UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE BIOCIÊNCIAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

RAPHAEL BRUNO AMEMIYA

Análise da ancestralidade genética da população de São Paulo

São Paulo
2024

Raphael Bruno Amemiya

Análise da ancestralidade genética da população de São Paulo

Versão Corrigida

Dissertação de Mestrado apresentada ao Programa Interunidades de Pós-Graduação em Bionformática da Universidade de São Paulo para obtenção do título de Mestre em Ciências.

Área de Concentração: Bioinformática
Orientador: Prof. Dr. Sergio Russo Matioli

São Paulo

2024

Ficha Catalográfica elaborada eletronicamente pelo autor, utilizando o programa desenvolvido pela Seção Técnica de Informática do ICMC/USP e adaptado para a Divisão de Biblioteca e Documentação do Conjunto das Químicas da USP

A498a

Amemiya, Raphael Bruno
Análise da ancestralidade genética da população de São Paulo / Raphael Bruno Amemiya. - São Paulo, 2024.
83 p.

Dissertação (mestrado) - Programa Interunidades de Pós-Graduação em Bioinformática - USP, .
Orientador: Matioli, Sergio Russo

1. Ancestralidade. 2. Genética. 3. Bioinformática. 4. Aprendizado de máquina. 5. Ciência de dados. I. T. II. Matioli, Sergio Russo, orientador.

Nome: AMEMIYA, Raphael Bruno
Título: Análise da ancestralidade genética da população de São Paulo
Dissertação de Mestrado apresentada ao Programa Interunidades de Pós-Graduação em Bionformática da Universidade de São Paulo para obtenção do título de Mestre em Ciências.

Aprovado em: 11 de junho de 2024

Banca Examinadora

Profa. Dra.	Julia Maria Pavan Soler
Instituição	IME - USP
Julgamento	A
Profa. Dra.	Regina Celia Mingroni Netto
Instituição	IB - USP
Julgamento	A
Profa. Dra.	Síntia Iole Nogueira Belangero
Instituição	UNIFESP(FM)
Julgamento	A

AGRADECIMENTOS

À minha mãe, Cleusa de Sousa Amemiya, por todo apoio e incentivo aos meus estudos. Seu exemplo de perseverança e dedicação foi essencial para que eu acreditasse no meu potencial e continuasse avançando em minha jornada acadêmica.

Ao Prof. Dr. Sergio Russo Matioli, por me aceitar como aluno e pela confiança no projeto de pesquisa.

Ao André Chinchio, Ricardo di Lazzaro Filho e toda a equipe Genera com quem tive oportunidade de trabalhar e aprender sobre pesquisa e inovação.

À Universidade de São Paulo, por todo conhecimento que recebi, o qual foi fundamental para o meu crescimento professional e para o desenvolvimento deste projeto.

RESUMO

AMEMIYA, R. B. **Análise da ancestralidade genética da população de São Paulo.** 2024. Dissertação (Mestrado) – Programa Interunidades de Pós-Graduação em Bioinformática, Universidade de São Paulo, São Paulo, 2024.

Os avanços da Biotecnologia possibilitaram a análise de milhares de marcadores genéticos, fornecendo informações importantes sobre ancestralidade e saúde. No contexto da ancestralidade genética, testes de DNA são importantes para identificar as origens de uma população e proporcionar aos indivíduos um maior conhecimento sobre seus antepassados. O Brasil é reconhecido por sua vasta diversidade étnica, com contribuições europeias, africanas, indígenas, asiáticas, entre outras. Essa diversidade étnica apresenta desafios na inferência da ancestralidade. Neste contexto, a aplicação de técnicas de bioinformática é importante para a análise de dados genéticos e a elaboração de modelos para inferir a ancestralidade. Tendo isso em mente, o objetivo deste projeto foi analisar a ancestralidade global de indivíduos da população de São Paulo utilizando técnicas de aprendizado de máquina. Para esta finalidade, buscou-se bancos genéticos públicos contendo dados de diferentes populações. Os dados foram processados e estruturados para serem aplicados em modelos não supervisionados e supervisionados. Neste projeto, foram implementados modelos supervisionados em Python com base na Estimativa de Máxima Verossimilhança. Os modelos criados também foram combinados usando abordagens de previsão em conjunto, que combinam os resultados de diferentes modelos. A capacidade dos modelos em inferir a ancestralidade de 23 grupos populacionais foi avaliada com validação cruzada estratificada e amostras simuladas. A raiz do erro quadrático médio (RMSE) foi calculada entre as proporções inferidas e esperadas de ancestralidade com amostras simuladas. O modelo com menor valor médio de RMSE teve uma média de precisão e sensibilidade na validação de 96,0% e 94,3%, respectivamente. Este modelo foi usado para inferir a ancestralidade de 411 indivíduos de São Paulo. Considerando apenas os grupos continentais com maiores proporções, foi inferida, em média 77,5% de ancestralidade europeia, 10,3% africana, 7,4% nativa americana e 4,1% de leste asiática. As análises realizadas neste projeto exemplificam a eficácia da combinação de mais de um modelo na inferência de ancestralidade genética, assim como o uso de técnicas de aprendizado de máquina como ferramenta para compreender a diversidade de populações complexas, como a de São Paulo.

Palavras-chave: Ancestralidade. Aprendizado de máquina. Bioinformática. Genética.

ABSTRACT

AMEMIYA, R. B. **Genetic ancestry analysis of the population of São Paulo.** 2024. Dissertação (Mestrado) – Programa Interunidades de Pós-Graduação em Bioinformática, Universidade de São Paulo, São Paulo, 2024.

Advances in biotechnology have enabled the analysis of thousands of genetic markers, providing important information about ancestry and health. In the context of genetic ancestry, DNA tests are important for identifying the origins of a population and providing individuals with better knowledge about their ancestors. Brazil is known for its vast ethnic diversity, with contributions from European, African, Native American, Asian, and other populations. This ethnic diversity represents a challenge in inferring ancestry. In this context, the development and application of bioinformatics techniques are important for genetic data analysis and the development of models to infer ancestry. With these in mind, the goal of this project was to analyze the genetic ancestry of individuals from the São Paulo population using machine learning models. For this purpose, public genetic databases with diverse populations were selected. The data were processed and structured to be applied with unsupervised and supervised models. In this project, supervised models were implemented in Python based on Maximum Likelihood Estimation. The developed models were also combined into ensemble models. The ability of the model to infer the ancestry of 23 population groups was evaluated using stratified cross validation, and simulated samples. The root mean squared error (RMSE) was calculated between the predicted and expected proportions of ancestry using the simulated samples. The model with the lowest value of mean RMSE had an average precision, and sensitivity of 96.0% and 94.3%, respectively. This model was used to infer ancestry of 411 individuals from São Paulo. Considering only continental groups with the highest proportions, it was inferred an average of 77.5% for European ancestry, 10.3% for African ancestry, 7.4% for Native American ancestry and 4.1% for East Asian ancestry. The analyses conducted in this project exemplify the effectiveness of the combination of models in inferring genetic ancestry, as well as the use of machine learning models as a tool to better comprehend the diversity of complex population, such as the population of São Paulo.

Keywords: Ancestry. Machine learning. Bioinformatics. Genetics.

LISTA DE SIGLAS

DTC	direct-to-consumer
STRs	Short Tandem Repeats
SNPs	Single Nucleotide Polymorphism
AIMs	Ancestry Informative Markers
HGDP	Human Genome Diversity Project
1kGP	1000 Genomes Project
AISNPs	Ancestry Informative SNPs
PCA	Principal Component Analysis
MDS	Multidimensional Scaling
HMM	Hidden Markov Model
GWAS	Genome Wide Association Study
MLE	Maximum Likelihood Estimation
TCLE	Termo de Consentimento Livre e Esclarecido
gnomAD	Genome Aggregation Database
KRGP	Korean Reference Genome Database
VCF	Variant Call Format
USP	Universidade de São Paulo
GSA	Global Screening Arrays
MAF	Minor Allele Frequency
HWE	Hardy-Weinberg Equilibrium
PC	Principal Component
LCN	Local Classifier per Node
LCPN	Local Classifier per Parent Node
LCL	Local Classifier per Level
RMSE	Root Mean Squared Error
SABE	Saúde, Bem-estar e Envelhecimento

SUMÁRIO

1 INTRODUÇÃO

O Brasil é um dos países mais heterogêneos do mundo, tanto do ponto de vista sociocultural quanto do ponto de vista genético. A história deste país é marcada por migrações continentais que influenciaram o processo de miscigenação da população brasileira. Europeus, africanos e ameríndios contribuíram significativamente para a complexa composição étnica da população (MYCHALECHKY *et al.*, 2017; SOUZA *et al.*, 2019). Quando os portugueses desembarcaram na Bahia em 1500, estima-se que havia cerca de três milhões de ameríndios no Brasil. Este número diminui drasticamente devido à escravidão, trabalho forçado, conflitos com colonizadores e epidemias de doenças europeias (MYCHALECHKY *et al.*, 2017). A partir da segunda metade do século XVI, os africanos foram trazidos como escravos para o Brasil para trabalharem no cultivo de cana-de-açúcar e, mais tarde, nas minas de ouro e plantio de café. Durante o período do tráfico negreiro, estima-se que aproximadamente 4 milhões de africanos tenham chegado ao Brasil, principalmente da Guiné, Congo, Angola, Moçambique e Nigéria. Durante a colonização, mais de 500.000 portugueses vieram ao país. Após a abertura dos portos, italianos, espanhóis e alemães também chegaram ao Brasil. As migrações da Ásia e do Oriente Médio começaram somente no século XX, principalmente do Japão, mas também da Síria e do Líbano. As variações no processo de colonização e ocupação do território brasileiro resultaram em uma diversa e extensa escala de miscigenação genética no país (PENA; SANTOS; TARAZONA-SANTOS, 2020; PEREIRA *et al.*, 2019; SOUZA *et al.*, 2019).

Por meio de testes genéticos é possível inferir a ancestralidade de um indivíduo. Pesquisadores têm utilizado esse tipo de teste por diversas razões, como genealogia, antropologia e epidemiologia (ROYAL *et al.*, 2010). Muitos desses estudos tiveram interesse em entender o padrão de ancestralidade de uma determinada população. Exemplos deste tipo de estudo no Brasil são os trabalhos dos pesquisadores Andrade *et al.* (2018) e Giolo *et al.* (2012). No exterior, podemos citar como exemplo o trabalho dos pesquisadores Hellenthal *et al.* (2014). Os testes genéticos de ancestralidade têm tido um rápido aumento de popularidade. Estimativas recentes indicam que mais de 26 milhões de pessoas no mundo já realizaram um teste de ancestralidade por meio de empresas de testes oferecidos direto ao consumidor (DTC, do inglês direct-to-consumer), sem a intermediação de

um profissional da saúde (JORDE; BAMSHAD, 2020). Fora do Brasil, as empresas mais conhecidas desse ramo são a 23andMe, Ancestry e MyHeritage; no Brasil, podemos citar a meuDNA e a Genera. Esses testes permitem obter informações sobre as raízes ancestrais de uma pessoa. Além disso, por meio da agregação de vários perfis genéticos em um banco de dados, também é possível encontrar parentes próximos e distantes. Essa informação é extremamente útil quando uma pessoa não tem conhecimento de sua genealogia, como é o caso de pessoas adotadas (JORDE; BAMSHAD, 2020; PHILLIPS, 2016; ROYAL *et al.*, 2010).

Para entender a miscigenação brasileira, vários estudos já foram conduzidos por meio de diferentes painéis de marcadores moleculares, como repetições curtas em tandem (STRs, do inglês Short Tandem Repeats), inserções e deleções (INDELs), e polimorfismos de nucleotídeo único (SNPs, do inglês Single Nucleotide Polymorphism). Marcadores moleculares presentes no cromossomo mitocondrial podem ser utilizados para estimar a linhagem materna de um indivíduo, uma vez que o DNA mitocondrial é herdado por herança materna. Quanto à linhagem paterna, são utilizados marcadores presentes no cromossomo Y, pois este é transmitido de pai para filho (JORDE; BAMSHAD, 2020; ROYAL *et al.*, 2010; SOUZA *et al.*, 2019).

O desenvolvimento de tecnologias de genotipagem em alta escala permitiu estudos inferirem a ancestralidade por meio de milhares de SNPs. Em geral, as empresas DTC utilizam a tecnologia de arranjo de SNPs para analisar milhares desses marcadores e determinar a ancestralidade de um indivíduo (JORDE; BAMSHAD, 2020). Uma alternativa para a análise de milhares de SNPs, tem sido a análise de marcadores informativos para ancestralidade (AIMs, do inglês Ancestry Informative Markers), que são marcadores genéticos que possuem grande variação de frequência entre diferentes populações. Muitos SNPs apresentam essa característica, tornando-os ferramentas poderosas para inferir a composição genética de populações miscigenadas. Estudos têm demonstrado a eficácia de um número reduzido desses marcadores na determinação da ancestralidade africana, asiática, europeia e nativo americana de indivíduos em populações miscigenadas (ANDRADE *et al.*, 2018; BARBOSA *et al.*, 2017).

Para inferir a ancestralidade é necessário comparar o perfil genético de um indivíduo com populações de diversas regiões do mundo. Há vários bancos de dados públicos como Human Genome Diversity Project (HGDP) e o 1000 Genomes Project (1kGP) que disponibilizam perfis genéticos de diversas populações do

mundo que podem ser usadas como populações de referência (BERGSTRÖM *et al.*, 2020; SUDMANT *et al.*, 2015). O 1kGP, por exemplo, analisou variantes genéticas de 26 populações que podem ser agrupadas em 5 grandes grupos (África, Americanos miscigenados/Latinos, Europa, Leste Asiático e Sul da Ásia) (SUDMANT *et al.*, 2015). Os dados genéticos obtidos nesse projeto permitem, entre outros tipos de teste, inferir a ancestralidade de qualquer indivíduo, seja pelo uso de milhares de SNPs ou pelo uso de poucos SNPs informativos para ancestralidade (AISNPs, do inglês Ancestry Informative SNPs) (HARISMENDY *et al.*, 2019; ROYAL *et al.*, 2010). Os testes de ancestralidades podem ser realizados em diferentes níveis de resolução, inferindo a ancestralidade de um indivíduo por continentes ou macrorregiões (Europa ocidental, Leste asiático etc.), ou com um detalhamento maior a nível de países. Independente da resolução, a acurácia do teste depende da seleção da população de referência e dos SNPs analisados (JORDE; BAMSHAD, 2020).

