como no caso dos dados sem imputação, a ausência de marcadores específicos pode resultar em classificações imprecisas, nas quais uma sequência é atribuída a um haplogrupo incorreto; subatribuição, quando a linhagem é identificada apenas em níveis mais amplos da árvore filogenética por falta de mutações diagnósticas suficientes; ou ainda superfragmentação artificial de linhagens, quando pequenas diferenças entre sequências são interpretadas como subclados distintos devido à ausência de marcadores que permitiriam agrupá-las corretamente (WEISSENSTEINER *et al*., 2016). A imputação aumenta o número de sítios informativos, permitindo que a classificação se aproxime da estrutura filogenética real.

Em síntese, a comparação entre metodologias demonstra que ambos os métodos de imputação melhoram a qualidade da inferência em relação ao conjunto sem imputação, sendo o MitoImpute o que apresentou desempenho superior em termos de escore médio, mediana e proporção de classificações confiáveis. Esses resultados reforçam a importância do aumento da densidade de marcadores e da coerência filogenética na análise de dados mitocondriais em estudos populacionais.

5.4.2 Diversidade de haplogrupos e macro-haplogrupos mitocondriais

A diversidade foi avaliada por meio do número de haplogrupos mitocondriais distintos inferidos em ambas populações, comparando os diferentes métodos: sem imputação, e os imputados com Beagle e MitoImpute (Figura 14). Observou-se diminuição no número de haplogrupos na população de Brasília, após imputação com ambos os métodos. Porém os valores aumentaram em Kalunga com o Beagle.

Figura 14 – Número de haplogrupos e macro-haplogrupos mitocondriais distintos identificados nas populações de Brasília e Kalunga.



Fonte: O autor (2026).

Nota: Considerando três condições analíticas: sem imputação, imputação pelo Beagle e imputação pelo MitoImpute.

O maior número de haplogrupos observado no conjunto sem imputação em Brasília reflete provável superfragmentação decorrente da baixa densidade de marcadores diagnósticos, levando à atribuição de sub-haplogrupos distintos a indivíduos pertencentes a uma mesma linhagem. A imputação, ao recuperar variantes filogeneticamente informativas, promoveu refinamento da classificação, reduzindo o número total de haplogrupos, mas aumentando a consistência evolutiva da inferência, aumentando também o número de macro-haplogrupos. Em Kalunga, o aumento observado com o Beagle sugere maior fragmentação haplotípica, enquanto o MitoImpute apresentou classificação mais conservadora, compatível com a estrutura filogenética mitocondrial conhecida (Tabela 8).

Tabela 8 - Número de haplogrupos e macro-haplogrupos, nas amostras de Brasília e Kalunga, nas metodologias distintas.

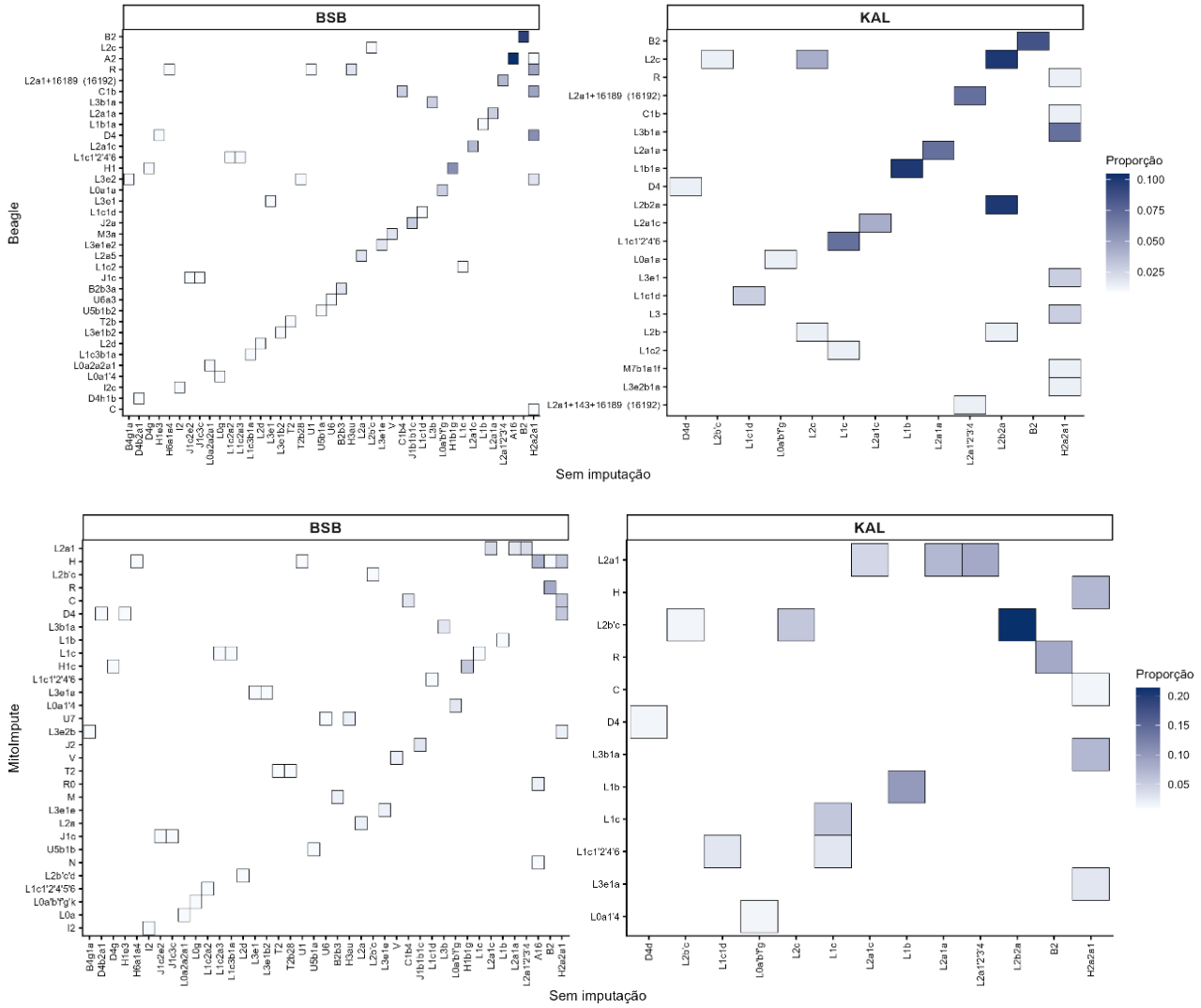
População	Método	Número de indivíduos	Número de haplogrupos	Número de macro-haplogrupos
Brasília	Beagle	105	35	12
Brasília	MitoImpute	105	30	12
Brasília	Sem imputação	105	41	11
Kalunga	Beagle	70	21	6
Kalunga	MitoImpute	70	12	5
Kalunga	Sem imputação	70	13	4

Fonte: O autor (2026). 5.4.3 Mudança de classificação por indivíduo e consistência entre métodos

A mudança de classificação de haplogrupos por indivíduo foi avaliada comparando as atribuições obtidas sem imputação com aquelas obtidas após imputação pelos dois métodos. Para cada indivíduo, registrou-se o haplogrupo inferido em cada condição e classificou como ausência de mudança, mudança de haplogrupo dentro do mesmo macro-haplogrupo ou mudança de macro-haplogrupo (Figura 15).

Observou-se que a imputação aumentou a resolução filogenética, promovendo transições principalmente para subclados mais específicos dentro dos macro-haplogrupos predominantes. As alterações de macro-haplogrupo foram menos frequentes, indicando que a imputação atua majoritariamente refinando, sem distorcer a estrutura ancestral ampla.

Figura 15 – Heatmaps de transição de haplogrupos mitocondriais por indivíduo, comparando as classificações obtidas sem imputação com aquelas obtidas após imputação pelos métodos Beagle (painéis superiores) e MitoImpute (painéis inferiores), nas populações de Brasília e Kalunga.



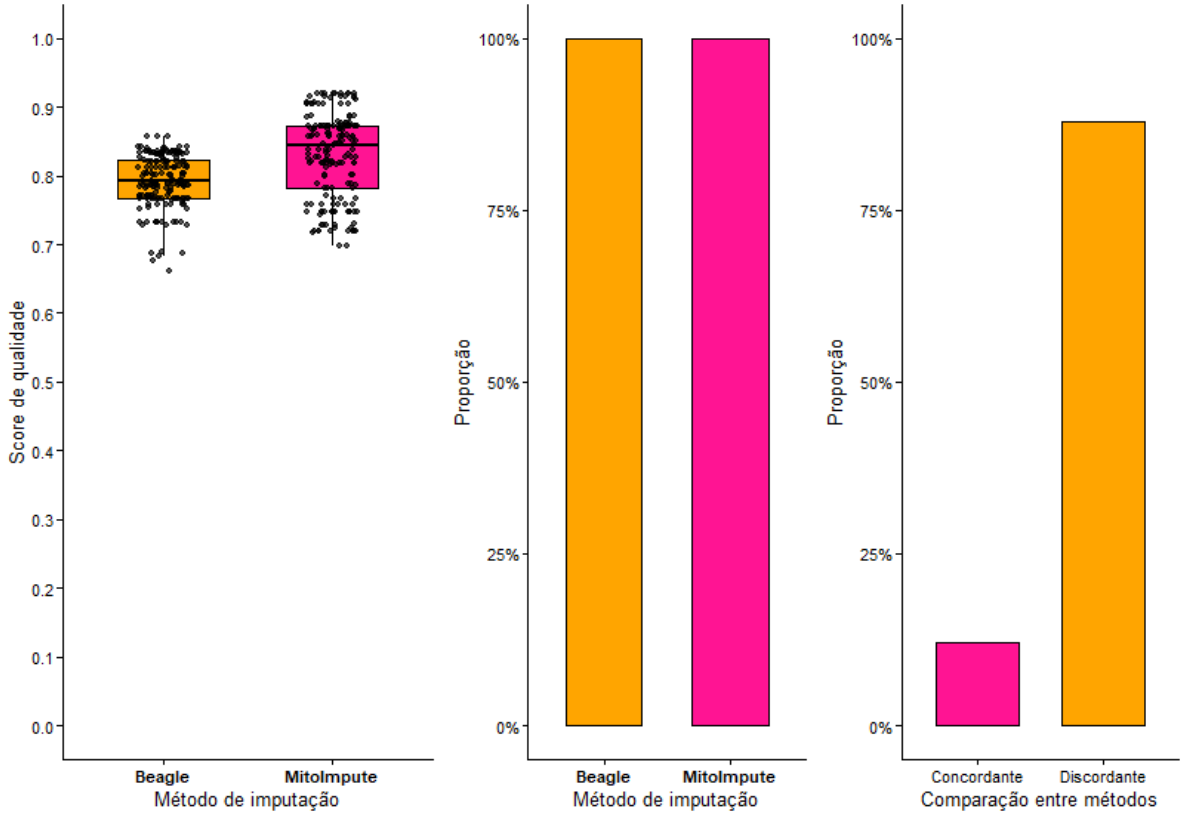
Fonte: O autor (2026).

Nota: As células representam a proporção de indivíduos que transitaram entre haplogrupos específicos, evidenciando principalmente mudanças intra-cládicas após imputação.

5.4.4 Comparação entre as duas ferramentas de imputação

A comparação entre os dois métodos de imputação mitocondrial evidenciou divergências, principalmente associadas a qualidade da imputação. Entretanto a taxa de atribuição de haplogrupos foi equivalente entre os dois métodos (Figura 16.B). A ferramenta IMPUTE2, usado pela pipeline MitoImpute, apresentou escores de qualidade mais elevados comparado ao Beagle (Figura 16.A). Além disso, conforme mostrado na Figura 16.C, a análise de concordância entre os haplogrupos inferidos revelou uma maior fração de discordâncias.

Figura 16 - Comparação entre os métodos de imputação mitocondrial Beagle e MitoImpute.



Fonte: O autor (2026), a partir dos resultados do HaploGrep.

Nota: (A) Distribuição dos scores de qualidade da imputação; (B) Proporção de amostras com haplogrupos mitocondriais atribuídos e não atribuídos; (C) Concordância direta dos haplogrupos mitocondriais inferidos por ambos os métodos.