Atualmente há duas abordagens para inferir a ancestralidade: ancestralidade global e ancestralidade local.

A ancestralidade global infere a proporção de componentes genéticos derivados de diferentes populações com base na análise de milhares de marcadores espalhados por todo o genoma. Esta abordagem tem sido dividida em duas categorias: “livre de modelos” (model-free) e “baseado em modelos” (model-based) (WOLLSTEIN; LAO, 2015).

Na categoria “baseado em modelos”, infere-se a proporção de ancestralidade de um indivíduo com base em um modelo estatístico. Nesse método, a proporção de ancestralidade é inferida a partir dos genótipos observados nos marcadores analisados e a frequência alélica destes marcadores em diferentes populações de referência, assumindo algumas premissas, como equilíbrio de Hardy-Weinberg e equilíbrio de ligação entre os marcadores (PADHUKASAHASRAM, 2014; YUAN *et al.*, 2017). Exemplos de métodos que usam essa abordagem são os programas Structure e Admixture (WOLLSTEIN; LAO, 2015).

A abordagem “livre de modelos” utiliza técnicas multivariadas, como a Análise de Componentes Principais (PCA, do inglês Principal Component Analysis) ou Escalonamento Multidimensional (MDS, do inglês Multidimensional Scaling). Basicamente, ambas as técnicas são métodos não supervisionados que permitem a redução dimensional dos dados, transformando as variáveis originais. Neste caso,

os marcadores genéticos, transformados em variáveis ortogonais, capturam a maior parte da variabilidade das variáveis originais (WOLLSTEIN; LAO, 2015). Quando aplicadas a dados genéticos com diferenças de ancestralidade entre os indivíduos, os eixos refletem a variação genética entre as populações. Amostras com ancestralidade similar tendem a possuir valores de componentes principais similares (BRISBIN *et al.*, 2012; CABREROS; STOREY, 2019; PADHUKASAHASRAM, 2014).

Uma desvantagem da ancestralidade global é o fato de considerar os marcadores genéticos isoladamente, e não a combinação de marcadores adjacentes. Marcadores genéticos no mesmo cromossomo tendem a ser herdados juntos na ausência de recombinação (PADHUKASAHASRAM, 2014). Devido aos eventos de recombinação, os cromossomos de um indivíduo miscigenado são um mosaico de segmentos originados de diferentes populações ancestrais. O objetivo da Ancestralidade Local é identificar a ancestralidade de cada um desses segmentos (PADHUKASAHASRAM, 2014; PASANIUC *et al,* 2013; SCHAEFER; SHAPIRO; GREEN, 2017; WOLLSTEIN; LAO, 2015; YUAN *et al.*, 2017). Há diversos programas disponíveis atualmente com essa finalidade. Os mais comuns utilizam o Modelo Oculto de Markov (HMM, do inglês Hidden Markov Model), Floresta Aleatória, Máquina de Vetores de Suporte, e XGBoost (DURAND *et al.*, 2014; PADHUKASAHASRAM, 2014; SHRINER, 2013; THORNTON; BERMEJO, 2014; YUAN *et al.*, 2017). Métodos de análise de ancestralidade local têm sido limitados em conseguir distinguir ancestralidade em poucas populações por vez e geralmente tomam agrupamentos no nível continental (DURAND *et al.*, 2014).

Em geral, os métodos mencionados podem ter o resultado prejudicado por diversos fatores. No que se refere ao conjunto de dados, a pouca representatividade das populações de referência, o desbalanceamento no número de amostras em cada população, a presença de indivíduos com parentesco muito próximo, a ocorrência de dados atípicos e número de marcadores genéticos são fatores que podem impactar negativamente o desempenho dos modelos, destacando a importância de cuidadosa atenção na preparação dos dados (SHRINGARPURE; XING, 2014; WOLLSTEIN; LAO, 2015).

Na área de aprendizado de máquina, tem-se sugerido atualmente o uso de técnicas de aprendizado em conjunto que combinam diferentes modelos base para obter um modelo final aprimorado, com melhor desempenho e generalização do que os modelos individuais. Entre os métodos de aprendizado em conjunto, os métodos

de “voting” (votação), “boosting” (incremento, ampliação), “bagging” (agregação ou empacotamento), e “stacking” (empilhamento) são os mais comuns, sendo considerados como o estado da arte. Atualmente são aplicados em diferentes áreas da saúde, análise de sentimentos, detecção de fraudes entre outros (MIENYE; SUN, 2022; SAGI; ROKACH, 2018). Outra abordagem pouco utilizada, é a classificação hierárquica, em que níveis hierárquicos são levados em consideração na definição das categorias a serem previstas (SILLA; FREITAS 2011). Por exemplo, no contexto de ancestralidade genética, pode-se desenvolver um algoritmo para treinar diferentes modelos considerando grupos continentais, subcontinentais, países ou grupos populacionais mais específicos, e assim por diante.

1.1 Justificativa

Os testes genéticos de ancestralidade, além de serem úteis para uma determinada pessoa conhecer mais sobre si mesma e suas origens (JORDE; BAMSHAD, 2020), também têm se demonstrado úteis na área da saúde. Estudos têm verificado que certos fatores genéticos associados a doenças possuem variações que tendem a ser mais prevalentes em determinados grupos étnicos. Variantes genéticas associadas à hipertensão, diabetes tipo 2, insuficiência renal, câncer de próstata e respostas a alguns tratamentos, por exemplo, possuem diferenças significativas em frequência entre etnias (ROTIMI; JORDE, 2010; ROYAL *et al.*, 2010; SHRINER, 2013). Além disso, por causa da acurácia ser maior em comparação com a autodeclaração de ancestralidade, os testes genéticos de ancestralidade podem aumentar o poder estatístico de estudos de associação de amplitude genômica (GWAS, do inglês “Genome Wide Association Study”) para detectar variantes genéticas de maior risco a determinadas doenças. Assim, se testes de escore de risco poligênico forem estabelecidos clinicamente, saber a ancestralidade é importante para uma interpretação mais acurada de risco.

Outro ponto a ser mencionado é que o fato de muitos estudos genéticos terem sido realizados em populações europeias, a patogenicidade de certas variantes é mais difícil de ser inferida em pessoas de origem não europeia. Sendo assim, a informação de ancestralidade pode ajudar a evitar falsas interpretações de resultados genéticos (GANNETT, 2014; JORDE; BAMSHAD, 2020; PEREIRA *et al.*, 2019). Dada a variabilidade genética da população brasileira e a importância de se

inferir a ancestralidade, faz-se necessária a utilização de ferramentas que possibilitem inferir ancestralidade com precisão e sensibilidade, inclusive em populações miscigenadas. Nesse sentido, foram empregadas aplicações de bioinformática para atender o objetivo deste projeto.

1.2 Objetivo

Analisar a ancestralidade global de indivíduos da população de São Paulo usando Estimativa de Máxima Verossimilhança (MLE, do inglês “Maximum Likelihood Estimation”) e Análise de Componentes Principais (PCA, do inglês “Principal Component Analysis”).

2 MATERIAIS E MÉTODOS

Seguem abaixo as principais etapas realizadas para atender o objetivo do projeto:

a) obtenção de dados genéticos;
b) processamento de qualidade;
c) aplicação de aprendizado não supervisionado;
d) aplicação de aprendizado supervisionado.

A coleta de dados genéticos envolveu a obtenção de genótipos de milhares de marcadores genéticos em bancos de dados públicos, como o 1kGP e o HGDP, para compor as populações de referência no cálculo de ancestralidade. Também foram obtidos dados genéticos de indivíduos do estado de São Paulo, fornecidos pela empresa Genera, com aprovação do Conselho de Ética e Termo de Consentimento Livre e Esclarecido (TCLE). Esses dados foram usados na análise da ancestralidade global dos indivíduos de São Paulo. Foi feito o processamento de qualidade e a estruturação dos dados para análises, incluindo PCA, análise não supervisionada de agrupamento e análise supervisionada de ancestralidade global. Na análise supervisionada, além do uso de pacotes conhecidos do Python, também foram desenvolvidos modelos supervisionados com base na literatura com intuito de inferir a ancestralidade de um indivíduo, obtendo resultados satisfatórios, inclusive em casos de miscigenação. Além de PCA e modelo com base em MLE, também foram exploradas outras técnicas de aprendizado de máquina como previsão em conjunto para inferir a ancestralidade global.

2.1 Obtenção de dados genéticos

Esta seção contém informações sobre os dados utilizados neste projeto.

2.1.1 Conjunto de referência

Para as análises de ancestralidade, foi necessário ter um conjunto de dados de populações de referência de diversas regiões do mundo, devido à diversidade

étnica da população brasileira. Este conjunto de dados foi construído utilizando dados genéticos disponíveis em base de dados públicos, como os do 1000 Genomes Project (1kGP), Human Genome Diversity Project (HGDP), Genome Aggregation Database (gnomAD) e Korean Reference Genome Database (KRGP) (BERGSTRÖM *et al.*, 2020; BYRSKA-BISHOP *et al.*, 2021; JUNG *et al.*, 2020; KARCZEWSKI *et al.*, 2020). Além dessas bases, foram incluídos dados genéticos de indivíduos de origens africana, europeia e nativa americana, provenientes dos estudos de Gnecchi-Ruscone *et al.* (2019), Raveane *et al.* (2019), Schlebusch *et al.* (2012) e Urnikyte *et al.* (2019), disponibilizados pelos autores para pesquisa.

Os dados do 1kGP incluem o sequenciamento genômico de cerca de 3.200 indivíduos provenientes de 26 populações diferentes, enquanto os dados do HGDP possuem o sequenciamento de cerca de 900 indivíduos distribuídos em 54 populações, conforme as Tabelas 1 e 2. Ambas as bases disponibilizam dados genéticos de milhares de SNPs em arquivos no formato VCF, em inglês Variant Call Format. Para conduzir este projeto, os dados dessas bases foram obtidos a partir do repositório do gnomAD (KOENIG *et al.*, 2023).

Tabela 1 – Populações do 1000 Genomes Project

(continua)

Grupos	Populações	Nº de indivíduos
Africana	Afro-caribenhos em Barbados	116
	Ancestralidade africana no sudoeste dos Estados Unidos	74
	Esan na Nigéria	149
	Gambianos na divisão ocidental – Mandinka	178
	Iorubá em Ibadan, Nigéria	178
	Luhya em Webuye, Quênia	99
	Mendê em Serra Leoa	99
Americana	Ancestralidade Mexicana em Los Angeles, California	97
	Colombianos em Medellin, Colômbia	132
	Peruanos em Lima, Peru	122
	Porto-riquenhos	139
Leste Asiática	Chineses Dai em Xishuangbanna, China	93
	Chineses Han do Sul	163
	Chineses Han em Beijing, China	103
	Japoneses em Tóquio, Japão	104
	Kinh na cidade Ho Chi Minh, Vietnã	122
Europeia	Ancestralidade do norte e oeste da Europa em Utah	179
	Britânicos na Inglaterra e Escócia	91
	Filândeses	99

Tabela 1 – Populações do 1000 Genomes Project

(conclusão)

Grupos	Populações	Nº de indivíduos
	Ibérios na Espanha	157
	Toscanos na Itália	107
Sul Asiática	Bengali em Bangladesh	131
	Indianos Gujarati em Houston, Texas	103
	Indianos Telugu no Reino Unido	107
	Punjabi em Lahore, Paquistão	146
	Tâmil do Sri Lanka no Reino Unido	114
Total		3202

Fonte: Byrska-Bishop *et al.* (2021)

Tabela 2 – Populações do Human Genome Diversity Project

(continua)

Grupos	Populações	Nº de indivíduos
Africano	Bantu, África do Sul	8
	Bantu, Quênia	11
	Biaka	22
	Iorubá	22
	Mandenka	22
	Mbuti	13
	San	6
Americano	Colombiano	7
	Karitiana	12
	Maia	21
	Pima	13
	Surui	8
Centro Sul Asiático	Balochi	24
	Brahui	25
	Burusho	24
	Hazara	19
	Kalash	22
	Makrani	25
	Pathan	24
	Sindi	24
	Uigures	10
Europeu	Adigue	16
	Basco	23
	Francês	28
	Italiano, Bérgamo	12
	Italiano, Tuscano	8
	Orcadenses	15
	Russo	25
	Sardenha	28
Leste Asiático	Cambojano	9
	Dai	9
	Daur	9
	Han	33

Tabela 2. Populações do Human Genome Diversity Project

(conclusão)

Grupos	Populações	Nº de indivíduos
	Han, Norte da China	10
	Hezhen	9
	Iacuto	25
	Japonês	27
	Lahu	8
	Miao	10
	Mongol	9
	Naxi	8
	Oroqen	9
	She	10
	Tujia	9
	Xibo	9
	Yi	10
	Tu	10
Oceania	Bougainville	11
	PapuanHighlands	9
	PapuanSepik	8
Oriente Médio	Beduínos	46
	Druso	42
	Mozabita	27
	Palestino	46
Total		929

Fonte: Bergström *et al.* (2020)

A versão 3 da base gnomAD também contém dados agregados de 76.156 genomas provenientes de indivíduos de diferentes continentes. Já a base KRGP contém dados agregados do genoma de 1.722 coreanos. Para este projeto, as frequências alélicas dos SNPs de judeus asquenazes no gnomAD e de coreanos no KRGP foram usadas para simular perfis genéticos, os quais foram incluídos ao conjunto de referência para análise de ancestralidade. A simulação foi feita usando o programa em Python, simuPOP (PENG; KIMMEL, 2005). Quanto aos dados dos artigos mencionados, a Tabela 3 contém a relação das populações. Em cada estudo, os indivíduos foram genotipados pela técnica de arranjo de SNPs e os dados foram disponibilizados em arquivos binários do programa PLINK (CHANG *et al.*, 2015; PURCELL; CHANG, 2024). A incorporação dos dados desses estudos no conjunto de referência teve o propósito de complementar a representatividade populacional, considerando diferentes grupos étnicos.