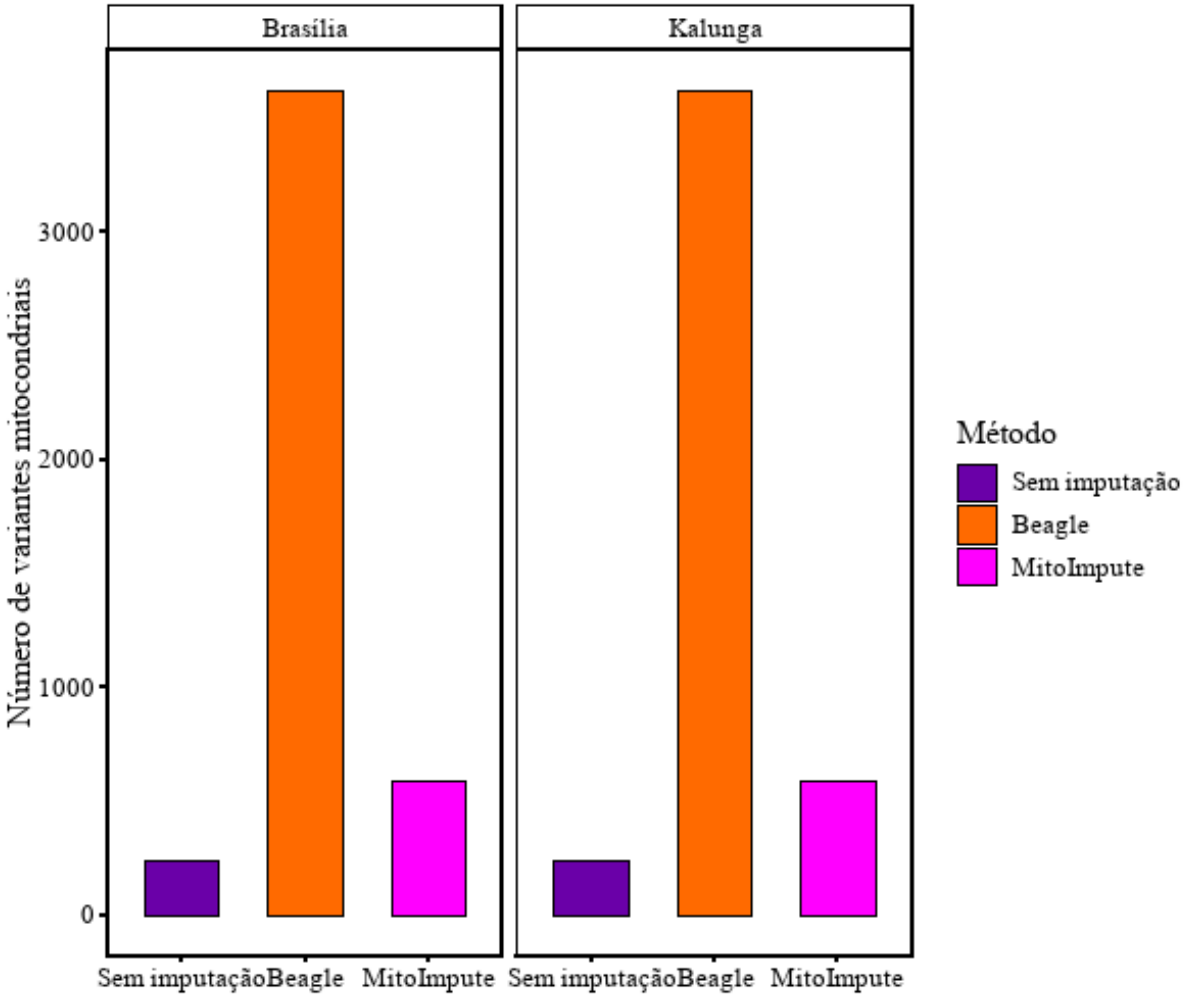
5.4.5 Caracterização das variantes mitocondriais

A caracterização das variantes mitocondriais demonstra diferenças palpáveis entre as metodologias: sem imputação x imputação com Beagle x imputação com MitoImpute. Inicialmente, constatou-se que os dados sem imputação continham um número reduzido de sítios polimórficos em ambas populações. A seguir, constatou-se que os dados imputados resultaram num aumento do número de variantes: o método Beagle apresentou o maior número de variantes recuperadas, enquanto o MitoImpute produziu número intermediário, porém maior que o conjunto sem imputação (Figura 17).

O Beagle é uma ferramenta de imputação de genótipos, tanto mitocondrial quanto nuclear, que amplia o número de posições genotípicas mesmo que nem todas essas posições

sejam informativas ou polimórficas na amostra. Ao contrário, a pipeline MitoImpute reconstrói haplótipos coerentes com a filogenia para mitogenomas.

Figura 17 – Número total de variantes mitocondriais detectadas por método de imputação nas populações de Brasília e Kalunga.

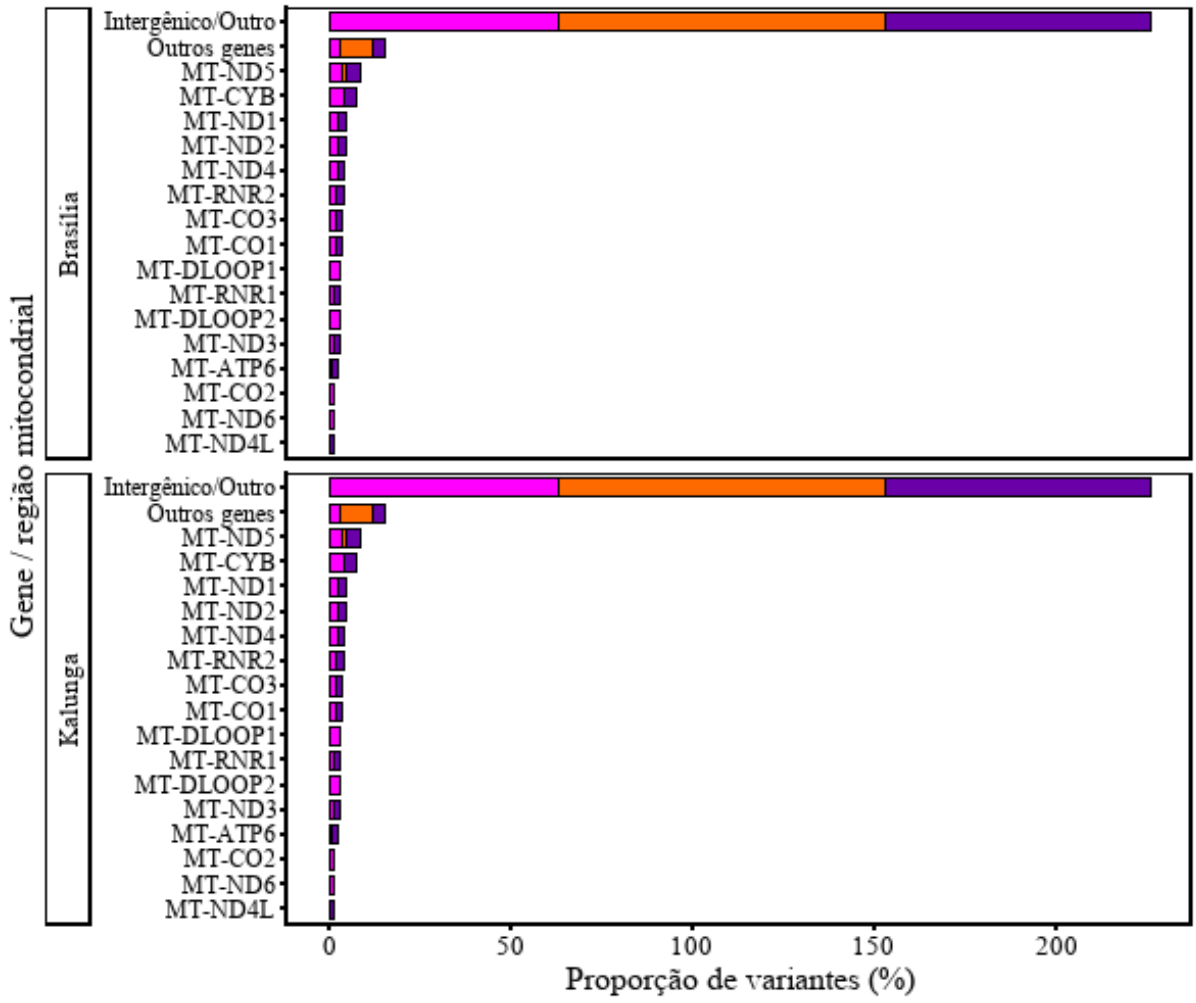


Fonte: O autor (2026).

Nota: O gráfico apresenta o número absoluto de sítios mitocondriais variantes identificados em cada população, comparando os conjuntos sem imputação e aqueles imputados pelos métodos Beagle e MitoImpute.

A distribuição das variantes ao longo do mtDNA aumentou principalmente com a imputação com o software Beagle (Figura 18). Houve uma maior concentração nos genes do complexo NADH desidrogenase (MT-ND), no gene MT-CYB e na região controle (D-loop), regiões conhecidas por apresentarem elevada variabilidade e maior densidade de sítios polimórficos em estudos mitocondriais globais (ANDERSON et al., 1981). Como mostra a Figura 18, há uma maior concentração de variantes em categorias intergênicas ou agregadas (PEREIRA *et al.*, 2009).

Figura 18 - Distribuição proporcional das variantes mitocondriais por gene/região nas populações de Brasília e Kalunga.



Fonte: O autor (2026).

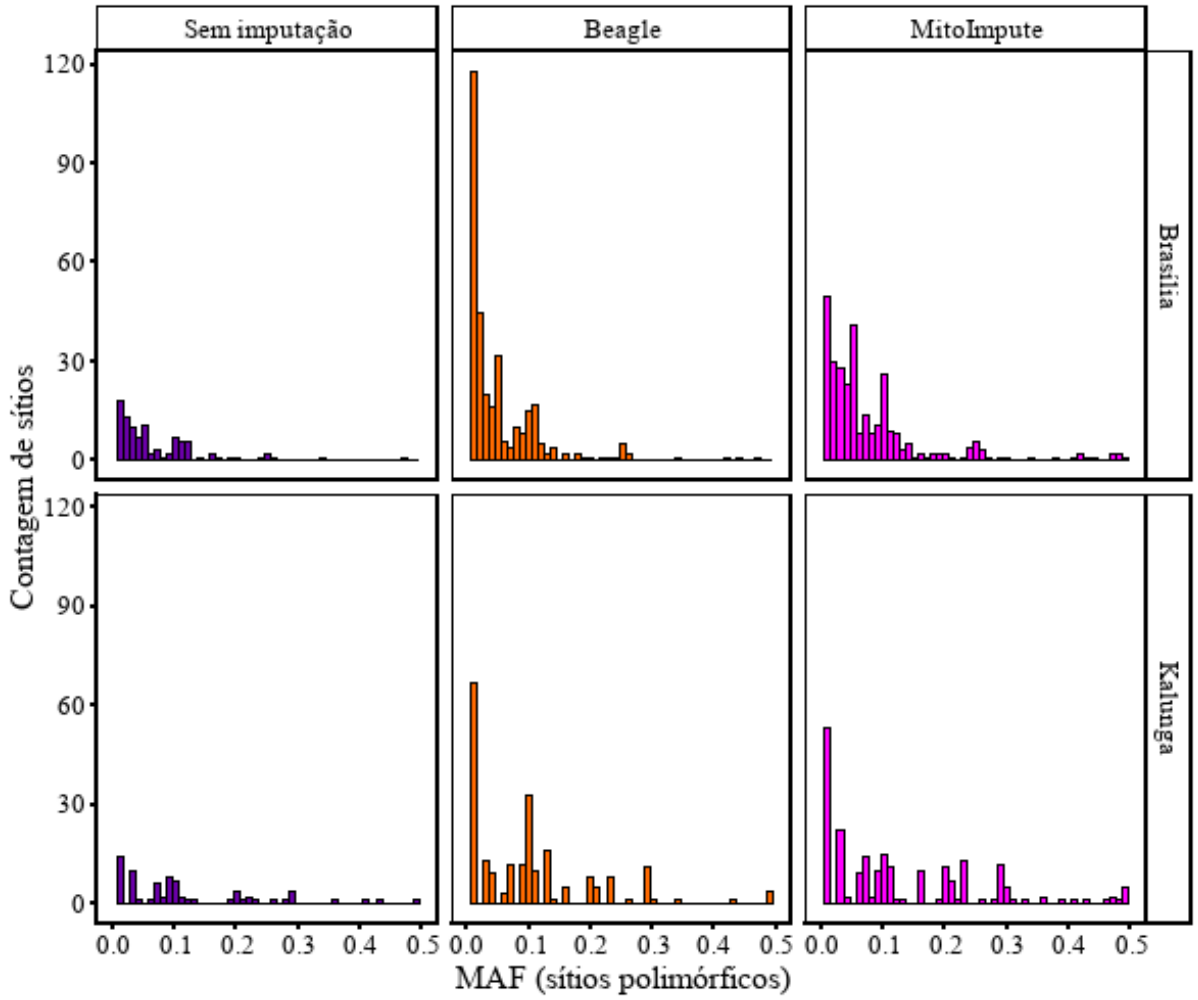
Nota: As cores representam os métodos de obtenção das variantes: laranja (Beagle), magenta (MitoImpute) e roxo (Sem imputação). As proporções são expressas em relação ao total de variantes observadas em cada população.

De acordo com a análise da distribuição da frequência alélica menor (MAF) nos dados imputados, observou-se predominância de variantes raras (MAF < 5%), com maior densidade de sítios de baixa frequência no MitoImpute, o que corresponde ao fato de que a maior parte da variação mitocondrial humana é composta por variantes raras, muitas delas específicas de linhagens recentes ou subclados filogenéticos (DE MARINO *et al*., 2022). A maior recuperação de variantes raras pelo MitoImpute pode ser reflexo maior da inferência de haplótipos específicos de referência, a identificação de variantes raras pode resultar da imputação de conjuntos de mutações que caracterizam determinados haplogrupos ou subclados

mitocondriais, ampliando a detecção de polimorfismos de baixa frequência nas amostras analisadas. Por outro lado, a imputação realizada com o software Beagle apresentou maior número absoluto de *singletons* (MAC = 1), isto é, variantes observadas em apenas um indivíduo da amostra. Esse padrão foi particularmente evidente na população Kalunga, o que pode indicar maior diversidade de variantes individuais ou, alternativamente, refletir diferenças na forma como cada método modela os haplótipos e infere variantes raras. Assim, enquanto o MitoImpute tende a recuperar variantes raras associadas a haplótipos presentes no painel de referência, o Beagle pode gerar maior número de variantes únicas, potencialmente relacionadas tanto à variabilidade real da população quanto às características do algoritmo de imputação utilizado. (Figura 19).

Em populações de menor tamanho amostral ou maior homogeneidade, como comunidades quilombolas, a imputação pode aumentar a detecção de variantes de frequência extremamente baixa devido à combinação de deriva genética e estrutura interna. Esse fenômeno é particularmente relevante para marcadores haploides como o mtDNA, cujo tamanho efetivo populacional reduzido intensifica flutuações alélicas (GAO *et al*., 1989).

Figura 19 - Distribuição da frequência do alelo minoritário (MAF) dos sítios mitocondriais polimórficos (MAF > 0) nas populações de Brasília e Kalunga.



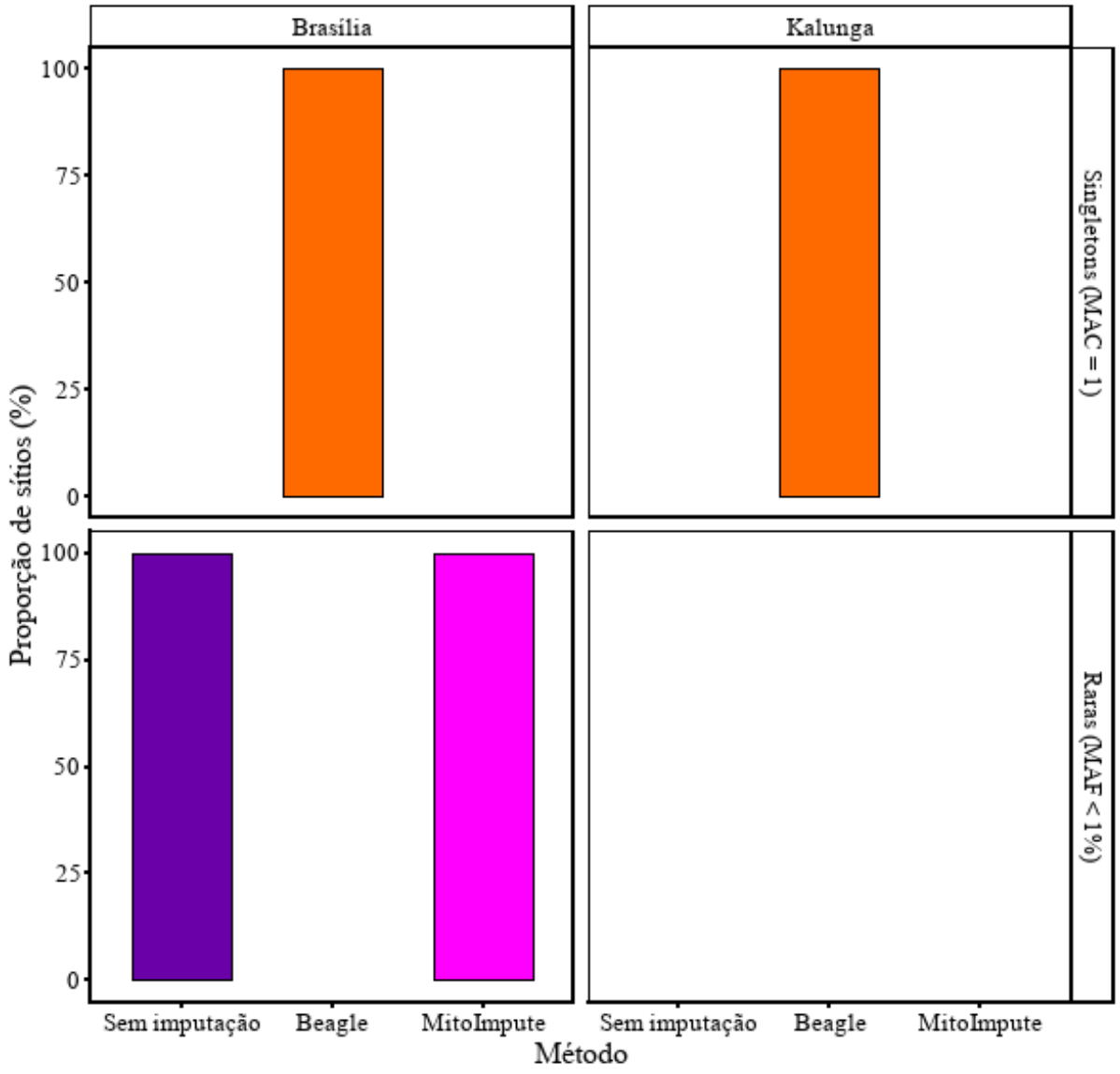
Fonte: O autor (2026).

Nota: Cada painel representa um método de inferência, com cores correspondentes aos métodos: laranja (Beagle), magenta (MitoImpute) e roxo (Sem imputação).

A Figura 20 evidencia divergências entre a quantificação de variantes raras (MAF < 1%) e de *singletons*, indicando que essas duas métricas capturam aspectos distintos da variabilidade genética presente nos dados. As variantes raras são definidas com base na frequência alélica menor que 1% na população analisada, podendo estar presentes em poucos indivíduos, mas ainda assim ocorrer mais de uma vez no conjunto amostral. Já os *singletons* correspondem a variantes observadas apenas uma única vez no conjunto de dados, ou seja, com contagem de alelo menor (*Minor Allele Count*, MAC) igual a 1. Dessa forma, é possível que uma análise apresente número relativamente elevado de variantes raras sem necessariamente apresentar grande quantidade de *singletons*, ou vice-versa. Essa diferença pode refletir tanto características da estrutura populacional quanto efeitos metodológicos associados à imputação e ao tamanho amostral, que influenciam a forma como variantes de baixa frequência são

detectadas e distribuídas entre os indivíduos analisados. Em Brasília, o MitoImpute apresentou maior proporção de variantes raras, enquanto o Beagle concentrou maior proporção de singletons. Em Kalunga, observou-se acúmulo mais expressivo de *singletons* no conjunto imputado pelo Beagle. Essas diferenças metodológicas têm implicações diretas para análises subsequentes, pois variantes extremamente raras influenciam estimativas de diversidade nucleotídica, Tajima's D e inferências demográficas. Estudos recentes mostram que a inclusão de variantes raras pode aumentar o poder para detectar eventos de expansão populacional recente, mas também pode introduzir ruído se a imputação gerar artefatos (BROWNING; BROWNING, 2016; DE MARINO *et al*., 2022). Portanto, a interpretação de padrões de frequência deve considerar tanto a genética populacional quanto as propriedades estatísticas do método de imputação.

Figura 20 – Frequência relativa de variantes raras no genoma mitocondrial segundo método e população.

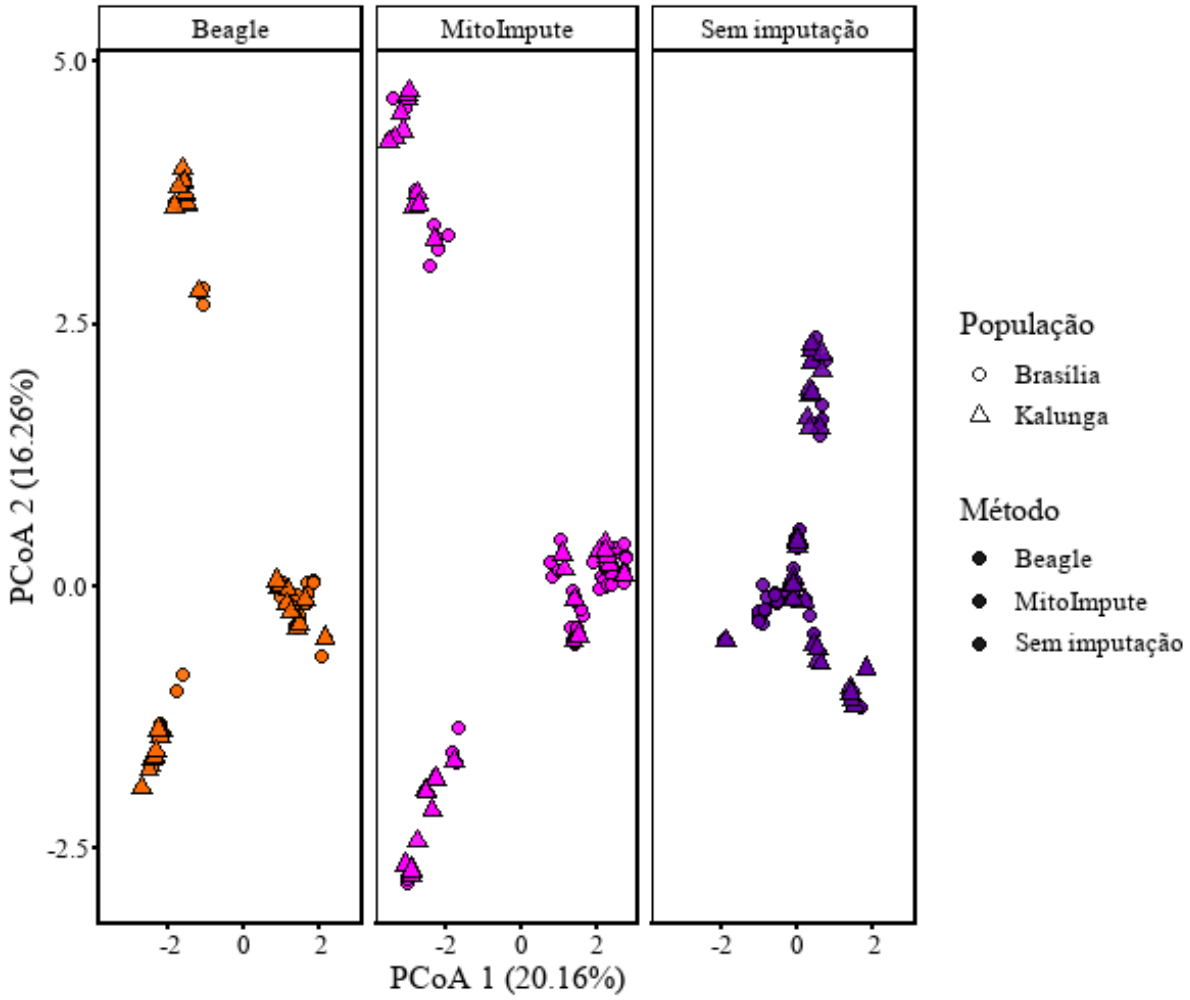


Fonte: O autor (2026).

Nota: Proporção de sítios singletons (MAC = 1) e raros (MAF < 1%) em Brasília e Kalunga, para os conjuntos sem imputação e imputados por Beagle e MitoImpute. As proporções (%) foram calculadas por população e método.

A PCoA (Figura 21) corresponde a distância genética e demonstrou melhora estrutural após a imputação. Esse aumento de resolução é consistente com o princípio de que maior número de sítios informativos melhora a estimativa de distâncias genéticas e a capacidade de detectar estrutura populacional sutil (EXCOFFIER *et al*., 1992)

Figura 21 – Análise de Coordenadas Principais (PCoA) baseada em variantes mitocondriais.



Fonte: O autor (2026).

Nota: Representação bidimensional da PCoA construída a partir de uma matriz de distância genética entre perfis mitocondriais, considerando presença e ausência de variantes. Cada ponto representa um indivíduo, com símbolos distintos para cada população (BSB = Brasília; KAL = Kalunga). Os valores entre parênteses indicam a proporção da variância explicada por cada eixo.