Tabela 3 – Populações disponibilizadas em artigos

Grupos	Populações	Nº de indivíduos
América - Gnecchi-Ruscone *et al.* (2019)	Ashaninka	10
	Bolivia_Aymara	26
	Cashibo	10
	Huambisa	8
	Shipibo	17
	Titicaca Aymara	21
	Titicaca Quechua	22
	Titicaca Uros	9
	Tzotzil	36
	Wichi	24
	Yanesha HighSelva	23
	Yanesha IntermediateSelva	23
Europa - Raveane *et al.* (2019)	Abruzo	8
	Albânia	6
	Apúlia	14
	Basilicata	9
	Calábria	2
	Campânia	14
	Friul-Veneza Júlia	15
	Lácio	4
	Ligúria	1
	Lombardia	5
	Marcas	16
	Molise	2
	Trentino-Alto Ádige	8
	Toscana	4
	Vale de Aosta	2
	Vêneto	16
	Úmbria	9
África - Schlebusch *et al.* (2012)	ColouredColesberg	20
	ColouredWellington	20
	GuiGhanaKgal	15
	Ju/'hoansi	18
	Karretjie	20
	Khomani	39
	Khwe	17
	Nama	20
	SEBantu	20
	SWBantu	12
	Xun	19
Europa - Urnikyte *et al.* (2019)	Lituânia	424
Total		1008

Fonte: Conforme coluna à esquerda

2.1.2 Coparticipação e perfis genéticos de indivíduos de São Paulo

Este projeto teve a coparticipação da empresa Genera. Fundada em 2010, a Genera tem o seu foco na realização de testes genéticos e em projetos de inovação para tornar os testes oferecidos mais acessíveis à população, mantendo a qualidade e o diferencial da marca. Nos seus primeiros anos, a empresa focou-se nos testes genéticos de paternidade e sexagem fetal. Atualmente, é especializada em testes de ancestralidade e outros testes com aplicação em genealogia, como o teste de Busca Parentes, que permite encontrar parentes próximos ou distantes na base de dados de clientes da empresa. A coparticipação da empresa no projeto é relevante devido à sua atuação com testes de ancestralidade genética e possuir uma extensa base de dados genéticos. O setor de Pesquisa e Desenvolvimento da empresa também conta com uma equipe técnica composta por biólogos, biomédicos e bioinformatas, incluindo o pesquisador responsável deste projeto, que coordena a equipe com o cargo atual de Coordenador de Projetos.

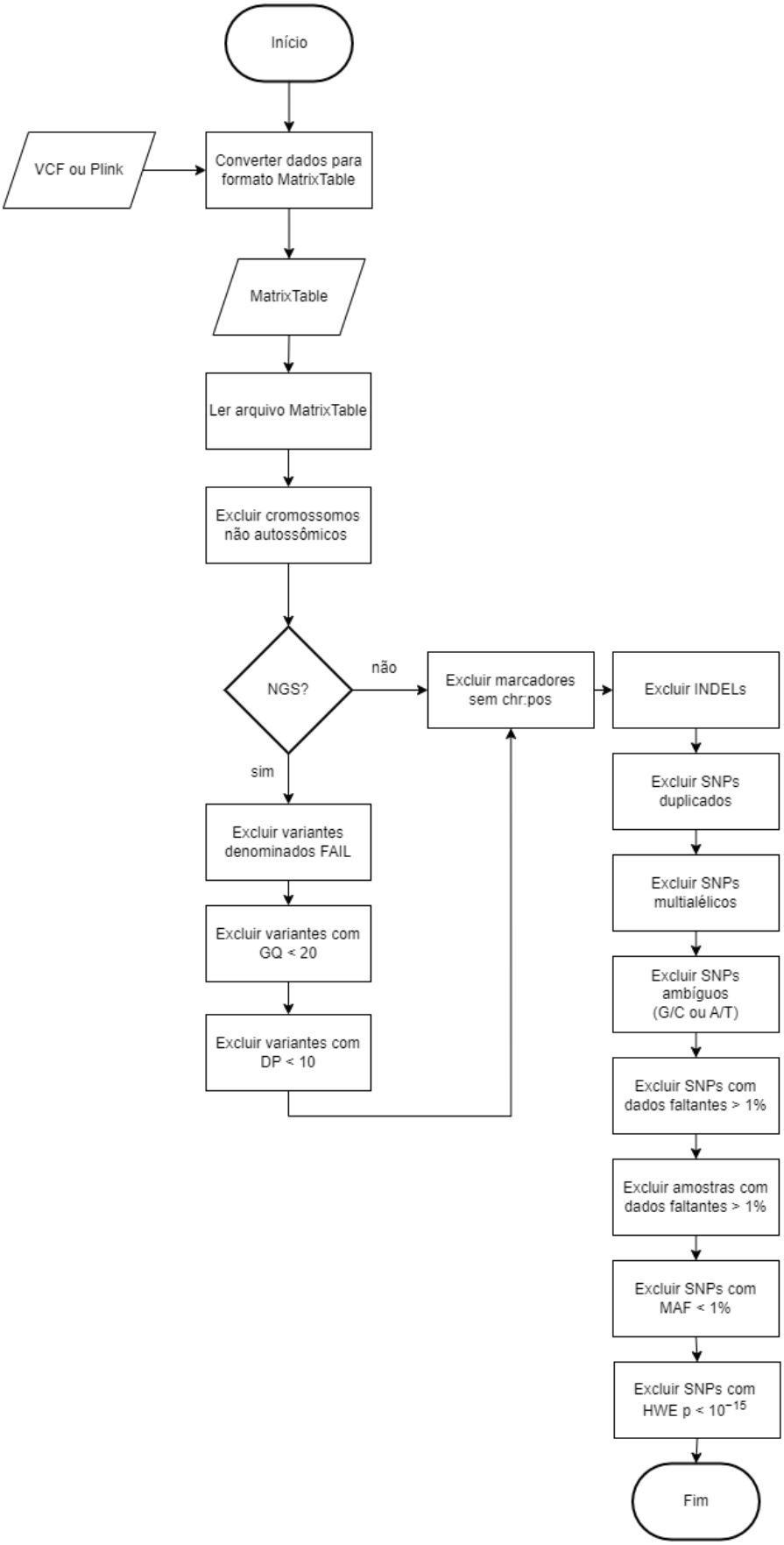
Para atender o objetivo principal deste projeto, foi necessário obter perfis genéticos de pessoas residentes no estado de São Paulo. Após a aprovação do projeto pelo Comitê de Ética em Pesquisa – Seres Humanos do Instituto de Biociências da USP, a empresa Genera concedeu acesso aos dados genéticos anonimizados de indivíduos de São Paulo, que foram contatados para serem convidados a participar da pesquisa por meio de e-mail enviado pela própria empresa, contendo o TCLE. A ideia inicial do projeto era analisar 1000 indivíduos de São Paulo, entretanto, apenas 425 clientes responderam com aceite à pesquisa.

Os dados foram entregues pela Genera em um arquivo contendo quatro colunas para o rsid, cromossomo, posição e genótipo das variantes analisadas. Esses arquivos foram posteriormente convertidos para o formato binário do programa PLINK para análise neste projeto. Dos 425 participantes, 411 foram incluídos nas análises. Os dados dos outros participantes não foram usados, pois adquiriram o “Método 1” da empresa, contendo um conjunto menor de SNPs, com baixa sobreposição em relação ao conjunto de referência populacional. Na época desta etapa do projeto, a Genera realizava a genotipagem por meio de dois métodos: Método 2, usando a técnica de arranjo de SNPs da Illumina, Global Screening Array (GSA) v.3, contendo cerca de 600 mil SNPs; e o Método 1 com um conjunto personalizado de aproximadamente 60 mil SNPs.

2.2 Processamento de qualidade

Após coleta, os dados foram preparados para as análises de ancestralidade. Os dados foram preparados seguindo as práticas comumente descritas na literatura para esse tipo de dado (BERGSTRÖM *et al.*, 2020; DI GAETANO *et al.*, 2012; GASPAR; BREEN, 2019; KOENIG *et al.*, 2023). A Figura 1 contém as etapas do processamento de qualidade. Estas etapas foram executadas com o auxílio do pacote Python, Hail 0.2 (Hail team, 2021).

Figura 1 – Etapas do processamento de qualidade



Fonte: Dados originais da pesquisa

Hail é uma ferramenta de análise de dados, de código aberto, e baseada em Python, contendo estruturas e métodos para trabalhar com dados genômicos. Após converter os dados para o formato MatrixTable, uma série de filtros foram aplicados. Este projeto focou-se apenas em marcadores autossômicos, pois a análises de marcadores nos cromossomos Y e mitocondrial requerem uma abordagem diferente que fogem do escopo deste projeto. Também foram excluídos INDELS, SNPs multialélicos, SNPs ambíguos e SNPs com mais de 1% de dados faltantes.

No caso dos conjuntos de dados de referência com mais de uma população. O filtro da frequência do alelo menos frequente (MAF, do inglês Minor Allele Frequency) excluiu SNPs com frequência relativa abaixo de 1% em todas as populações. O filtro do Equilíbrio de Hardy-Weinberg (HWE) excluiu SNPs com p valor menor que 10^{-15} em todas as populações. Estes dois filtros costumam ser usados para remover SNPs com possíveis erros de genotipagem. O desvio do HWE também pode ocorrer devido ao efeito Wahlund, que se refere à redução da heterozigosidade em uma população contendo subpopulações com frequências alélicas diferentes (ABRAMOVS; BRASS; TASSABEHJI, 2020; PEARMAN; URBAN; ALEXANDER, 2022). Como este é o caso dos conjuntos de referência com mais de uma população, a exclusão dos SNPs pelo filtro do HWE foi feita apenas quando o desvio do HWE ocorreu em todas elas. Diferentes estudos adotam valores diferentes de corte, variando entre p valor menor que 10^{-3} à menor que 10^{-7}, mas também há recomendações de apenas excluir valores extremos para evitar a exclusão de bons SNPs (CHANG *et al.*, 2015; PURCELL; CHANG, 2024; ZHAO *et al.*, 2018).

Em seguida, amostras com grau de parentesco muito alto foram removidas das análises de ancestralidade para evitar viés nos resultados. A determinação de parentesco entre as amostras foi feita utilizando o programa em R, SNPRelate (ZHENG *et al.*, 2012), usando valores de coeficiente de “kinship” acima de 0,125 como corte, exceto em nativos americanos que foram excluídos relações de primeiro grau.

Após os filtros, os dados do KRGP, dos artigos e da Genera foram convertidos da versão GRCh37 para a versão GRCh38 usando Hail 0.2.

O programa PLINK 1.9 (CHANG *et al.*, 2015; PURCELL; CHANG, 2024) também foi usado para excluir SNPs em desequilíbrio de ligação no conjunto de dados da Genera, usando a opção e os parâmetros --indep-pairwise 50 5 0,5. O desequilíbrio de ligação refere-se à correlação de alelos próximos em um mesmo

cromossomo, que segregam juntos com maior frequência do que o esperado (HOLLOWAY; PRESCOTT, 2017). O filtro para desequilíbrio de ligação foi necessário, pois os modelos de ancestralidade global usados neste projeto assumem equilíbrio de ligação, como o programa Admixture 1.3 (ALEXANDER; NOVEMBRE; LANGE, 2009). Segundo o manual deste programa, é possível que o filtro para desequilíbrio de ligação não remova o desequilíbrio totalmente, principalmente em populações de miscigenação recente. Miscigenações recentes podem introduzir covariância entre marcadores previamente não ligados nas populações de origem. Isso pode ocorrer quando haplótipos são mais comuns em uma das populações (TAN; ATKINSON, 2023).

Após processamento de qualidade, foram mantido os SNPs em comum entre os conjuntos de dados. Em seguida, os conjuntos de dados populacionais foram unidos em um único arquivo no formato binário do PLINK.

Em seguida, os dados foram carregados para análise com Python no Jupyter Notebook, juntamente com Kedro v0.18.8, uma ferramenta que incorpora boas práticas de engenharia de dados para criação de fluxo de trabalho. Os dados foram carregados no formato “DataArray” usando o pacote do Python, chamado pandas-plink v2.2.9. Neste formato, os dados ficam estruturados numa matriz em que cada linha representa um indivíduo e cada coluna contém contagem de um dos alelos de um determinado SNP. Além dessa matriz, esse “DataArray” também contém informações acerca dos dados carregados. Abaixo estão listadas as informações que foram mais relevantes para as análises deste projeto:

a) iid - código de cada indivíduo;
b) snp - código de cada SNP no formato rsid;
c) chrom - número do cromossomo de cada SNP;
d) pos - posição de cada SNP no cromossomo;
e) a1 - um dos alelos de cada SNP, cuja contagem está na matriz.

A Tabela 4 ilustra a estrutura dos dados na matriz. Supondo que os respectivos alelos (a1) dos SNP1, SNP2 e SNP3 são G, A e T. O Indivíduo 1 teria dois alelos G no SNP1, um alelo A no SNP2 e nenhum alelo T no SNP3. Essa mesma lógica seria aplicada aos demais indivíduos e SNPs analisados. Essa estrutura de dados é bastante utilizada na área da bioinformática, por exemplo, o programa SNPRelate, usado para PCA e análises de parentesco, utiliza formato de dados similar (ZHENG *et al.*, 2012). Os dados faltantes estão representados como

nan. Como após o processamento de qualidade, a proporção de dados faltantes era mínima, os dados faltantes foram imputados com o valor mais frequente usando as classes SimpleImputer e Pipeline da biblioteca do Python scikit-learn v1.3.2 (PEDREGOSA *et al.*, 2011).