O Beagle maximiza o número total de sítios e tende a aumentar a ocorrência de *singletons*, enquanto o MitoImpute apresenta maior sensibilidade para variantes raras distribuídas ao longo do genoma. Do ponto de vista evolutivo, a recuperação de variantes raras é particularmente relevante para estudos de filogenia e história demográfica recente, dado que a maioria das mutações mitocondriais humanas é de baixa frequência (DE MARINO *et al.*, 2022).

5.5 Estrutura populacional e diferenciação genética

A estrutura genética foi avaliada entre as duas populações, Brasília (BSB) e Kalunga (KAL), com base em 591 sítios mitocondriais imputados, utilizando como metodologia a distância genética, frequências alélicas e particionamento da variância.

A diversidade mitocondrial observada nas duas populações foi elevada, porém com padrões distintos que refletem suas histórias demográficas. A população de Brasília apresentou diversidade haplotípica alta (Hd = 0,9949), com 90 haplótipos distintos em 105 indivíduos, valor compatível com populações urbanas brasileiras altamente miscigenadas e formadas por múltiplos fluxos migratórios maternos ao longo do tempo. Na população Kalunga, embora a diversidade haplotípica também tenha sido alta (Hd = 0,9830), o número de haplótipos foi menor (51 haplótipos em 70 indivíduos), sugerindo maior compartilhamento de linhagens maternas entre os indivíduos. Esse padrão é compatível com o histórico de relativo isolamento geográfico e sociocultural das comunidades quilombolas.

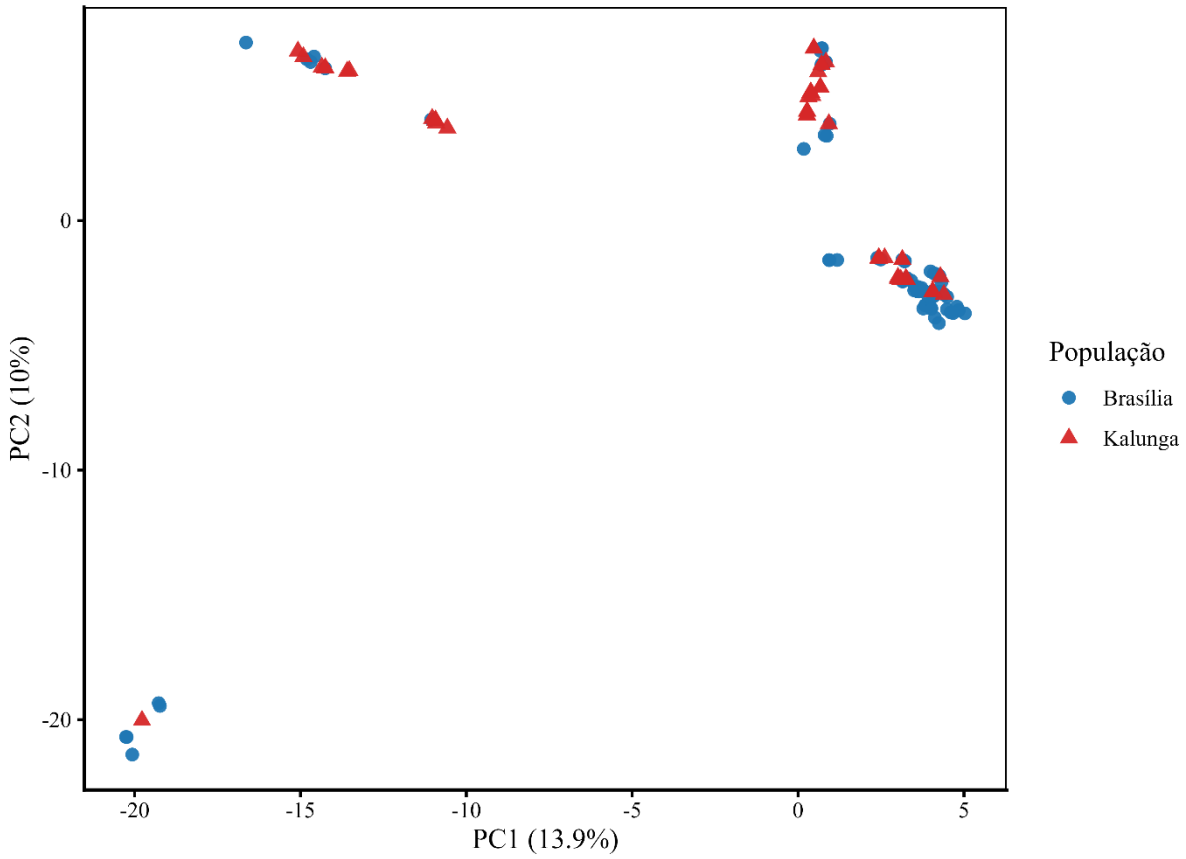
A diversidade nucleotídica foi ligeiramente maior em Kalunga ($\pi = 0,07284$; $k = 43,05$) do que em Brasília ($\pi = 0,06849$; $k = 40,48$), indicando maior divergência média entre as sequências mitocondriais dessa população. Esse resultado é particularmente informativo, pois revela que, embora Kalunga apresente menor número de haplótipos, as linhagens presentes são, em média, mais diferenciadas entre si. Esse padrão é frequentemente associado a populações com maior contribuição de haplogrupos africanos do macro-haplogrupo L, que apresentam maior profundidade temporal e maior distância filogenética entre suas sublinhagens quando comparados a o haplogrupos (SOARES *et al*., 2012). Em contraste, populações altamente miscigenadas, como Brasília, tendem a apresentar muitos haplótipos derivados de um número maior de ramos filogenéticos, porém com menor divergência média entre as sequências, em função de expansões populacionais recentes e do aporte contínuo de linhagens relacionadas.

A interpretação dos valores de Tajima's D permite integrar esses achados em um modelo histórico-demográfico. Valores negativos desse índice são compatíveis com expansão populacional recente ou seleção purificadora, resultando em excesso de variantes raras, enquanto valores positivos indicam redução populacional, estruturação interna ou efeitos mais pronunciados de deriva genética (TAJIMA, 1989). Considerando o contexto das populações analisadas, um valor negativo em Brasília é esperado e biologicamente plausível, refletindo crescimento populacional recente associado à formação da capital federal e à intensa migração de diferentes regiões do país, o que gera aumento do número de haplótipos e acúmulo de variantes de baixa frequência. Por outro lado, valores próximos de zero ou positivos em Kalunga são compatíveis com a história de isolamento relativo, menor tamanho efetivo e

estrutura populacional, condições que favorecem a manutenção de variantes em frequências intermediárias e reduzem o sinal de expansão recente.

A análise de componentes principais (PCA) baseada nas variantes mitocondriais imputadas (Figura 22) revelou uma estrutura genética caracterizada por diferenciação parcial entre as populações de Brasília e Kalunga, com os dois primeiros eixos explicando 23,9% da variância total (PC1 = 13,9%; PC2 = 10,0%). Observa-se a formação de agrupamentos com sobreposição entre os indivíduos das duas populações, indicando compartilhamento de linhagens maternas e refletindo a história de miscigenação que caracteriza a formação da população brasileira. Ainda assim, nota-se uma tendência de separação ao longo do PC1 e uma maior dispersão dos indivíduos Kalunga no espaço multivariado, enquanto parte dos indivíduos de Brasília se concentra em um cluster mais compacto. Esse padrão sugere maior estruturação interna e possível efeito de deriva genética na população quilombola, enquanto que a população de Brasília apresenta maior homogeneidade decorrente do fluxo gênico contínuo e da diversidade de origens maternas. A presença de pontos extremos em ambas as populações indica a ocorrência de linhagens mitocondriais mais divergentes.

Figura 22 – Análise de Componentes Principais (PCA) das amostras mitocondriais das populações de Brasília e Kalunga. utilizando SNPs mitocondriais imputados pelo MitoImpute.

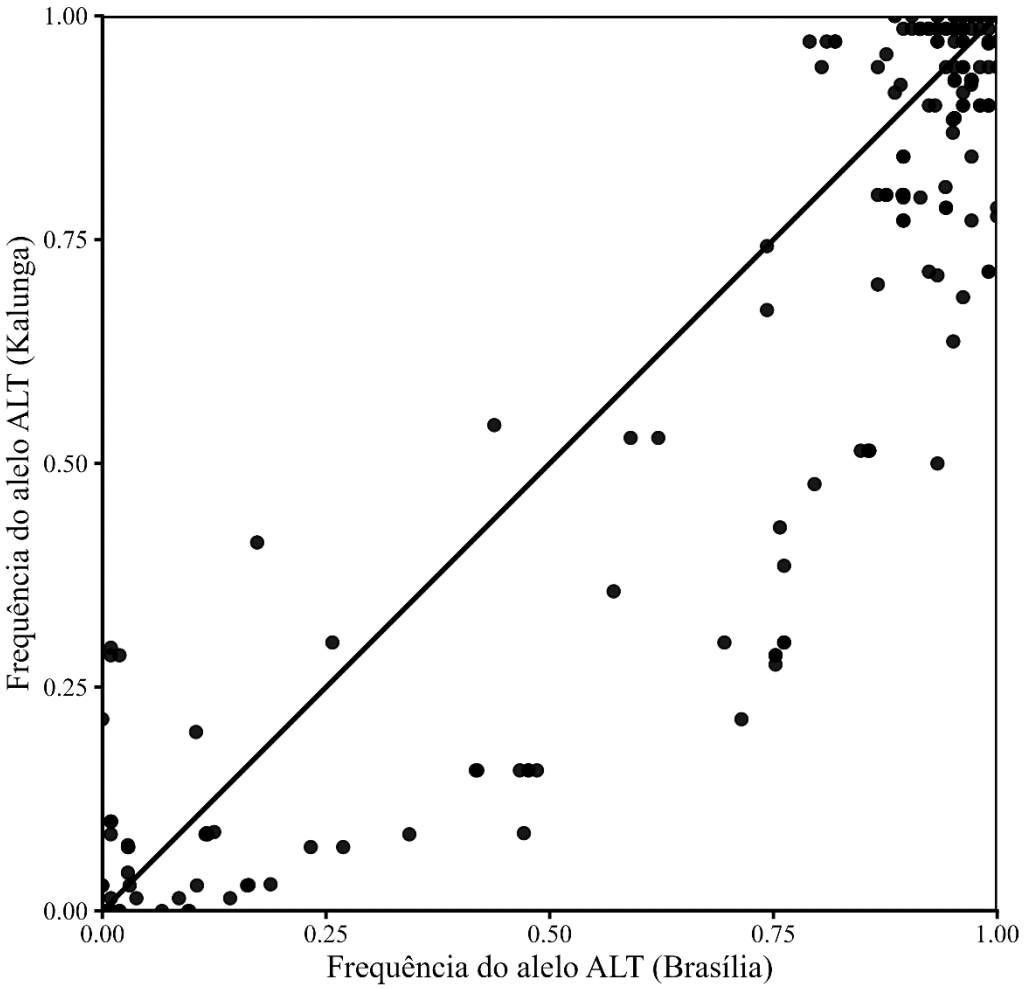


Fonte: O autor (2026).

Nota: Os pontos representam os indivíduos analisados, sendo os círculos azuis correspondentes à população de Brasília e os triângulos vermelhos à população Kalunga. Os eixos indicam as duas primeiras componentes principais, PC1 (13,9%) e PC2 (10%), que juntas explicam 23,9% da variação total dos dados.

A comparação das frequências alélicas entre os grupos analisados (Figura 23) expõe uma correlação positiva, ainda que haja uma dispersão ampla em torno da linha de identidade. Vários SNVs apresentam frequências discrepantes entre as populações, indicando que parte da diferenciação é atribuível a variações específicas de frequência, e não apenas à composição global de haplogrupos.

Figura 23 – Comparação das frequências do alelo alternativo (ALT) entre Brasília (BSB) e Kalunga (KAL) para SNPs mitocondriais imputados.

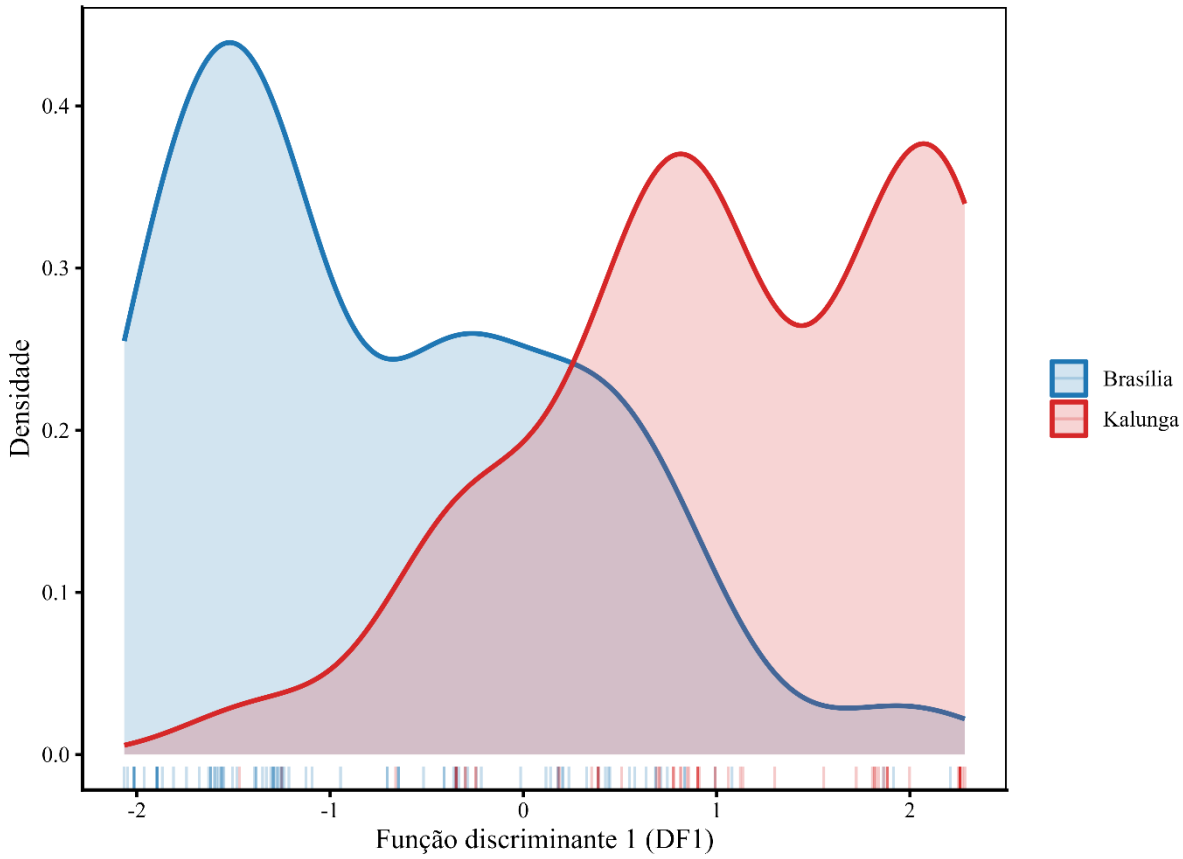


Fonte: O autor (2026).

Nota: Cada ponto representa um sítio SNV mitocondrial plotado de acordo com sua frequência do alelo alternativo (ALT) em Brasília (eixo X) e em Kalunga (eixo Y). A linha diagonal representa a identidade (y = x), indicando frequências iguais nas duas populações.

A análise discriminante (Figura 24) reforça esse padrão estrutural. As distribuições da função discriminante 1 (DF1) mostram deslocamento claro entre as populações, com sobreposição parcial. A separação das densidades indica que a composição mitocondrial permite discriminação estatística entre BSB e KAL, embora não absoluta.

Figura 24 – Distribuição da função discriminante (DF1) para as populações de Brasília (BSB) e Kalunga (KAL) com base em SNPs mitocondriais imputados.



Fonte: O autor (2026).

Nota: A figura apresenta as curvas de densidade da primeira função discriminante (DF1) obtida a partir de análise discriminante baseada nas frequências alélicas mitocondriais. A área azul representa Brasília (BSB) e a área vermelha representa Kalunga (KAL). As marcas na base do gráfico indicam a posição individual das amostras ao longo do eixo discriminante.

A AMOVA resultou 8,42% da variação genética total é atribuível à diferença entre populações (F_{ST} = 0,0842; p = 0,001), enquanto 91,58% ocorre dentro das populações. A PERMANOVA apresentou resultado convergente (R^2 = 0,067; F = 12,45; p = 0,001), indicando que aproximadamente 6,7% da variação na matriz de distâncias é explicada pela população de origem. A estimativa baseada em frequências alélicas simples (GST = 0,0350; p = 0,001) também foi significativa, embora numericamente inferior ao F_{ST} — diferença esperada, pois F_{ST} incorpora informação de distância genética entre haplótipos, enquanto GST considera apenas diferenças de frequência.

Valores de F_{ST} entre 0,05 e 0,15 são classicamente interpretados como diferenciação moderada (WRIGHT, 1978), compatível com marcadores uniparentais. Assim, o valor observado ($F_{ST} \approx 0{,}084$) indica que cerca de 8% da variação mitocondrial reflete separação histórica entre BSB e KAL.

A comunidade Kalunga apresenta maior isolamento geográfico e social, o que resulta num diferencial de determinadas linhagens maternas. Já Brasília, fundada em 1960, resulta de intenso fluxo migratório interno no país, reunindo indivíduos miscigenados de diferentes regiões. Esse contraste demográfico explica a combinação de alta diversidade intrapopulacional com diferenciação moderada entre os grupos. Os resultados também são coerentes com a literatura que demonstram elevada diversidade interna combinada a diferenciação regional detectável (ALVES-SILVA *et al.*, 2000; PENA *et al*., 2011). Globalmente, análises de variância molecular mostram que a maior parte da diversidade humana é compartilhada dentro das populações, mas diferenças históricas regionais permanecem detectáveis, especialmente em marcadores uniparentais (WALLACE *et al*., 1999).

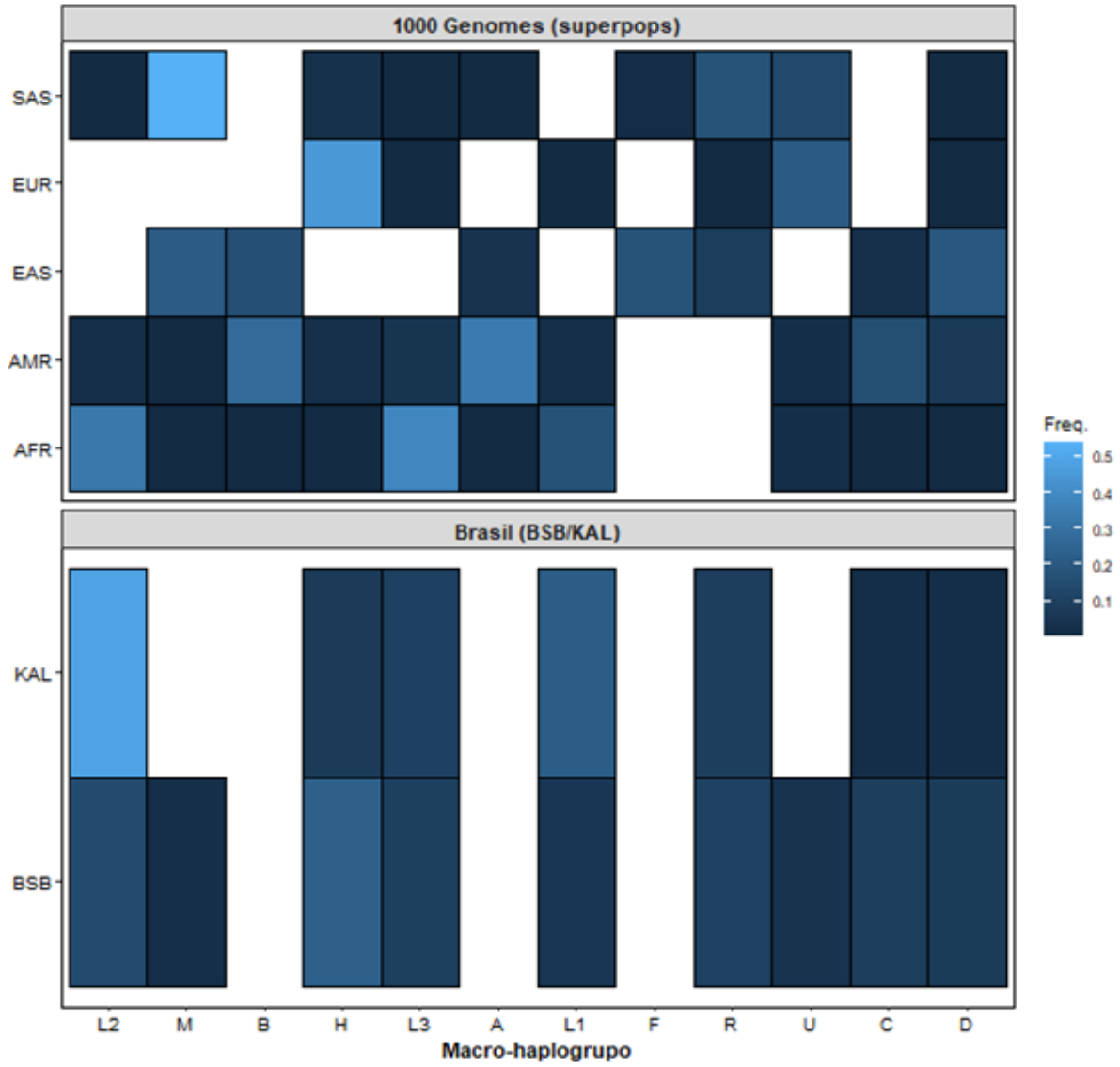
5.6 Comparação com populações de referência

Para realizar a comparação com outras populações, empregou-se os dados imputados pela pipeline MitoImpute, com as superpopulações do 1000 Genomes Project, a fim de contextualizar as coortes brasileiras, Brasília e Kalunga, dentro do panorama global de diversidade mitocondrial. Os dados imputados foram escolhidos para essa análise visto que a imputação aumentou a confiabilidade da atribuição de linhagens, especialmente em indivíduos classificados em haplogrupos amplos devido ao baixo número de SNPs observados.

Na comparação das frequências de macro-haplogrupos mitocondriais utilizou cinco superpopulações do consórcio (AFR, EUR, EAS, SAS e AMR) (1000 GENOMES PROJECT CONSORTIUM, 2015). O padrão do conjunto de referência atribui filogeneticamente os macro-haplogrupos L a população africana (AFR), macro-haplogrupos de H, U, J, T e R a europeia (EUA), macro-haplogrupos M e seus derivados a leste e sul asiática (EAS e SAS), e macro-haplogrupos A, B, C, D americana (AMR).

Na Figura 25 observa-se o resultado da comparação revelando uma maior contribuição africana e europeia, e menor frequência indígena e asiática, compatível com a história demográfica das duas populações analisadas. Kalunga apresentou predominância do macro-haplogrupo africano L2, reforçando evidências de continuidade genética em comunidades quilombolas, apesar de séculos de miscigenação. Por outro lado, Brasília apresentou uma composição heterogênea, porém com frequência elevada do macro-haplogrupo europeu H.

Figura 25 – Frequência relativa de macro-haplogrupos mitocondriais nas superpopulações do 1000 Genomes e nas coortes brasileiras (BSB e KAL).



Fonte: O autor (2026).

Nota: As siglas referem as superpopulações do 1000 Genomes Project (AFR – africana; AMR – americana; EAS – leste asiática; EUR – europeia; SAS – sul asiática). O painel inferior mostra as frequências correspondentes nas populações brasileiras analisadas: Brasília (BSB) e Kalunga (KAL). A intensidade da coloração azul representa a frequência relativa de cada macro-haplogrupo dentro de cada população, conforme escala indicada (Freq.). Células em branco indicam ausência do macro-haplogrupo na respectiva população.