Tabela 4 – Exemplo de como os dados pós-processamento estão estruturados para análise

	SNP1	SNP2	SNP3
Indivíduo1	2	1	0
Indivíduo2	1	2	nan

Fonte: Dados originais da pesquisa

Após processamento de qualidade, foram feitas as análises abaixo:

1. Análise não supervisionada com o programa Admixture 1.3;
2. PCA com a biblioteca do Python scikit-learn v1.3.2.

Essas análises foram feitas para verificar se o processamento de qualidade foi feito adequadamente, comparando os resultados obtidos com o que se costuma verificar na literatura. Também foi usado o algoritmo LocalOutlierFactor do scikit-learn v1.3.2 (PEDREGOSA *et al.*, 2011) para identificar possíveis dados atípicos. Além disso, apesar de já se saber as populações de origem das amostras dos bancos de referência, essas análises foram úteis para confirmar as possibilidades de grupos populacionais que poderiam ser usadas nos modelos supervisionados.

2.3 Aprendizado de Máquina

Aprendizado de máquina é um ramo da Inteligência Artificial com diversas aplicações em resoluções de problemas, inclusive na Bioinformática. Como definido por Arthur Samuel em 1959, aprendizado de máquina é a área de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados. Os algoritmos de aprendizado de máquina podem ser categorizados em três grupos principais: supervisionado, não supervisionado e por reforço (GÉRON, 2022; OLSON *et al.*, 2018). No aprendizado supervisionado, são disponibilizados ao algoritmo os

valores esperados de previsão para cada amostra, para que possa determinar a função que melhor se ajusta aos dados fornecidos no treinamento.

Dependendo do que pretende prever, há modelos chamados de regressão para previsões numéricas e classificação para previsões categóricas. Nos métodos não supervisionados, não são fornecidos valores esperados ao algoritmo, neste caso, as amostras são agrupadas de acordo com padrões de similaridade encontrados. No aprendizado por reforço, o computador, chamado de “agente”, aprende a executar uma determinada ação por uma série de tentativa e erro, recebendo recompensas por cada acerto e punições por cada erro. O computador aprende ao chegar na estratégia que otimiza o ganho de recompensas (GÉRON, 2022).

2.3.1 Admixture

O programa Admixture 1.3 está dentro da categoria “baseado em modelos” de ancestralidade global e implementa duas abordagens: uma não supervisionada e outra supervisionada. Na abordagem não supervisionada, o usuário define o número de k populações, permitindo ao algoritmo calcular as frequências alélicas de cada SNP nessas k populações. Em seguida, o algoritmo infere as proporções que cada populações contribui para a composição de ancestralidade de cada amostra. Na abordagem supervisionada, o usuário fornece as populações conhecidas de cada amostra do conjunto de referência em um dos arquivos de entrada. Então, o algoritmo calcula as frequências alélicas em cada população determinada pelo usuário, que podem ser usadas para inferir as proporções de ancestralidade de cada amostra em comparação com as populações de referência (ALEXANDER; LANGE, 2009, 2011).

Neste projeto, o método não supervisionado do programa foi usado para avaliar se não houve erros na etapa de processamento de qualidade e verificar se as proporções de ancestralidade para as k populações fazem sentido com o esperado para as amostras com ancestralidade conhecida no conjunto de referência. Nessa análise, o programa foi executado várias vezes, variando os valores de k de 1 à 10. O método supervisionado também foi utilizada para comparar com o desempenho dos modelos em Python desenvolvidos neste projeto.

2.3.2 PCA

PCA é uma técnica não supervisionada usada para redução de dimensionalidade dos dados, na qual se obtêm novas variáveis ortogonais, denominadas componentes principais (PCs, do inglês Principal Components), que capturam o máximo da variabilidade das variáveis originais. PCA é tradicionalmente usada em genética de populações para redução de dimensionalidade e visualização geral da estrutura populacional. Para aplicar essa técnica com dados genéticos é necessária uma matriz G com dimensões N x D, onde N é o número de indivíduos e D é o número de marcadores genéticos, neste caso, SNPs. Esta matriz é preenchida com a contagem de um dos alelos que cada amostra possui em cada SNP. Na PCA, os autovalores e autovetores são obtidos da matriz de covariância (ou correlação) dos dados. Estes são usados para transformar os dados e obter uma nova matriz X de dimensões N x C, onde C é o número de componentes principais (DUFORET-FREBOURG *et al.*, 2016; GASPAR; BREEN, 2019; PASCHOU *et al.*, 2007; RASCHKA; LIU; MIRJALILI, 2022).

A escolha do número de componentes principais a serem usados em análises subsequentes pode ser feita de diferentes maneiras. Com base no critério a priori, quando o pesquisador já tem uma ideia de quantos PCs são necessários; com base no critério da raiz latente, no qual é escolhido um número de PCs, cuja soma de autovalores é maior que 1; e com base no critério da proporção da variabilidade explicada, no qual é selecionado o número de PCs que capturam uma proporção desejada da variabilidade original dos dados (FÁVERO; BELFIORE, 2019). Quando os componentes principais são usados como entrada de modelos não supervisionados ou supervisionados, o número de componentes pode ser definido através de uma busca em que se testa cada quantidade de componentes em uma validação cruzada e escolhe-se a quantidade de componentes com a qual o modelo obtém o melhor desempenho para uma determinada métrica. Quando o interesse é visualização, geralmente, são selecionados os dois ou os três primeiros componentes principais para a construção dos eixos do gráfico de dispersão.

Neste projeto, a PCA foi feita usando a biblioteca em Python, scikit-learn v1.3.2 (PEDREGOSA *et al.*, 2011), que usa a estratégia de Halko, Martinsson e Tropp (2009), Decomposição em Valores Singulares Aleatorizado, quando o conjunto de dados contém mais de 500 linhas e 500 colunas, e a quantidade de PCs

a serem obtidos é menor que 80% do menor valor entre o número de linhas e colunas na matriz de dados. Essa estratégia utiliza algoritmos randomizados para obter uma matriz menor que a matriz original, a qual é utilizada no cálculo de Decomposição em Valores Singulares.

A PCA foi usada para reduzir a dimensionalidade dos dados obtidos após o processamento de qualidade. Os três primeiros PCs extraídos foram usados para visualização do padrão geral dos dados. Como a origem populacional dos indivíduos no conjunto de referência era conhecida, o objetivo dessa visualização foi de verificar se indivíduos de uma mesma população ficariam mais próximos entre si do que de indivíduos de outras populações.

2.3.3 Modelo supervisionado de Ancestralidade global – MLEMix

Na análise de ancestralidade global com MLE, a proporção de ancestralidade é inferida a partir dos genótipos observados nos marcadores analisados e da frequência alélica desses marcadores em diferentes populações de referência, assumindo algumas premissas, como equilíbrio de ligação entre os marcadores (PADHUKASAHASRAM, 2014; YUAB *et al.*, 2017). Esta mesma abordagem é usada no programa Admixture (ALEXANDER; NOVEMBRE; LANGE, 2009). Baseado na explicação em Alexander, Novembre e Lange (2009) e Frudakis (2010), a inferência de ancestralidade pode ser calculada conforme função log-verossimilhança (1):

$$\mathcal{L}(Q,F) = \sum_i \sum_j \{g_{ij} \ln [\sum_k q_{ik} f_{kj}] + (2 - g_{ij}) \ln [\sum_k q_{ik} (1 - f_{kj})]\} \quad (1)$$

Onde, g_{ij} representa o número de cópias do alelo1 no SNP *j* de um indivíduo *i*, podendo assumir os valores 2, 1, ou 0, considerando genótipos AA, AB e BB, respectivamente. Quanto ao q_{ik}, refere-se à proporção contribuída pela população *k* ao genoma do indivíduo *i*. E f_{kj}, refere-se à frequência do alelo 1 no SNP *j* na população *k*. A ideia é buscar os valores de q_{ik} que maximizam a função, tendo como restrições as fórmulas (2) abaixo:

$$0 \leq f_{kj} \leq 1;\ q_{ik} \geq 0;\ \sum_k q_{ik} = 1 \quad (2)$$

Também pode ser incluído um termo de penalização conforme em Alexander e Lange (2011) que consiste na equação (4) abaixo:

$$\mathcal{G}(Q,F) = \mathcal{L}(Q,F) - \lambda\mathcal{P}(Q) \quad (3)$$

$$\mathcal{P}(Q) = \sum_{i,k} \frac{\log\left(1+{}^{q_{ik}}/_{\Upsilon}\right)}{\log\left(1+{}^{1}/_{\Upsilon}\right)} \quad (4)$$

Neste caso, ao invés de maximizar a função log-verossimilhança, a função G(Q,F) é maximizada. Esta função consiste na função (1) menos o termo de penalização λP(Q). As constantes λ e ϒ determinam a intensidade dessa penalização. Esta penalização controla as proporções de ancestralidade q_{ik}, evitando excesso de proporções residuais e de miscigenação. Isso é útil principalmente em situações de análise de populações muito similares ou quando há uma quantidade limitada de SNPs para análise.

Para chegar na função (1), tem-se como base as equações abaixo:

$$\Pr(AA\ para\ i\ no\ SNP\ j) = [\textstyle\sum_k q_{ik} f_{kj}]^2 \quad (5)$$

$$\Pr(AB\ para\ i\ no\ SNP\ j) = 2\left[\textstyle\sum_k q_{ik} f_{kj}\right][\textstyle\sum_k q_{ik}(1-f_{kj})] \quad (6)$$

$$\Pr(AA\ para\ i\ no\ SNP\ j) = [\textstyle\sum_k q_{ik}(1-f_{kj})]^2 \quad (7)$$

Onde as equações (5), (6), e (7) são adaptações da Teoria de Hardy-Weinberg, incluindo q_{ik}, pois o indivíduo pode ser miscigenado para as k populações.

Para este projeto, foi desenvolvido um script em Python seguindo as recomendações da biblioteca scikit-learn (“Developing scikit-learn estimators”) para a construção de um classificador. Dessa forma, foi criado uma classe do Python, nomeada de “MLEMix”, contendo os métodos “fit”, “partial_fit”, “predict_proba” e “predict”, possibilitando a integração do classificador à pipeline do scikit-learn.

Foi criado um script próprio ao invés de usar a opção supervisionada do programa Admixture, por dois motivos: 1) ter maior flexibilidade para o treino e avaliação do modelo, podendo aproveitar funcionalidades já prontas do scikit-learn como validação cruzada, otimização de hiperparâmetros, entre outros; 2) poder utilizar este modelo nos modelos de previsão em conjunto do scikit-learn, como o “StackingClassifier” e o “VotingClassifier”.

Em resumo, o método “fit” do modelo, criado para este projeto, recebe uma matriz X (n_samples, n_features) com a dosagem de alelos e um vetor y (n_samples) com os nomes das populações de cada amostra do conjunto de referência. O método contém o código que calcula as frequências alélicas de cada SNP para cada uma das populações fornecidas. O “método predict_proba” recebe uma matriz X de genótipos (n_samples, n_features) e contém o código que irá retornar as proporções de ancestralidade para cada uma das amostras, conforme a matriz de genótipos fornecida e as frequências calculadas no “fit”. Para maximizar a função de máxima verossimilhança foi utilizado a biblioteca scipy do Python (VIRTANEN *et al.*, 2020). Após o treinamento, o modelo pode ser salvo usando pacotes do Python, como joblib (VAROQUAUX; GRISEL, 2009), podendo ser usado novamente para fazer previsões com novos dados. A Figura 2 ilustra que a inicialização de uma instância do modelo, treinamento e previsão, podem ser feitos com algumas linhas de código.

Figura 2 – Exemplo de uso do modelo MLEMix

```
>>> from mlemix.mlemix import MLEMix
>>> model = MLEMix()
>>> model.fit(X, y)
    MLEMix()
>>> print(model.predict(X_new))
    [2]
>>> print(model.predict_proba(X_new))
    [[0.30, 0,  0.70]]
```

Fonte: Dados originais da pesquisa

Além dessa classe do Python, também foi criada uma classe chamada “RegMLEMix” que herda a classe “MLEMix” e modifica o método que contém a função log-verossimilhança para incluir o termo de penalização.

Neste projeto, este modelo foi utilizado para previsão de ancestralidade de três formas:

a) Diretamente como um modelo único;
b) Em combinação de mais de um modelo, usando o “VotingClassifier” do scikit-learn;
c) Em um algoritmo de classificação hierárquico, também criado neste projeto e explicado mais adiante.

Repositório dos modelos no Github:
https://github.com/Raphael-Amemiya/mlemix

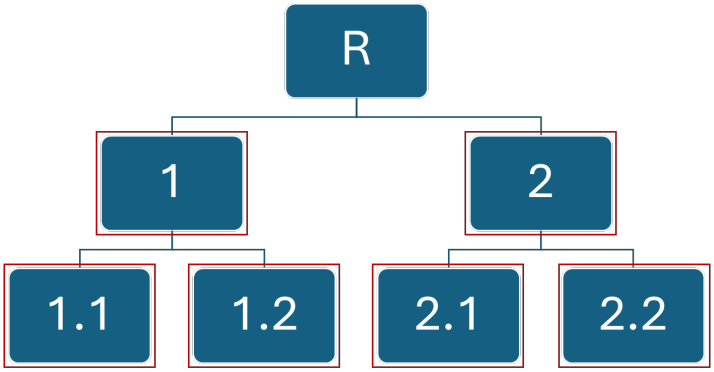
2.3.4 Classificação hierárquica

Este tipo de classificação, embora pouco mencionado, apresenta uma proposta interessante para dados cujas classes possuem certa estrutura hierárquica. No contexto da previsão de ancestralidade, é possível considerar diferentes níveis, como continentais, subcontinentais, países, sub-regiões ou subgrupos específicos. Segundo Silla e Freitas (2011), pode-se considerar as abordagens abaixo:

a) Plano: abordagem tradicional, na qual se ignora qualquer hierarquia, tendo apenas um único modelo com todas as classes terminais da árvore;
b) Local: quando a hierarquia é considerada e há classificadores específicos para cada situação. Esta abordagem também é chamada de “top-down”, pois a cada nova amostra teste, primeiro é feita a previsão das classes mais gerais do primeiro nível, então, usa a classe prevista para restringir as classes a serem previstas nos níveis subsequentes. O problema que pode surgir dessa abordagem é a propagação de erros de um nível para os seguintes. Entretanto, pode-se usar limites de confiança do resultado para evitar que o modelo continue propagando o erro.