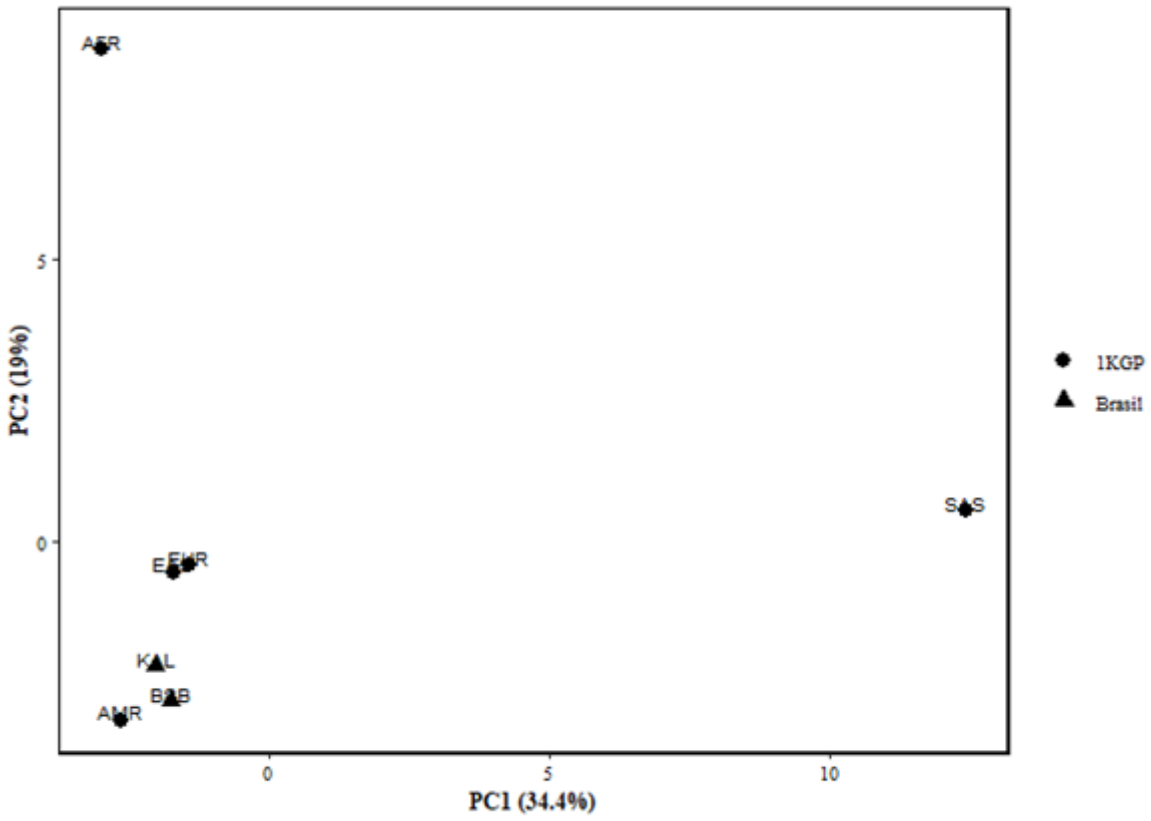
A Análise de Componentes Principais (PCA), baseada nas frequências de macro-haplogrupos, evidenciou uma separação entre as populações do 1000 Genomes, com PC1 explicando 34,4% da variância e PC2 explicando 19% (Figura 26). O eixo PC1, principalmente população sul asiática (SAS), diverge das demais, enquanto o eixo PC2 destacou a diferenciação africana (AFR). Esse padrão é compatível com a profunda divergência

filogenética das linhagens L em relação aos macro-haplogrupos não africanos, que derivam de L3 e representam expansões pós a saída do *Homo sapiens* da África (*Out of Africa*) (SOARES *et al*., 2012).

Ainda que se esperasse a proximidade da comunidade Kalunga ao cluster AFR, devido a elevada frequência de macro-haplogrupos L, isso não aconteceu. Isso se dá ao fato de AFR ter um padrão homogêneo, sem contribuição de outros haplogrupos, enquanto o AMR que está relativamente próximo já possui um padrão mais miscigenado. Além disso, a distribuição relativa entre L1, L2 e L3 difere do padrão médio africano do 1000 Genomes. Portanto, o padrão é consistente com gargalos populacionais e isolamento relativo, correspondente da história demográfica dessa população.

Outrossim, a população de Brasília se posiciona próxima de KAL, levemente deslocada ao centro, por causa da contribuição europeia. Vale ressaltar que a PCA baseada em frequências capta estrutura populacional agregada, não diversidade intra-populacional. Avila *et al*. (2019) ao analisar mitogenoma de algumas regiões brasileiras, como o Centro-Oeste, teve como resultado alta proximidade com a população afro-americana, baixa proximidade com as populações europeias e proximidade intermediária com as populações do Sul da Ásia.

Figura 26 – Análise de Componentes Principais (PCA) baseada nas frequências de macro-haplogrupos mitocondriais das superpopulações do 1000 Genomes e das populações brasileiras (BSB e KAL).



Fonte: O autor (2026).

Nota: Os círculos indicam as superpopulações do 1000 Genomes Project (AFR – africana; AMR – americana; EAS – leste asiática; EUR – europeia; SAS – sul asiática), enquanto os triângulos representam as populações brasileiras analisadas (Brasília – BSB; Kalunga – KAL). A proximidade entre os pontos reflete maior similaridade na composição de macro-haplogrupos mitocondriais.

A análise da diferenciação populacional mitocondrial, estimada pelo índice F_{ST} resultou em torno de 0,08 quando comparada ao conjunto global das populações do 1000 Genomes, situando dentro da distribuição observada entre pares intercontinentais e próximo a média dessas comparações. Assim, o F_{ST} observado anui que a população brasileira possui identidade mitocondrial própria, moldada por eventos de fundação, miscigenação e deriva genética, mantendo, entretanto, grau de diferenciação compatível com o espectro de variação intercontinental descrito na literatura (SILVA *et al*., 2000; NUNES *et al*., 2025).

6 **CONCLUSÃO**

Este trabalho é o primeiro estudo a analisar a imputação de mitogenoma na população brasileira com intuito de avaliar a diversidade e estrutura genética mitocondrial. De forma geral, os resultados demonstraram que a baixa densidade de marcadores mitocondriais compromete a resolução filogenética e a qualidade da inferência dos haplogrupos, enquanto que o método de imputação MitoImpute promoveu melhorias significativas tanto na confiabilidade das inferências quanto na recuperação da diversidade materna.

A partir do conjunto reduzido de variantes, inicialmente, analisado com o HaploGrep3 evidenciou-se limitações refletidas nas baixas métricas de qualidade, elevada proporção de falhas e na supergeneralização das classificações haplogrupais. Isso confirma que painéis com número restrito de SNPs mitocondriais são insuficientes para atribuições filogenética, resultando em inconsistências com a variabilidade esperada para populações humanas.

Ambos os programas utilizados aumentaram significativamente o número de variantes e elevando as métricas de qualidade das inferências. No entanto, o Beagle destacou-se pela elevada densidade de variantes imputadas, enquanto o MitoImpute apresentou maior consistência qualitativa nas classificações. O fato da pipeline MitoImpute ser desenvolvida para imputação do mitogenoma pode explicar esse desempenho superior em termos de estabilidade classificatória.

A análise das transições de classificação por indivíduo evidenciou que a imputação modificou os níveis mais gerais para subclados. Esse resultado demonstra que a imputação reduz a perda de informações causada pela baixa densidade de marcadores e aumenta a precisão da inferência, sem introduzir vieses na ancestralidade materna.

As estimativas de ancestralidade materna demonstraram que as populações analisadas possuem padrões distintos e de acordo com seus processos históricos de formação. A população de Brasília apresentou uma composição heterogênea, com contribuição de linhagens europeias, africanas, indígenas e, minoritariamente, asiática. Em contraste, a população Kalunga apresentou predominância de linhagens africanas, mas com contribuição indígena e europeia, o que está em consonância com sua história de formação e relativo isolamento inicial, seguido por interações com regiões adjacentes. A presença de haplogrupos compartilhados entre as duas populações indica a existência de um componente comum da formação genética brasileira, ainda que em frequências distintas.

As análises de diversidade genética e estrutura populacional reforçaram que ambas possuem alta diversidade haplotípica, porém com padrões distintos de variação nucleotídica e diferenciação genética. Os valores de F_{ST} indicaram estruturação moderada, corroborada pelos resultados da AMOVA, demonstrando que, embora compartilhe parte da história demográfica, Kalunga e Brasília possuem composições maternas diferentes.

Dessa maneira, este trabalho ressalva que, a imputação com MitoImpute é uma abordagem eficiente para estudos populacionais, além de evidenciar que a variabilidade do mtDNA reflete os processos históricos e demográficos que moldaram as populações brasileiras, ainda que sejam divergentes. Assim, este estudo contribui metodologicamente ao demonstrar o desempenho e aplicabilidade da imputação mitogenômica em populações sub-representadas e que tenham poucos marcadores informativos para análise.

REFERÊNCIAS

ABÊ-SANDES, K. *Diversidade genética de afro-brasileiros: DNA mitocondrial e cromossomo Y*. 2002. Tese (Doutorado) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2002.

ACHILLI, A.; PEREGO, U. A.; BRAVI, C. M.; COBLE, M. D.; KONG, Q. P.; WOODWARD, S. R.; SALAS, A.; TORRONI, A.; BANDELT, H. J. *The phylogeny of the four pan-American mtDNA haplogroups: implications for evolutionary and disease studies*. PLOS ONE, v. 3, n. 3, e1764, 2008. DOI: https://doi.org/10.1371/journal.pone.0001764.

ALVES-SILVA, J.; DA SILVA SANTOS, M.; GUIMARÃES, P. E.; FERREIRA, A. C.; BANDELT, H. J.; PENA, S. D.; PRADO, V. F. *The ancestry of Brazilian mtDNA lineages.* American Journal of Human Genetics, v. 67, n. 2, p. 444–461, ago. 2000. DOI: https://doi.org/10.1086/303004 .

ANAIS do III Encontro da Rede da Comunidade Kalunga. Pontifícia Universidade Católica de Goiás (PUC-Goiás), Associação Quilombo Kalunga, Rede Kalunga, Universidade Federal de Goiás (UFG), Universidade Federal do Tocantins (UFT) – Campus de Arraias-TO, 2017. Disponível em: https://files.cercomp.ufg.br/weby/up/133/o/Anais_3_Kalunga_18_11_%281%29.pdf. Acesso em: 23 nov. 2025.

ANDERSON, S.; BANKIER, A. T.; BARRELL, B. G.; DE BRUIJN, M. H. L.; COULSON, A. R.; DROUIN, J.; EPERON, I. C.; NIERLICH, D. P.; ROE, B. A.; SANGER, F.; SCHREIER, P. H.; SMITH, A. J. H.; STADEN, R.; YOUNG, I. G. *Sequence and organization of the human mitochondrial genome.* Nature, v. 290, n. 5806, p. 457–465, 1981. DOI: https://doi.org/10.1038/290457a0 .

AVILA, E.; GRAEBIN, P.; CHEMALE, G.; FREITAS, J.; KAHMANN, A.; ALHO, C. S. *Full mtDNA genome sequencing of Brazilian admixed populations: a forensic-focused evaluation of a MPS application as an alternative to Sanger sequencing methods*. Forensic Science International: Genetics, v. 42, p. 154–164, set. 2019. DOI: https://doi.org/10.1016/j.fsigen.2019.07.004.

BAIOCCHI, M. de N. *Kalunga: povo da terra.* Goiânia: UFG, 2006. 132 p.

BEDOYA, G.; MONTINARO, F.; GARCÍA, J.; SOTO, I.; BOURGUET, C.; CARVAJAL-CARMONA, L. G.; MORENO-ESTRADA, A.; RUIZ-LINARES, A. *Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate.* Proceedings of the National Academy of Sciences of the United States of America, v. 103, n. 19, p. 7234–7239, 2006. DOI: https://doi.org/10.1073/pnas.0508716103 .

BETHELL, L. Nota sobre as populações americanas às vésperas das invasões europeias. In: BETHELL, L. (org.). *América Latina colonial.* São Paulo: Editora da Universidade de São Paulo, 1997.

BROWNING, B. L.; BROWNING, S. R. *Genotype imputation with millions of reference samples.* American Journal of Human Genetics, v. 98, n. 1, p. 116–126, jan. 2016. DOI: https://doi.org/10.1016/j.ajhg.2015.11.020 .

BROWNING, B. L.; ZHOU, Y.; BROWNING, S. R. *A one-penny imputed genome from next-generation reference panels.* American Journal of Human Genetics, v. 103, n. 3, p. 338–348, 2018. DOI: https://doi.org/10.1016/j.ajhg.2018.07.015 .

BUDOWLE, B.; ALLARD, M. W.; WILSON, M. R.; CHAKRABORTY, R. *Forensics and mitochondrial DNA: applications, debates, and foundations.* Annual Review of Genomics and Human Genetics, v. 4, p. 119–141, 2003. DOI: https://doi.org/10.1146/annurev.genom.4.070802.110352 .

CAMPBELL, M. C.; TISHKOFF, S. A. *African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping.* Annual Review of Genomics and Human Genetics, v. 9, p. 403–433, 2008. DOI: https://doi.org/10.1146/annurev.genom.9.081307.164258 .

CANN, R. L.; STONEKING, M.; WILSON, A. C. *Mitochondrial DNA and human evolution.* Nature, v. 325, n. 6099, p. 31–36, jan. 1987. DOI: https://doi.org/10.1038/325031a0.

CHEN, Y. S.; TORRONI, A.; EXCOFFIER, L.; SANTACHIARA-BENERECETTI, A. S.; WALLACE, D. C. *Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups.* American Journal of Human Genetics, v. 57, n. 1, p. 133–149, jul. 1995. PMID: 7611282. Disponível em: https://pubmed.ncbi.nlm.nih.gov/7611282/. Acesso em: 26 fev. 2025

COSTA, V. S. *A luta pelo território: histórias e memórias do povo Kalunga.* 2013. 75 p. Trabalho de Conclusão de Curso (Licenciatura em Educação do Campo) – Universidade de Brasília, Brasília, 2013.