O Classificador Local por Nó (LCN, do inglês Local Classifier per Node) consiste no treinamento de um classificador binário para cada nó, exceto o nó raiz (SILLA; FREITAS, 2011), como ilustrado na Figura 3:

Figura 3 – Classificador Local por Nó: cada retângulo representa uma classe e as linhas vermelhas representam os classificadores binários

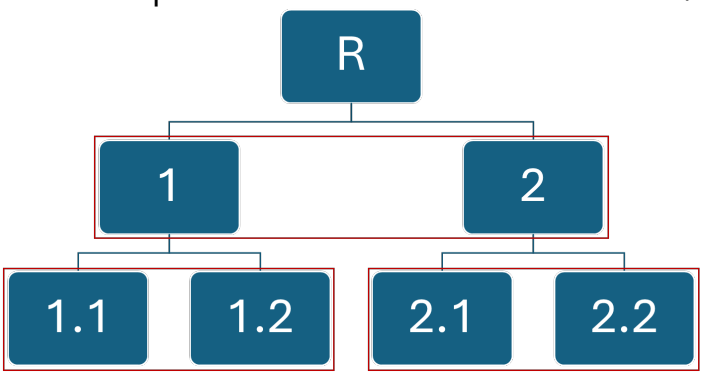


Fonte: Dados originais da pesquisa

Para evitar inconsistências no LCN, quando um classificador prevê “Falso” para uma classe, os classificadores dos nós filhos são ignorados.

O Classificador Local Por Nó Parental (LCPN, do inglês Local Classifier per Parent Node) consiste no treinamento de um classificador multiclasse para cada nó parental para prever as classes dos nós filhos (SILLA; FREITAS, 2011), como ilustrado na Figura 4:

Figura 4 – Classificador Local Por Nó Parental: cada retângulo representa uma classe, e as linhas vermelhas representam os classificadores binários/multiclasse

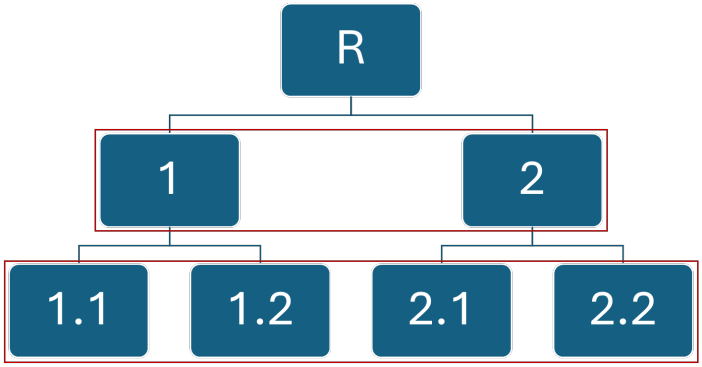


Fonte: Dados originais da pesquisa

Para evitar inconsistências no LCPN, sugere-se controles que permitam que os classificadores continuem nos nós filhos apenas se as classes dos nós parentais correspondentes forem previstas. Além disso, Silla e Freitas (2011) sugerem uma “seleção de classificadores”, onde para cada nó seria selecionado um tipo de classificador que melhor se ajustar aos dados.

O Classificador Local por Nível (LCL, do inglês Local Classifier per Level) treina um classificador por nível, conforme a Figura 5:

Figura 5 – Classificador Local por Nível: cada retângulo representa uma classe, e as linhas vermelhas os classificadores binários/multiclasses



Fonte: Dados originais da pesquisa

Como cada classificador é independente do outro, o LCL também está sujeito a inconsistências, mas isso pode ser amenizado com restrições das previsões dos nós filhos apenas se houver previsão de classe correspondente no nó parental.

Neste projeto, uma classe do Python foi escrita para implementar esta abordagem de previsão. O algoritmo cria um grafo dirigido para organizar o treinamento hierárquico. Para facilitar essa funcionalidade, foi usado a biblioteca do Python, NetworkX v3.1 (HAGBERG; SWART; CHULT, 2008).

Esta classe do Python também foi criada para ser integrável com o “scikit-learn”, então, contém os métodos “fit”, “predict_proba”, “predict” e outros métodos internos. Além disso, ao inicializar o modelo, o usuário deve fornecer uma lista contendo os modelos que serão usados em cada nível. O método “fit” consiste no treinamento e recebe uma matriz de genótipos X (n_samples, n_features) e uma matriz y (n_samples) que contém os grupos populacionais a serem previstos em cada nível. O método “predict_proba” retorna as probabilidades previstas pelo classificador. Para evitar inconsistências, apenas os grupos populacionais que obtiverem um valor de probabilidade acima de 1% têm os grupos populacionais subsequentes incluídos no modelo do nível seguinte.

Como as populações de origem das amostras dos artigos e bancos de dados consultados são conhecidas, foi possível usar essa informação para construir os grupos populacionais que o modelo usou nos treinamentos deste projeto.

Repositório dos modelos no Github:
https://github.com/Raphael-Amemiya/mlemix

2.3.5 VotingClassifier

VotingClassifier é uma técnica de previsão por conjuntos implementada no “scikit-learn”, que tem como estratégia a combinação de diferentes classificadores para fazer previsões. A previsão para uma amostra pode ser feita por meio da classe com maior número de previsões entre os classificadores, ou por meio da média (ponderada ou não) das probabilidades previstas pelos classificadores (PEDREGOSA *et al.*, 2011).

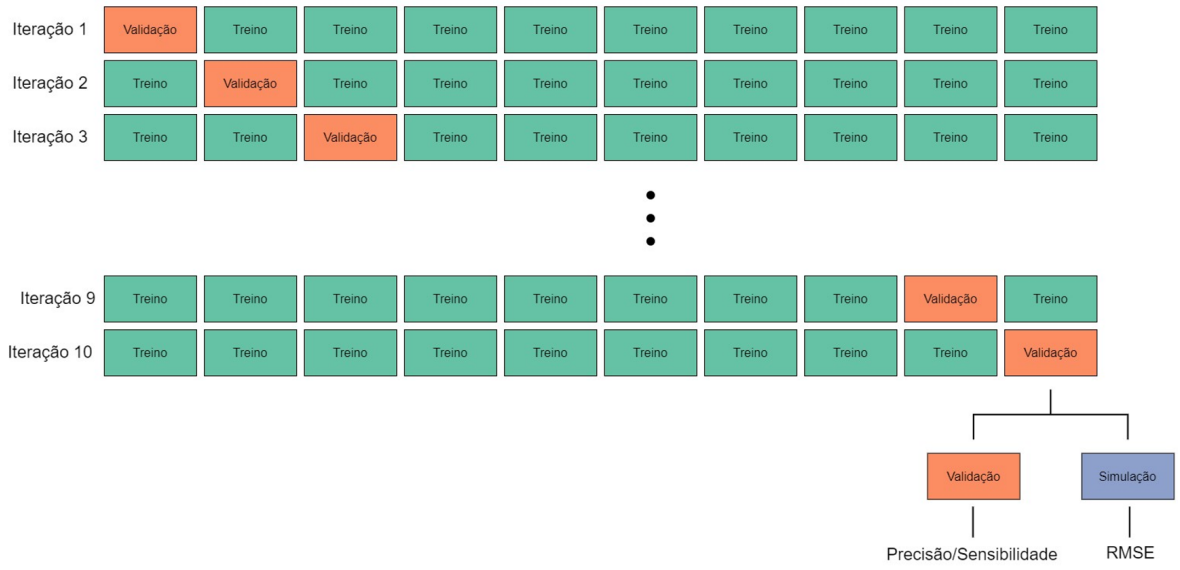
Neste projeto, esta técnica foi usada para combinar os resultados previstos pelos modelos supervisionados.

2.3.6 Validação dos modelos

Neste projeto, foram criados diferentes modelos baseados em MLE (RegMLEMix) com valores diferentes do parâmetro λ (0, 25, 50, 100, 200 e 300). Também foram criados modelos de previsão em conjunto combinando RegMLEMix (λ=0) e outros modelos com valores diferentes de λ. Versões semelhantes destes modelos também foram criados utilizando a abordagem hierárquica, HierarchicalLocalClassifier.

Para validar os modelos supervisionados, foi utilizado validação cruzada estratificado com 10 partições. Nesta validação, o conjunto de dados é dividido em 10 partes mantendo a proporção de indivíduos de cada grupo em cada partição. Uma dessas partes é reservada para validação, enquanto as outras partes são utilizadas para treino. O procedimento se repete utilizando outra parte para validação e as demais para treino, até que todas as partes tenham sido usadas como validação. Neste projeto, as amostras de validação em cada iteração também foram usadas para simular amostras miscigenadas com a biblioteca Hail 0.2 (Hail Team). Em cada iteração, foram simuladas 700 amostras. A Figura 6 ilustra o esquema da validação:

Figura 6 – Validação cruzada com 10 partições



Fonte: Dados originais da pesquisa

A Raiz do Erro Quadrático Médio (RMSE, do inglês “Root Mean Squared Error”) foi usada para avaliar os modelos. Os valores de RMSE foram calculados

entre as proporções de ancestralidade previstas e esperadas das amostras simuladas. A Fórmula 8 usada de RMSE foi baseada no trabalho de Bansal e Libiger (2015) e Wathen *et al.,* (2019) que também avaliaram o desempenho de suas metodologias com amostras simuladas. Após a validação cruzada, o modelo com menor valor médio de RMSE foi selecionado para calcular a ancestralidade das amostras de São Paulo. Métricas de precisão e sensibilidade também foram calculadas em cada iteração, usando os indivíduos do conjunto de referência das partições de validação. Ambas as métricas foram calculadas usando a biblioteca scikit-learn.

$$RMSE(\hat{a},\ a) = \sqrt{\frac{1}{nK}\sum_i \sum_k (\hat{a}_{ik} - a_{ik})^2} \quad (8)$$

Sendo *n* o número de indivíduos e K o número de grupos populacionais. Para cada indivíduo *i*, a_{ik} representa a proporção de ancestralidade esperada, enquanto $\hat{a}_{ik}$ representa a proporção de ancestralidade inferida.

2.3.7 Inferência da ancestralidade da população de São Paulo

A ancestralidade global dos Indivíduos de São Paulo foi calculada com o modelo supervisionado com menor valor de RMSE na validação cruzada. PCA também foi usado para visualizar a posição dos indivíduos em comparação com as populações recentes no conjunto de referência.

3. RESULTADOS

Como a intenção era analisar a ancestralidade genética da população de São Paulo com os dados da Genera, foram usados apenas os SNPs em comum entre os dados da Genera e dos dados públicos das populações de referência. Após o procedimento de qualidade, ficaram um número final de 52.609 SNPs nos 22 cromossomos autossômicos, com uma densidade aproximada de 1 a 2 a cada 100 mil pares de bases.

Os conjuntos de dados do 1kGP e HGDP, disponíveis através da base do gnomAD, continham inicialmente um total de 4.091 indivíduos, enquanto os artigos continham um total de 1.008 indivíduos. Após o processamento de qualidade, incluindo a exclusão de indivíduos com relações de parentesco elevadas, ficaram 4.316 indivíduos. Populações miscigenadas, como Colômbia e Porto Rico, não foram selecionadas para compor o conjunto de referência, permanecendo 3.416 indivíduos. Essas populações são conhecidas por apresentar uma mistura genética diversificada de ancestralidade europeia, africana e nativa americana, o que poderia introduzir um viés nos cálculos de ancestralidade para a população de São Paulo. Os dados de frequências alélicas do KRGDB e gnomAD foram usados para gerar dados simulados de 500 perfis genéticos de coreanos e de judeus asquenazes. Na análise de PCA, apenas 200 amostras dessa simulação foram usadas. No total, 4.416 indivíduos de diferentes populações foram mantidos para compor o conjunto de referência populacional. Os indivíduos foram organizados com base nas origens das populações, localização geográfica, cultura, história e possíveis semelhanças genéticas entre as populações, além de consultas prévias na literatura, como por exemplo, Budiarto *et al.* (2020), Gaspar e Breen (2019), e Kumar *et al.* (2020).