CROOCK, D.; UREN, C.; MÖLLER, M.; PETERSEN, D. C. *Mitochondrial DNA imputation accuracy and its application in a Southern African mitochondrial genome-wide association study.* bioRxiv, 2025. DOI: https://doi.org/10.64898/2025.12.11.693816 .

DAIGO, M. *Pequena história da imigração japonesa no Brasil.* Tradução: M. Ninomiya. São Paulo: Gráfica Paulos, 2008.

DE MARINO, A.; MAHMOUD, A. A.; BOSE, M.; BIRCAN, K. O.; TERPOLOVSKY, A.; BAMUNUSINGHE, V.; HYSI, P. G.; HAMMOND, C. J.; YOUNG, T. L.; KHAWAJA, A. P. *A comparative analysis of current phasing and imputation software*. PLOS ONE, v. 17, n. 10, e0260177, 2022. DOI: https://doi.org/10.1371/journal.pone.0260177.

DORJI, J.; CHAMBERLAIN, A. J.; REICH, C. M.; VANDERJAGT, C. J.; NGUYEN, T. V.; DAETWYLER, H. D.; MACLEOD, I. M. *Mitochondrial sequence variants: testing imputation accuracy and their association with dairy cattle milk traits*. Genetics Selection Evolution, v. 56, n. 1, p. 62, 2024. DOI: https://doi.org/10.1186/s12711-024-00931-5.

ESCHER, L. M.; NASLAVSKY, M. S.; SCLIAR, M. O.; HUTZ, M. H.; RIBEIRO-DOS-SANTOS, A. K.; SALZANO, F. M.; SEUÁNEZ, H. N.; PENA, S. D. J. *Challenges in selecting admixture models and marker sets to infer genetic ancestry in a Brazilian admixed population*. Scientific Reports, v. 12, art. 21240, 2022. DOI: https://doi.org/10.1038/s41598-022-25521-7.

EXCOFFIER, L.; LISCHER, H. E. L. *Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows.* Molecular Ecology Resources, v. 10, n. 3, p. 564–567, maio 2010. DOI: https://doi.org/10.1111/j.1755-0998.2010.02847.x.

EXCOFFIER, L.; SMOUSE, P. E.; QUATTRO, J. M. *Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data.* Genetics, v. 131, n. 2, p. 479–491, jun. 1992. DOI: https://doi.org/10.1093/genetics/131.2.479.

FERREIRA, L. B. *Diversidade do DNA mitocondrial de populações brasileiras ameríndias e afrodescendentes*. 2006. 197 f. Tese (Doutorado em Genética e Biologia Molecular) – Universidade de Brasília, Brasília, 2006. Disponível em: https://www.oasisbr.ibict.br/vufind/Record/BRCRIS_1a920b6bc5207dcb9cfb181d0f194845. Acesso em: 04 out. 2025.

FREITAS, J. M.; FASSIO, L. H.; BRAGANHOLI, D. F.; CHEMALE, G. *Mitochondrial DNA control region haplotypes and haplogroup diversity in a sample from Brasília, Federal District, Brazil.* Forensic Science International: Genetics, v. 40, p. e228–e230, maio 2019. DOI: https://doi.org/10.1016/j.fsigen.2019.02.006.

GAO, G. *Practical consideration of genotype imputation: sample size, window size, reference choice, and untyped rate.* Statistics and Its Interface, v. 4, n. 3, p. 339–351, 2011. DOI: https://doi.org/10.4310/SII.2011.v4.n3.a8.

GILES, R. E.; BLANC, H.; CANN, H. M.; WALLACE, D. C. *Maternal inheritance of human mitochondrial DNA.* Proceedings of the National Academy of Sciences of the United States of America, v. 77, n. 11, p. 6715–6719, nov. 1980. DOI: https://doi.org/10.1073/pnas.77.11.6715 .

GONÇALVES, Anna Beatriz Rodrigues. *Análise genética das linhagens matrilíneas de indivíduos maranhenses para fins de bancos de dados populacional e forense*. 2020. Dissertação (Mestrado em Biociências) — Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2020. Disponível em:

https://www.bdtd.uerj.br:8443/bitstream/1/21551/2/Disserta%C3%A7%C3%A3o%20-%20Anna%20Beatriz%20Rodrigues%20Gon%C3%A7alves%20-%202020%20-%20Completa.pdf. Acesso em: 07 out. 2025.

GONÇALVES, V. F.; CARVALHO, C. M. B.; BORTOLINI, M. C.; BYDLOWSKI, S. P.; PENA, S. D. J. *The phylogeography of African Brazilians.* Human Heredity, v. 57, n. 1, p. 42–50, 2007. DOI: https://doi.org/10.1159/000106059.

GONÇALVES, V. F.; CARVALHO, C. M. B.; BORTOLINI, M. C.; BYDLOWSKI, S. P.; PENA, S. D. J. *The phylogeography of African Brazilians*. Human Heredity, v. 65, n. 1, p. 23–32, 2008. DOI: https://doi.org/10.1159/000106059.

GONTIJO, C. C.; MENDES, F. M.; SANTOS, C. A.; KLAUTAU-GUIMARÃES, M. N.; LAREU, M. V.; CARRACEDO, Á.; PHILLIPS, C.; OLIVEIRA, S. F. *Ancestry analysis in rural Brazilian populations of African descent.* Forensic Science International: Genetics, v. 36, p. 160–166, 2018. DOI: https://doi.org/10.1016/j.fsigen.2018.06.007.

HARTL, D. L.; CLARK, A. G. *Princípios de genética de populações*. 4. ed. Porto Alegre: Artmed, 2010. 660 p.

HOLLAND, M. M.; PARSONS, T. J. *Mitochondrial DNA sequence analysis: validation and use for forensic casework.* Forensic Science Review, v. 11, n. 1, p. 21–50, jun. 1999. PMID: 26255820. Disponível em: https://pubmed.ncbi.nlm.nih.gov/26255820/. Acesso em: 12 jun. 2025.

JOERIN-LUQUE, I. A.; AUGUSTO, D. G.; CALONGA-SOLÍS, V.; DE ALMEIDA, R. C.; LOPES, C. V. G.; PETZL-ERLER, M. L.; BELTRAME, M. H. *Uniparental markers reveal new insights on subcontinental ancestry and sex-biased admixture in Brazil.* Molecular Genetics and Genomics, v. 297, n. 2, p. 419–435, mar. 2022. DOI: https://doi.org/10.1007/s00438-022-01857-7.

KABISCH, M.; HAMANN, U.; BERMEJO, J. L. *Imputation of missing genotypes within LD-blocks relying on the basic coalescent and beyond: consideration of population growth and structure.* BMC Genomics, v. 18, 798, 2017. DOI: https://doi.org/10.1186/s12864-017-4208-2.

KEHDY, F. S. G.; GOUVEIA, M. H.; MACHADO, M.; MAGALHÃES, W. C. S.; HORIMOTO, A. R.; HORTA, B. L.; MOREIRA, R. G.; LEAL, T. P.; SCLIAR, M. O.; SOARES-SOUZA, G. B.; RODRIGUES-SOARES, F.; ARAÚJO, G. S.; ZAMUDIO, R.; SANT ANNA, H. P.; SANTOS, H. C.; DUARTE, N. E.; FIACCONE, R. L.; FIGUEIREDO, C. A.; SILVA, T. M.; COSTA, G. N. O.; BELEZA, S.; BERG, D. E.; CABRERA, L.; DEBORTOLI, G.; DUARTE, D.; GHIROTTO, S.; GILMAN, R. H.; GONÇALVES, V. F.; MARRERO, A. R.; MUNIZ, Y. C.; WEISSENSTEINER, H.; YEAGER, M.; RODRIGUES, L. C.; BARRETO, M. L.; LIMA-COSTA, M. F.; PEREIRA, A. C.; RODRIGUES, M. R.; TARAZONA-SANTOS, E.; THE BRAZILIAN EPIGEN PROJECT CONSORTIUM. *Origin*

and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proceedings of the National Academy of Sciences of the United States of America, v. 112, n. 28, p. 8696–8701, 2015. DOI: https://doi.org/10.1073/pnas.1504447112 .

KIMURA, L.; RIBEIRO-RODRIGUES, E. M.; DE MELLO AURICCHIO, M. T.; VICENTE, J. P.; BATISTA SANTOS, S. E.; MINGRONI-NETTO, R. C. *Genomic ancestry of rural African-derived populations from Southeastern Brazil.* American Journal of Human Biology, v. 25, n. 1, p. 35–41, 2013. DOI: https://doi.org/10.1002/ajhb.22349.

LARICCHIA, K. M.; LAKE, N. J.; WATTS, N. A.; SHAND, M.; HAESSLY, A.; GAUTHIER, L.; BENJAMIN, D.; BANKS, E.; SOTO, J.; GARIMELLA, K.; EMERY, J.; GENOME AGGREGATION DATABASE CONSORTIUM; REHM, H. L.; MACARTHUR, D. G.; TIAO, G.; LEK, M.; MOOTHA, V. K.; CALVO, S. E. *Mitochondrial DNA variation across 56,434 individuals in gnomAD*. Genome Research, v. 32, n. 3, p. 569–582, mar. 2022. DOI: https://doi.org/10.1101/gr.276013.121.

LI, N.; STEPHENS, M. *Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data.* Genetics, v. 165, n. 4, p. 2213–2233, dez. 2003. DOI: https://doi.org/10.1093/genetics/165.4.2213.

LU, Y.; et al. *Technical design document for a SNP array that is optimized for population genetics*. Relatório técnico. Disponível em: ftp://ftp.cephb.fr/hgdp_supp10/8_12_2011_Technical_Array_Design_Document.pdf. Acesso em: 18 fev. 2026.

LUO, S.; VALENCIA, C. A.; ZHANG, J.; et al. *Biparental inheritance of mitochondrial DNA in humans*. Proceedings of the National Academy of Sciences, v. 115, n. 51, p. 13039–13044, 2018. DOI: https://doi.org/10.1073/pnas.1810946115.

MARCHINI, J.; HOWIE, B. *Genotype imputation for genome-wide association studies.* Nature Reviews Genetics, v. 11, n. 7, p. 499–511, 2010. DOI: https://doi.org/10.1038/nrg2796 .

MARCHINI, J.; HOWIE, B.; MYERS, S.; MCVEAN, G.; DONNELLY, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, v. 39, n. 7, p. 906–913, 2007. DOI: https://doi.org/10.1038/ng2088.

MCINERNEY, T. W.; FULTON-HOWARD, B.; PATTERSON, C.; et al. *A globally diverse reference alignment and panel for imputation of mitochondrial DNA variants.* BMC Bioinformatics, v. 22, 417, 2021. DOI: https://doi.org/10.1186/s12859-021-04337-8.

MISHMAR, D.; RUIZ-PESINI, E.; GOLIK, P.; MACAULAY, V.; CLARK, A. G.; HOSSEINI, S.; BRANDON, M.; EASLEY, K.; CHEN, E.; BROWN, M. D.; SUKERNIK, R. I.; OLCKERS, A.; WALLACE, D. C. *Natural selection shaped regional mtDNA variation in*

humans. Proceedings of the National Academy of Sciences of the United States of America, v. 100, n. 1, p. 171–176, jan. 2003. DOI: https://doi.org/10.1073/pnas.0136972100.

NASS, M. M. K.; NASS, S. *Intramitochondrial fibers with DNA characteristics. I. Fixation and electron staining reactions.* The Journal of Cell Biology, v. 19, n. 3, p. 593–611, 1 dez. 1963. DOI: https://doi.org/10.1083/jcb.19.3.593 **.**

NISHIDA, M. *Japanese immigration to Brazil.* Oxford Research Encyclopedia of Latin American History, 26 set. 2017. DOI: https://doi.org/10.1093/acrefore/9780199366439.013.423.

NONIN-LECOMTE, S.; DARDEL, F.; LESTIENNE, P. P. *Self-organisation of an oligodeoxynucleotide containing the G- and C-rich stretches of the direct repeats of the human mitochondrial DNA.* Biochimie, v. 87, n. 8, p. 725–735, 2005. DOI: https://doi.org/10.1016/j.biochi.2005.03.009 .

NUNES, K.; KIMURA, L.; GONTIJO, C. C.; ARCANJO, A. C.; MINGRONI-NETTO, R. C.; OLIVEIRA, S. F. de. *Estudos de dinâmica populacional, ancestralidade genética e saúde em comunidades quilombolas: relato de uma experiência.* Tessituras: Revista de Antropologia e Arqueologia, v. 8, n. 2, p. 218–251, dez. 2020. DOI: https://doi.org/10.15210/tes.v8i2.18396.

OJALA, D.; MONTOYA, J.; ATTARDI, G. *tRNA punctuation model of RNA processing in human mitochondria.* Nature, v. 290, n. 5806, p. 470–474, 1981. DOI: https://doi.org/10.1038/290470a0 .

OLIVEIRA JÚNIOR, A. N.; STUCCHI, D.; CHAGAS, M. F.; BRASILEIRO, S. S. *Laudo antropológico: comunidades negras de Ivaporunduva, São Pedro, Pedro Cubas, Sapatu, Nhunguara, André Lopes, Maria Rosa e Pilões – Vale do Rio Ribeira de Iguape – SP*. In: ANDRADE, T.; PEREIRA, C. A. C.; ANDRADE, M. R. O. (orgs.). *Negros do Ribeira: reconhecimento étnico e conquista do território.* São Paulo: ITESP, 2000. p. 39–192.

PALACIN, L. *O século do ouro em Goiás.* Goiânia: Universidade Católica de Goiás, 1994. 150 p.

PEDROSA, M. A. F. *Composição genética de quatro populações remanescentes de quilombos do Brasil com base em microssatélites e marcadores de ancestralidade.* 2006. Dissertação (Mestrado em Biologia Animal) – Universidade de Brasília, Brasília, 2006. Disponível em: https://files.cercomp.ufg.br/weby/up/133/o/2006_MariaAng%C3%A9licaFlorianoPedrosa.pdf#:~:text=de%20quilombos%20do%20Nordeste%20em%201998%2C%20e,autorizaram%20o%20trabalho%20e%20foi%20feita%20uma. Acesso em: 14 mar. 2025.

PENA, S. D.; DI PIETRO, G.; FUCHSHUBER-MORAES, M.; GENRO, J. P.; HUTZ, M. H.; KEHDY, F. D. S.; KOHLRAUSCH, F.; MAGNO, L. A.; MONTENEGRO, R. C.; MORAES, M. O.; MORAES, M. E.; MORAES, M. R.; OJOPI, E. B.; PERINI, J. A.; RACCIOLI, C.;

RIBEIRO-DOS-SANTOS, A. K.; RIOS-SANTOS, F.; ROMANO-SILVA, M. A.; SORTICA, V. A.; SUAREZ-KURTZ, G. *The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected.* PLOS ONE, v. 6, n. 2, e17063, fev. 2011. DOI: https://doi.org/10.1371/journal.pone.0017063 .

PEREIRA, L.; FREITAS, F.; FERNANDES, V.; PEREIRA, J. B.; COSTA, M. D.; COSTA, S.; MÁXIMO, V.; MACAULAY, V.; ROCHA, R.; SAMUELS, D. C. *The diversity present in 5140 human mitochondrial genomes*. The American Journal of Human Genetics, v. 84, n. 5, p. 628–640, 2009. DOI: https://doi.org/10.1016/j.ajhg.2009.04.013.

RIBEIRO-DOS-SANTOS, Ândrea K.; PEREIRA, Jaciléa M.; LOBATO, Mario R. L.; CARVALHO, Bruno M.; GUERREIRO, João F.; SANTOS, Sidney E. B. *Dissimilarities in the process of formation of Curiaú, a semi-isolated Afro-Brazilian population of the Amazon region*. American Journal of Human Biology, v. 14, n. 4, p. 440–447, jul.-ago. 2002. DOI: https://doi.org/10.1002/ajhb.10059

ROGERS, A. R.; HARPENDING, H. C. *Population growth makes waves in the* 552–569, 1992. PMID: 1316531. Disponível em: https://www.researchgate.net/publication/21838240_Rogers_AR_Harpending_HC_Population_growth_makes_waves_in_the_distribution_of_pairwise_genetic_differences_Mol_Biol_Evol_9_552-569., Acesso em: 29 ago. 2025.*distribution of pairwise genetic differences.* Molecular Biology and Evolution, v. 9, n. 3, p.

SALAS, A.; RICHARDS, M.; DE LA FE, T.; LAREU, M. V.; SOBRINO, B.; SÁNCHEZ-DIZ, P.; MACAULAY, V.; CARRACEDO, Á.; LALUEZA-FOX, C. *The making of the African mtDNA landscape*. American Journal of Human Genetics, v. 70, n. 5, p. 1155–1166, nov. 2002. DOI: https://doi.org/10.1086/344348.

SALAS, A.; RICHARDS, M.; LAREU, M. V.; SCOZZARI, R.; COPPA, A.; TORRONI, A.; MACAULAY, V.; CARRACEDO, A. *The African diaspora: mitochondrial DNA and the Atlantic slave trade.* American Journal of Human Genetics, v. 74, n. 3, p. 454–465, mar. 2004. DOI: https://doi.org/10.1086/382194.

SANTOS, L. M. E. dos. *Ancestralidade genética em populações miscigenadas: desafios, aplicações e um olhar sobre a história da formação do Distrito Federal.* 2023. Tese (Doutorado em Ciências Genômicas e Biotecnologia) – Universidade de Brasília, Brasília, 2023.

SARIYA, S.; LEE, J. H.; MAYEUX, R.; VARDARAJAN, B. N.; REYES-DUMEYER, D.; MANLY, J. J.; BRICKMAN, A. M.; LANTIGUA, R.; MEDRANO, M.; JIMENEZ-VELAZQUEZ, I. Z.; TOSTO, G. *Rare variants imputation in admixed populations: comparison across reference panels and bioinformatics tools.* Frontiers in Genetics, v. 10, 239, abr. 2019. DOI: https://doi.org/10.3389/fgene.2019.00239.

SCHÖNHERR, S.; WEISSENSTEINER, H.; KRONENBERG, F.; FORER, L. *HaploGrep 3 - an interactive haplogroup classification and analysis platform.* Nucleic Acids Research, v. 51, n. W1, p. W263–W268, jul. 2023. DOI: https://doi.org/10.1093/nar/gkad284 **.**

SILVA, A. C. A. *Na saúde e na doença: variabilidade genética humana estimada por marcadores genéticos neutros e em genes.* 2016. 154 f. Tese (Doutorado em Biologia Animal) – Universidade de Brasília, Brasília, DF, 2016.

SILVA, M. A. C. e; FERRAZ, T.; COUTO-SILVA, C. M.; LEMES, R. B.; NUNES, K.; COMAS, D.; HÜNEMEIER, T. *Population histories and genomic diversity of South American natives.* Molecular Biology and Evolution, v. 39, n. 1, 2022. DOI: https://doi.org/10.1093/molbev/msab339.

SILVA, M.; ALSHAMALI, F.; SILVA, P.; CARRILHO, C.; MANDLATE, F.; JESUS TROVOADA, M.; CERNÝ, V.; PEREIRA, L.; SOARES, P. *60,000 years of interactions between Central and Eastern Africa documented by major African mitochondrial haplogroup L2*. Scientific Reports, v. 5, art. 12526, 2015. DOI: https://doi.org/10.1038/srep12526.

SOARES, P.; ALSHAMALI, F.; PEREIRA, J. B.; FERNANDES, V.; SILVA, N. M.; AFONSO, C.; COSTA, M. D.; MUSILOVÁ, E.; MACAULAY, V.; RICHARDS, M. B.; CERNÝ, V.; PEREIRA, L. *The expansion of mtDNA haplogroup L3 within and out of Africa*. Molecular Biology and Evolution, v. 29, n. 3, p. 915–927, mar. 2012. DOI: https://doi.org/10.1093/molbev/msr245.

SOUZA, F. G.; MATOS, G. B.; SENA SANTOS, C.; et al. *Mitochondrial ancestry from complete mitogenomes highlights a lack of characterization of indigenous haplogroups in Brazilian Amazon population*. Communications Biology, v. 8, e835, 2025. DOI: https://doi.org/10.1038/s42003-025-08126-4.

STEWART, J.; CHINNERY, P. *The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease.* Nature Reviews Genetics, v. 16, n. 9, p. 530–542, 2015. DOI: https://doi.org/10.1038/nrg3966 **.**

TAJIMA, F. *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.* Genetics, v. 123, n. 3, p. 585–595, nov. 1989. DOI: https://doi.org/10.1093/genetics/123.3.585.

TAMM, E.; KIVISILD, T.; REIDLA, M.; METSPALU, M.; SMITH, D. G.; MULLIGAN, C. J.; BRAVI, C. M.; RICKARDS, O.; MARTINEZ-LABARGA, C.; KHUSNUTDINOVA, E. K.; FEDOROVA, S. A.; GOLUBENKO, M. V.; STEPANOV, V. A.; GUBINA, M. A.; ZHADANOV, S. I.; OSSIPOVA, L. P.; DAMBA, L.; VOEVODA, M. I.; DIPIERRI, J. E.; VILLEMS, R.; MALHI, R. S. *Beringian standstill and spread of Native American founders*. PLOS ONE, v. 2, n. 9, e829, 2007. DOI: https://doi.org/10.1371/journal.pone.0000829.

TANAKA, M.; OZAWA, T. *Strand asymmetry in human mitochondrial DNA mutations.* Genomics, v. 22, n. 2, p. 327–335, jul. 1994. DOI: https://doi.org/10.1006/geno.1994.1391 .

THE 1000 GENOMES PROJECT CONSORTIUM. *A global reference for human genetic variation*. Nature, v. 526, n. 7571, p. 68–74, out. 2015. DOI: https://doi.org/10.1038/nature15393.

TORRONI, A.; ACHILLI, A.; MACAULAY, V.; RICHARDS, M.; BANDELT, H. J. *Harvesting the fruit of the human mtDNA tree.* Trends in Genetics, v. 22, n. 6, p. 339–345, jun. 2006. DOI: https://doi.org/10.1016/j.tig.2006.04.001 .

TRECCANI, M.; LOCATELLI, E.; PATUZZO, C.; MALERBA, G. *A broad overview of genotype imputation: standard guidelines, approaches, and future investigations in genomic association studies.* BIOCELL, v. 47, n. 6, p. 1225–1241, 2023. DOI: https://doi.org/10.32604/biocell.2023.027884.

VAN OVEN, M. *PhyloTree Build 17: Growing the human mitochondrial DNA tree*. Forensic Science International: Genetics Supplement Series, v. 5, p. e392–e394, 2015. DOI: https://doi.org/10.1016/j.fsigss.2015.09.155.

VAN OVEN, M.; KAYSER, M. *Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation.* Human Mutation, v. 30, n. 2, p. E386–E394, fev. 2009. DOI: https://doi.org/10.1002/humu.20921.

VIGILANT, L.; STONEKING, M.; HARPENDING, H.; HAWKES, K.; WILSON, A. C. *African populations and the evolution of human mitochondrial DNA.* Science, v. 253, n. 5027, p. 1503–1507, set. 1991. DOI: https://doi.org/10.1126/science.1840702.

WALLACE, D. C. *A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine*. Annual Review of Genetics, v. 39, p. 359–407, 2005. DOI: https://doi.org/10.1146/annurev.genet.39.110304.095751.

WALLACE, D. C.; BROWN, M. D.; LOTT, M. T. *Mitochondrial DNA variation in human evolution and disease*. Gene, v. 238, p. 211–230, 1999. DOI: https://doi.org/10.1016/S0378-1119(99)00295-4.

WATSON, E.; FORSTER, P.; RICHARDS, M.; BANDELT, H. J. *Mitochondrial footprints of human expansions in Africa.* American Journal of Human Genetics, v. 61, n. 3, p. 691–704, set. 1997. DOI: https://doi.org/10.1086/515503.

WEIR, B. S.; COCKERHAM, C. C. *Estimating F-statistics for the analysis of population structure*. Evolution, v. 38, n. 6, p. 1358–1370, nov. 1984. DOI: https://doi.org/10.2307/2408641.

WEISSENSTEINER, H.; PACHER, D.; KLOSS-BRANDSTÄTTER, A.; FORER, L.; SPECHT, G.; BANDELT, H. J.; KRONENBERG, F.; SALAS, A.; SCHÖNHERR, S.

HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Research, v. 44, n. W1, p. W58–W63, jul. 2016. DOI: https://doi.org/10.1093/nar/gkw233 **.**

WRIGHT, S. *Evolution and the genetics of populations. Volume 4: variability within and among natural populations.* Chicago: University of Chicago Press, 1978. 511 p.

WRIGHT, S. *The genetical structure of populations.* Annals of Eugenics, v. 15, p. 323–354, 1951. DOI: https://doi.org/10.1111/j.1469-1809.1949.tb02451.x.

ZHANG, B.; ZHI, D.; ZHANG, K.; GAO, G.; LIMDI, N. N.; LIU, N. *Practical consideration of genotype imputation: sample size, window size, reference choice, and untyped rate*. Statistical Interface, v. 4, n. 3, p. 339–352, 2011. DOI: https://doi.org/10.4310/SII.2011.v4.n3.a8.