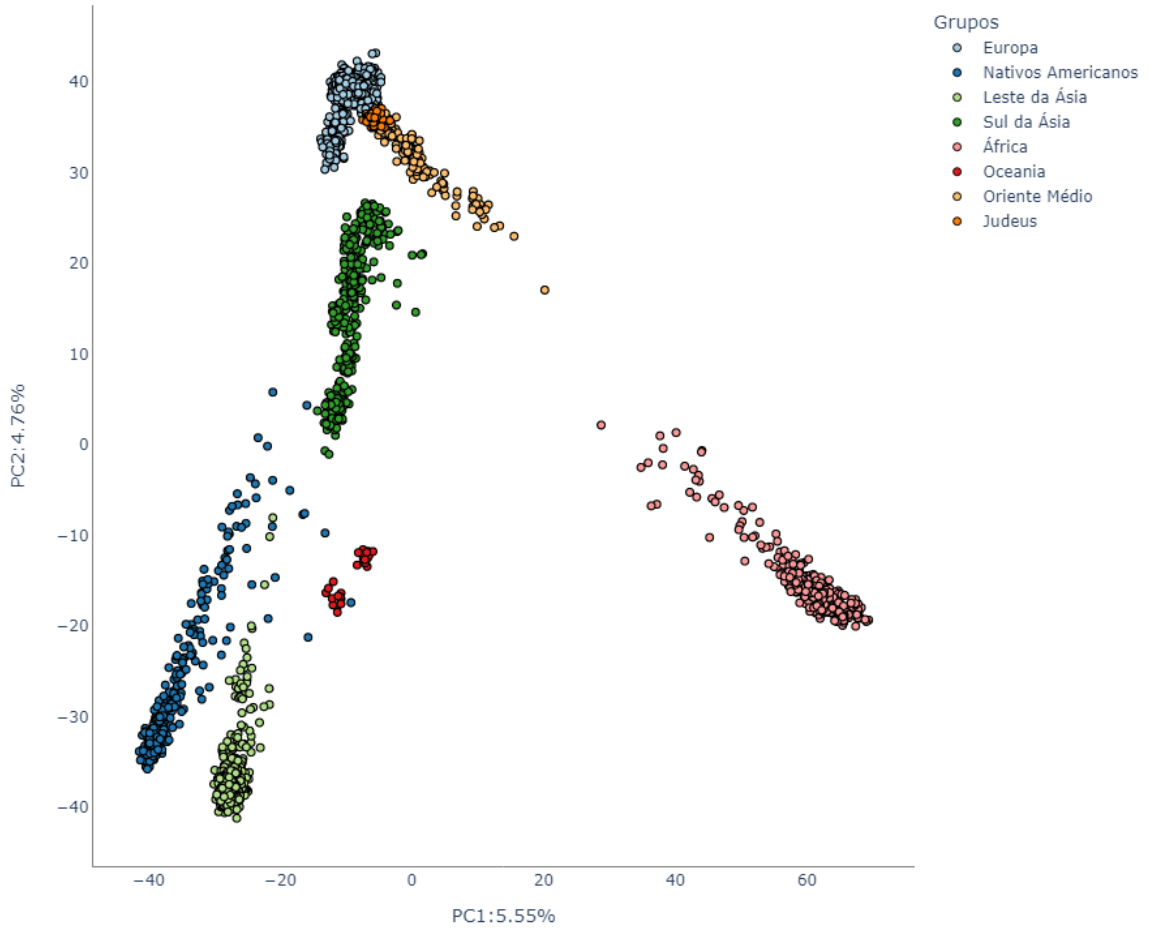
As seções seguintes contêm os resultados das análises não supervisionadas e supervisionadas de ancestralidade global.

3.1 PCA

A Figura 7 foi construída usando os valores dos dois primeiros componentes principais (PC1 e PC2) obtidos da PCA com o conjunto de referência. As cores no gráfico foram adicionadas de acordo com as informações de origem dos indivíduos disponibilizadas com os dados. Nesta figura, é possível verificar que alguns grupos

populacionais ficam distantes entre si, enquanto outros ficam bem próximos. Em geral, os indivíduos de um mesmo grupo de origem tendem a ficar mais próximos. Este resultado sugere que é possível capturar a variabilidade da ancestralidade genética por meio dos valores dos componentes principais, obtidos a partir de uma matriz de dosagem de alelos.

Figura 7 – PCA do conjunto de referência, PC1 x PC2



Fonte: Resultados originais da pesquisa

Gráficos gerados a partir dos PC1 e PC2 da análise de componentes principais, utilizando dados como os do 1kGP, geralmente apresentam uma forma triangular característica. Na análise dos dados deste projeto, o PC1 diferencia grupos africanos de não africanos, e o PC2 diferencia grupos populacionais do leste da Ásia e nativos americanos de grupos europeus. Nas figuras do Apêndice A e B, o PC3 diferencia grupos nativos americanos dos outros grupos.

Neste projeto, além dos dados do 1kGP e HGDP, também foram analisados indivíduos de outros grupos populacionais da Europa, América e África provenientes

de diferentes artigos. Além disso, os perfis genéticos de judeus e coreanos foram simulados a partir de frequências alélicas. As posições dos indivíduos dessas outras fontes de dados estão coerentes com as posições dos indivíduos do 1kGP e HGDP, o que indica que o processamento de qualidade e a junção dos dados dos diferentes estudos foram feitos adequadamente.

Na PCA, os primeiros componentes extraem a maior variabilidade dos dados originais. Na PCA realizada com todos os grupos continentais, a variabilidade dos PC1 e PC2 foi de 5,55% e 4,76%, respectivamente.

Também foi realizada a PCA com os grupos de cada grupo continental (Apêndice C à F). Por exemplo, no Apêndice F das populações asiáticas, o PC1 separa os grupos do sul da Ásia dos outros grupos, enquanto ao longo do PC2, os indivíduos do sudeste asiático estão mais próximos dos indivíduos da China do que dos indivíduos do Japão e da Coreia.

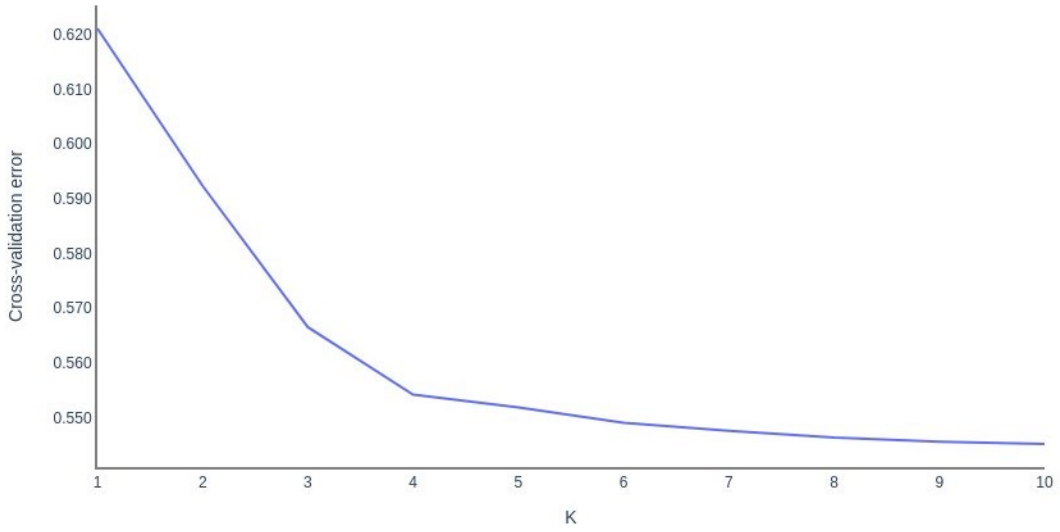
3.2 Admixture

O método não supervisionado do programa Admixture 1.3 foi usado para verificar o padrão de ancestralidade dos dados. Essa análise foi feita com todos os indivíduos do conjunto de referência e com cada grupo continental separadamente, usando diferentes valores de k grupos. Os resultados foram representados em gráficos utilizando o programa em Python, Pong v1.5 (BEHR *et al.*, 2016). Nestes gráficos, as proporções dos componentes dos k grupos são representados em cores diferentes.

A técnica do cotovelo é geralmente usada em problemas não supervisionados para determinar o número ideal de grupos em um conjunto de dados. Primeiro, o algoritmo de agrupamento é executado com diferentes números de k grupos. Em seguida, um gráfico é criado, no qual o eixo x representa o número de grupos e o eixo y representa o erro. À medida que o número de grupos aumenta, o valor de erro diminui significativamente até o ponto onde o aumento no número de grupos não resulta em uma melhora significativa. Nessa técnica, o número ideal de grupos é escolhido no ponto onde a curva se estabiliza. A Figura 9 contém os resultados do programa com k=5, que seria o valor escolhido considerando a redução de erro na Figura 8. Na Figura, 9 verifica-se que há componentes (cores) que predominam nos grupos continentais africanos, nativos americanos, leste asiáticos e da Oceania.

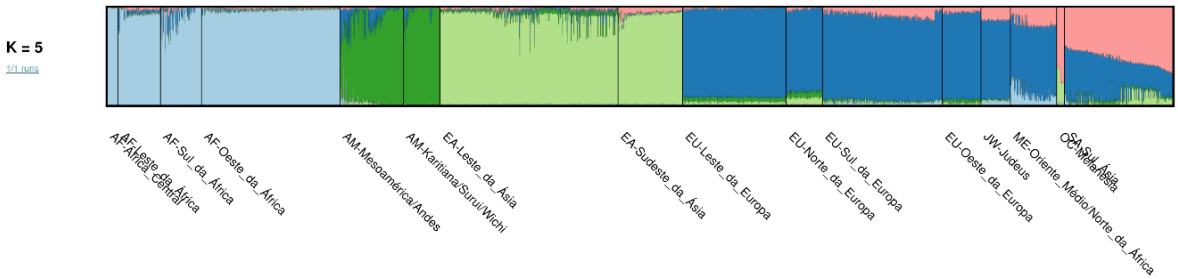
Neste conjunto de dados com k=5, os grupos da Europa compartilham predominantemente o mesmo componente com Judeus e Oriente Médio. Enquanto o Sul da Ásia ficou com um perfil intermediário, compartilhando dois componentes.

Figura 8 – Erros obtidos variando o número de k grupos no Admixture



Fonte: Resultados originais da pesquisa

Figura 9 – Resultado do Admixture com k=5



Fonte: Resultados originais da pesquisa

Neste tipo de análise, é importante visualizar o gráfico de mais de um valor de k, a fim de averiguar se o padrão observado em um k é consistente com outros valores de k, ou se há um padrão atípico que só ocorre quando o programa é executado com um determinado valor de k. No Apêndice G, no resultado com k=2, verifica-se a separação dos grupos africanos dos demais grupos, assim como ocorreu no PC1 da PCA. Com k=3, os grupos de nativos americanos e leste asiáticos se separaram. Com k=4, os grupos de nativo americanos se separaram dos grupos do leste da Ásia. Os Apêndices H à L contêm os gráficos das análises com cada grupo continental. De modo geral, conforme aumenta o k, há um maior

detalhamento de alguns grupos, enquanto outros grupos mais similares não apresentam muita mudança com aumento do k, o resultado fica apenas “ruídoso”.

Os resultados destes gráficos juntamente com os gráficos da PCA auxiliaram na escolha dos grupos populacionais para análise supervisionada. Por exemplo, pelo padrão observado nos gráficos, os grupos asiáticos foram organizados em Bengali/Punjabi, Coreia, Japão, Mongol, Paquistão, Sudeste Asiático e Sul da China.

3.3 Análise supervisionada

Informações prévias sobre origens das populações, sua localização geográfica e consultas prévias na literatura foram usadas em conjunto com os resultados do Admixture e PCA para organizar os grupos populacionais a serem usados nos modelos supervisionados de ancestralidade global. Esta organização em grupos foi feita por uma questão técnica da metodologia, pois, como visto nos resultados da PCA e do Admixture, a ancestralidade genética se estende ao longo de um continuum, no qual os indivíduos compartilham componentes entre si. A organização dos grupos populacionais foi feita em três níveis: um com seis grupos mais abrangentes, outro intermediário com dezesseis grupos, e um mais específico com vinte e três grupos, conforme a Tabela 5:

Tabela 5 – Número de indivíduos nos grupos dos três níveis

(continua)

Nível1	Nível2	Nível3	Nº
África	África Central	Biaka/Mbuti	36
	Sul da África	Khoisan	123
	Leste da África	Bantos/Quênia	160
	Oeste da África	Mandê	228
		Esan/Iorubá	243
América	Andes/Mesoamérica	Andes	156
		Mesoamérica	59
	Amazônia/Patagônia	Karitiana/Surui/Wichi	124
Leste da Ásia	China/Sudeste Asiático	Sudeste Asiático	219
		Sul da China	295
	Nordeste da Ásia	Coreia	500
		Japão	130
	Norte da Ásia	Mongol	79
Europa/Oriente Médio	Judeus	Asquenaze	500
	Leste/Norte/Oeste da Europa	Norte da Europa	98
		Leste da Europa	377
		Oeste da Europa	131

Tabela 5 – Número de indivíduos nos grupos dos três níveis

(conclusão)

Nível1	Nível2	Nível3	Nº
	Oriente Médio/Norte da África	Oriente Médio/Norte da África	157
	Sudoeste da Europa	Basco/Ibéria	127
	Sul da Europa	Albânia/Itália/Sardenha	280
Oceania	Melanesia	Papua/Bougainville	27
Sul da Ásia	Sul da Ásia	Bengali/Punjabi	204
		Paquistão	163
Total			4416

Fonte: Resultados originais da pesquisa

Diferentes modelos baseados em MLE (RegMLEMix) foram criados com valores diferentes do parâmetro λ (0, 25, 50, 100, 200 e 300). Também foram criados modelos de previsão em conjunto (VotingClassifier) combinando o RegMLEMix (λ=0) com outros modelos RegMLEMix com diferentes valores de λ (25, 50, 100, 200 e 300). Versões semelhantes destes modelos também foram criadas usando a abordagem hierárquica (HierarchicalLocalClassifier). Como explicado na seção sobre o método, este algoritmo recebe do usuário uma lista contendo os modelos para serem utilizados em cada nível. Neste caso foram três níveis, conforme a Tabela 5, um nível mais geral, outro intermediário e um mais específico. Note que em algumas partes não há subdivisões, apenas a repetição do grupo. Optou-se por essa alternativa, ao invés de criar grupos com poucos indivíduos, ou criar grupos que trariam mais complexidade para os modelos sem muito ganho de informação.

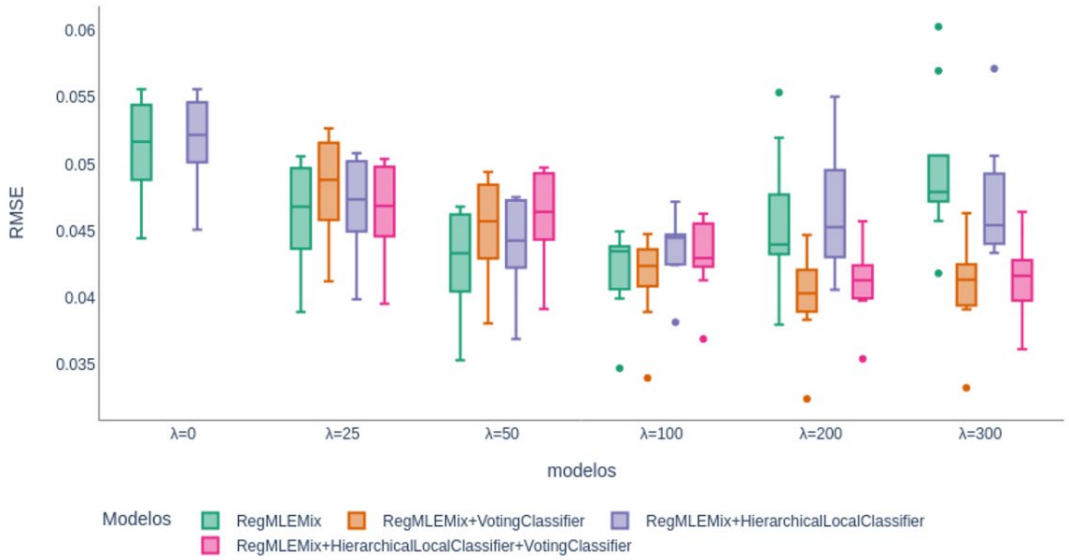
A Figura 10 contém o resultado do RMSE obtido na validação cruzada estratificada com 10 partições. Os valores de RMSE foram calculados entre as proporções de ancestralidade previstas e esperadas das amostras simuladas para os grupos do nível 3. A Tabela 6 contém os três modelos com os menores valores de RMSE na validação cruzada. O modelo com o menor valor de RMSE foi o Modelo1 que usa o VotingClassifier, cuja previsão é a média das previsões dos modelos RegMLEMix (λ=0) e RegMLEMix (λ=200). Os outros dois modelos com menores valores de RMSE também foram baseados no VotingClassifier. O modelo com o terceiro menor valor de RMSE foi montado com a abordagem de classificação hierárquica, usando o VotingClassifier nos níveis 2 e 3. O Apêndice M contém a configuração de todos os modelos usados.

Tabela 6 – Modelos com menor valor de RMSE

	Modelos	RMSE (std)
1	VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=200)	0,040 (0,003)
2	VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=300)	0,041 (0,003)
3	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=200) N3 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=200)	0,041 (0,003)

Fonte: Resultados originais da pesquisa

Figura 10 – RMSE dos modelos na validação cruzada



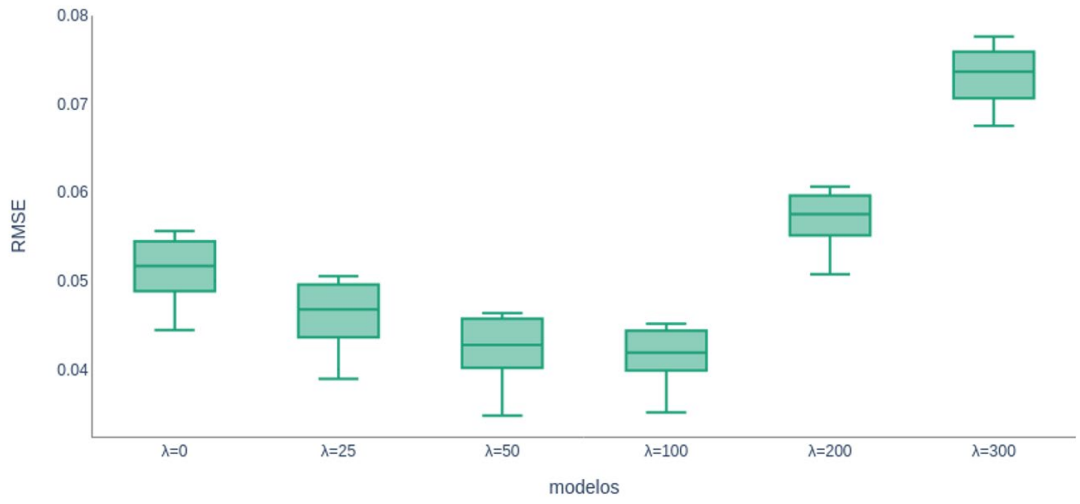
Fonte: Resultados originais da pesquisa

Em geral, para cada valor do parâmetro λ, os modelos com classificação hierárquica ficaram com RMSE próximo do obtido pelos modelos sem classificação hierárquica. Em comparação com os modelos com λ=0, em geral os modelos com λ maior que zero tendem a ter valores de RMSE um pouco menores. Os modelos com VotingClassifier, combinando um modelo com λ=0 e outro com λ diferente de 0, parecem ser mais vantajosos com valores altos de λ (200 e 300).

Separadamente, o método supervisionado do programa Admixture 1.3 também foi avaliado em validação cruzada, com diferentes valores do parâmetro de regularização λ. Os valores de RMSE calculados entre as proporções de ancestralidades obtidas e esperadas estão na Figura 11. O programa obteve o menor valor médio de RMSE, 0,041 (std=0,003), quando executado com λ=100. Este

valor de RMSE é similar ao obtido pelos modelos supervisionados deste projeto que tiveram o melhor desempenho:

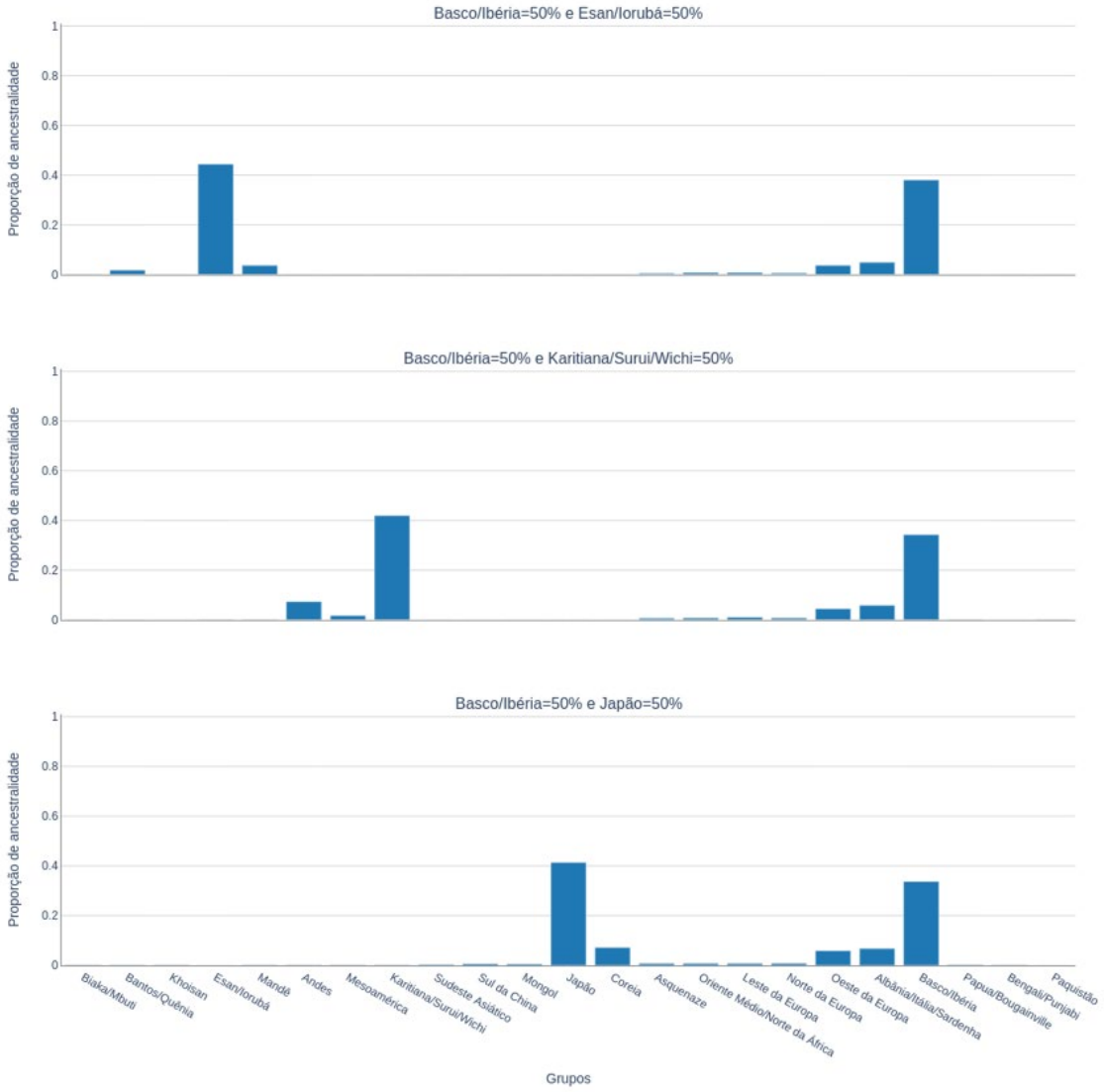
Figura 11 – RMSE do programa Admixture 1.3 com diferentes valores de regularização



Fonte: Resultados originais da pesquisa

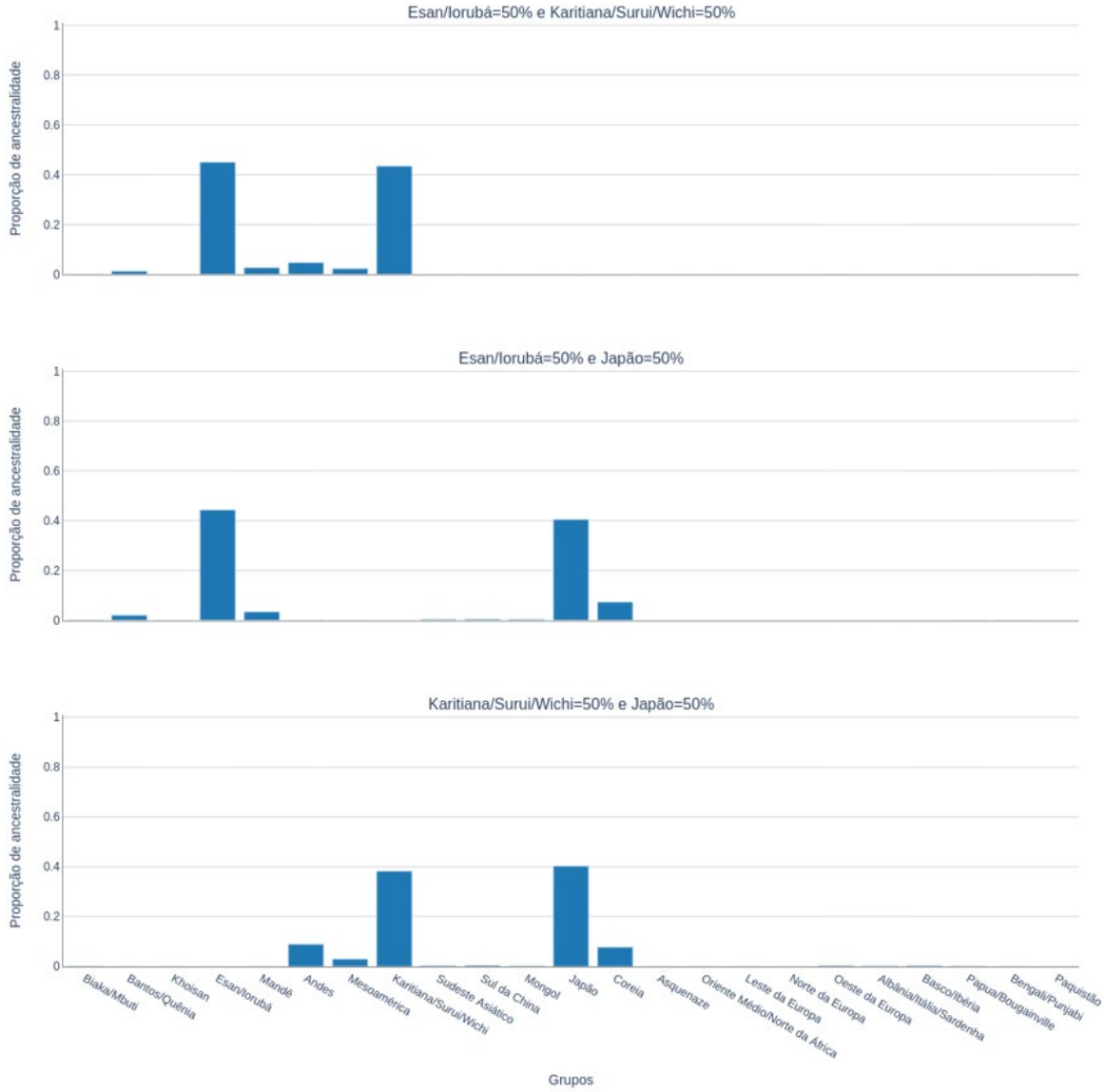
Considerando os resultados das amostras simuladas separadas nas partições de validação, as Figuras 12 e 13 foram criadas com a média das proporções de ancestralidade calculadas pelo Modelo1 para amostras simuladas entre grupos continentais. Por exemplo, o primeiro gráfico da Figura 12 mostra a média de ancestralidade de amostras simuladas (Esan/Iorubá=50% e Basco/Ibéria=50%). A média dos valores estimados para Esan/Iorubá foi de 44,42% e a média para a Basco/Ibéria foi de 38,06%. Em geral o Modelo1 atribui a maior parte das proporções de ancestralidade para os respectivos grupos populacionais esperados, mas também inferiu certa proporção a outros grupos similares dentro do mesmo grupo continental, por exemplo entre Japão e Coreia do Leste da Ásia.

Figura 12 – Média de ancestralidade prevista para simulação (Europa, Nativos Americanos, África e Ásia)



Fonte: Resultados originais da pesquisa

Figura 13 – Média de ancestralidade prevista para simulação (Ásia, África e Nativos Americanos)

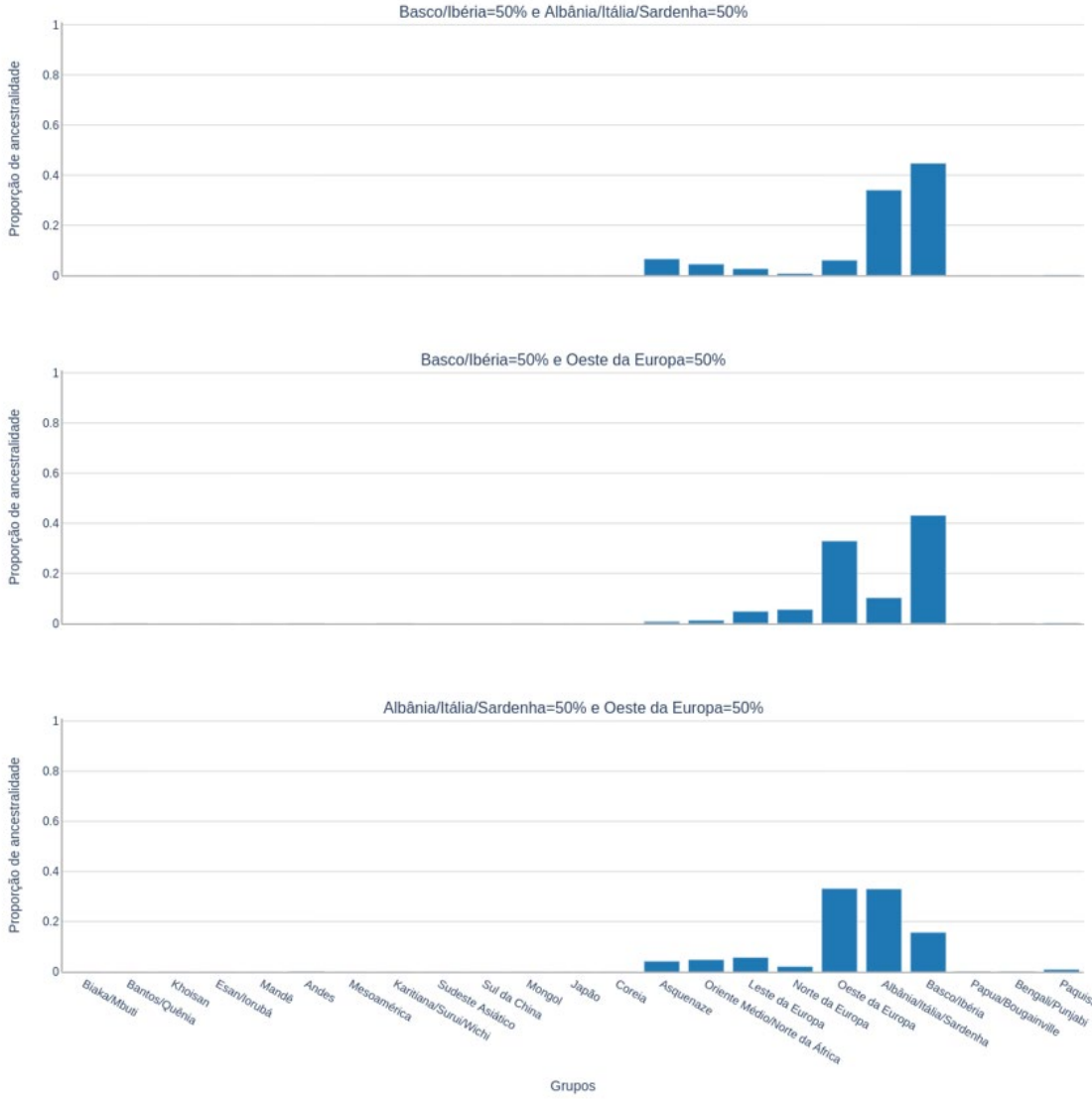


Fonte: Resultados originais da pesquisa

A Figura 14 mostra a média das proporções de ancestralidade calculadas pelo Modelo1 para as amostras simuladas entre grupos da Europa. Nas amostras simuladas entre Albânia/Itália/Sardenha (50%) e Basco/Ibéria (50%), a média estimada de Basco/Ibéria foi de 44,77% e a média de Albânia/Itália/Sardenha foi de 34,03%. Nas amostras simuladas entre Oeste da Europa (50%) e Basco/Ibéria (50%), a média estimada para Basco/Ibéria foi de 43,17% e a média para Europa Ocidental foi de 32,96%. Nas amostras simuladas entre Europa Ocidental (50%) e Albânia/Itália/Sardenha (50%), a média de Albânia/Itália/Sardenha foi de 33,01% e a média de Europa Ocidental foi de 33,18%. Mesmo entre grupos geneticamente

similares, o modelo conseguiu atribuir as proporções de ancestralidade, mas nestes três conjuntos de amostras simuladas, o Modelo1 também inferiu proporções de ancestralidade de outros grupos da Europa e do Oriente Médio. Além da capacidade do modelo em distinguir grupos similares, esse padrão pode ser explicado pelo passado de migrações e miscigenações entre populações europeias. Além disso, a composição e organização dos grupos populacionais no conjunto de referência também podem ter influenciado o desempenho do modelo.

Figura 14 – Média de ancestralidade prevista para simulação (Oeste da Europa, Albânia/Itália/Sardenha, e Basco/Ibéria)

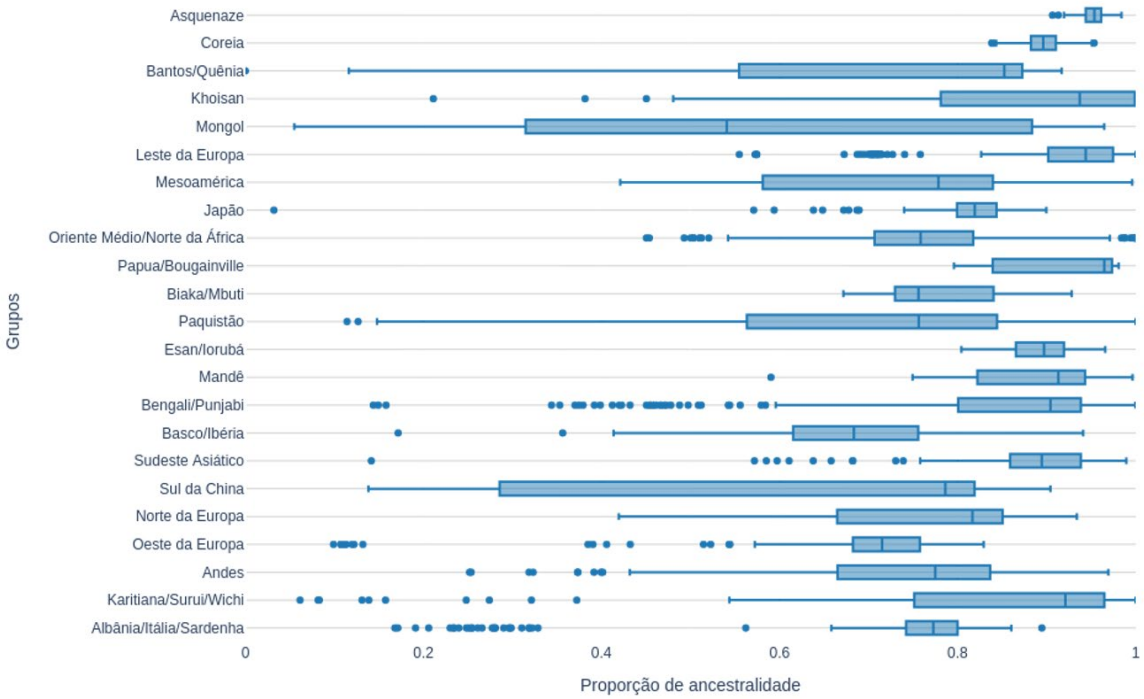


Fonte: Resultados originais da pesquisa

A Figura 15 apresenta as distribuições dos valores de ancestralidade previstos pelo Modelo1 para os indivíduos do conjunto de referência nas partições

de validação. Cada boxplot representa a distribuição de ancestralidade dos indivíduos pertencentes aos respectivos grupos. Em média, a mediana dos valores previstos para cada grupo foi de 82,77%. O Modelo 1 teve dificuldade com alguns grupos, como Mongol e Sul da China, que tiveram valores do primeiro quartil iguais a 31,56% e 28,59%, respectivamente.

Figura 15 – Distribuição dos valores previstos pelo Modelo1 para cada grupo do conjunto referência



Fonte: Resultados originais da pesquisa

A Tabela 7 contém os valores de precisão e sensibilidade do Modelo1, calculados na validação cruzada para cada um dos grupos populacionais do conjunto de referência. Mongol e Sul da China tiveram os menores valores de sensibilidade. Em geral, todos os grupos tiveram bons valores de precisão, todos acima de 80%. Em média a precisão do Modelo1 foi 0,960 e a sensibilidade foi de 0,943.

Tabela 7 – Valores de precisão e sensibilidade do Modelo1 para cada um dos grupos populacionais

(continua)

Grupos	Precisão	Sensibilidade
Biaka/Mbuti	1,00	1,00
Bantos/Quênia	1,00	0,84

Tabela 7 – Valores de precisão e sensibilidade do Modelo1 para cada um dos grupos populacionais

(conclusão)

Grupos	Precisão	Sensibilidade
Khoisan	0,93	0,99
Esan/Iorubá	0,96	1,00
Mandê	1,00	1,00
Andes	0,96	0,98
Mesoamérica	0,99	1,00
Karitiana/Surui/Wichi	0,99	0,92
Sudeste Asiático	0,84	0,99
Sul da China	0,97	0,71
Mongol	0,90	0,63
Japão	1,00	0,99
Coreia	0,90	1,00
Asquenaze	1,00	1,00
Oriente Médio/Norte da África	1,00	1,00
Leste da Europa	1,00	1,00
Norte da Europa	1,00	1,00
Oeste da Europa	0,99	0,92
Albânia/Itália/Sardenha	0,98	0,90
Basco/Ibéria	0,83	0,98
Papua/Bougainville	1,00	1,00
Bengali/Punjabi	0,96	0,88
Paquistão	0,88	0,95

Fonte: Resultados originais da pesquisa

3.4 Ancestralidade de indivíduos de São Paulo

Considerando o conjunto de referência com populações recentes usadas neste projeto, a média da ancestralidade global estimada pelo Modelo1 para os 411 indivíduos de São Paulo deste projeto pode ser verificada na Tabela 8. Lembrando que o Modelo1 usa o VotingClassifier, cuja previsão de cada amostra é a média das previsões dos modelos RegMLEMix (λ=0) e RegMLEMix (λ=200), e foi o modelo com o menor valor de RMSE na validação cruzada.

Tabela 8. Ancestralidade média dos indivíduos de São Paulo

(continua)

Grupos populacionais	Ancestralidade média (IC)
Basco/Ibéria	0,359 ±0,020
Albânia/Itália/Sardenha	0,245 ±0,022
Oeste da Europa	0,067 ±0,005
Esan/Iorubá	0,063 ±0,009

Tabela 8. Ancestralidade média dos indivíduos de São Paulo

(conclusão)

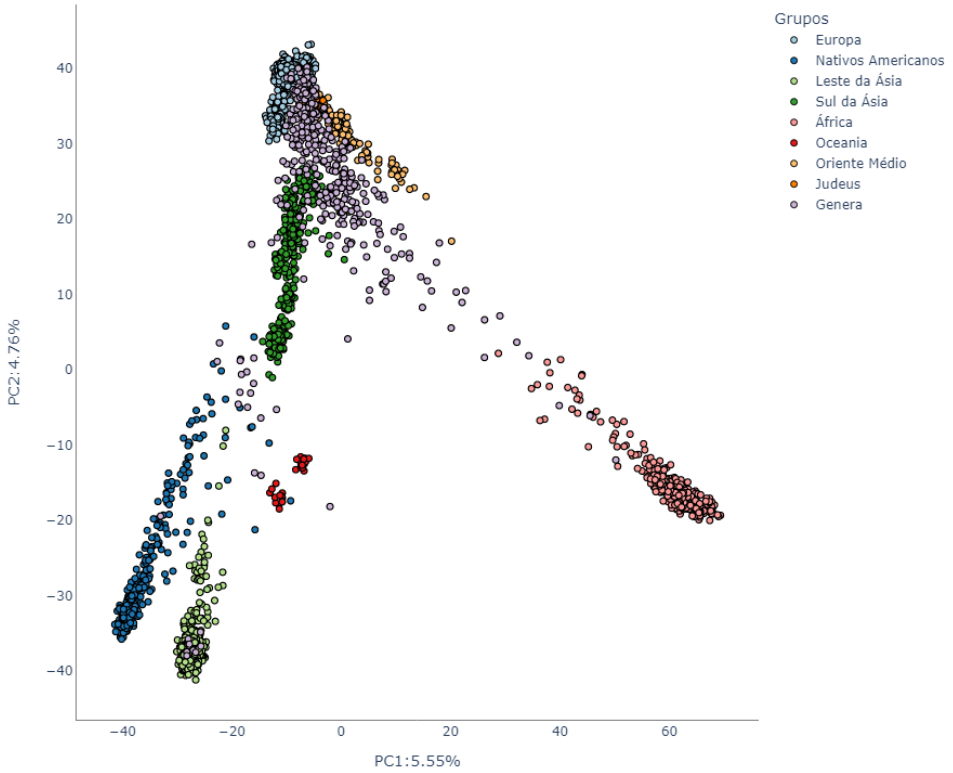
Grupos populacionais	Ancestralidade média (IC)
Karitiana/Surui/Wichi	0,041 ±0,004
Oriente Médio/Norte da África	0,034 ±0,004
Japão	0,033 ±0,014
Leste da Europa	0,032 ±0,006
Asquenaze	0,029 ±0,006
Mandê	0,021 ±0,003
Andes	0,017 ±0,005
Bantos/Quênia	0,016 ±0,003
Mesoamérica	0,016 ±0,002
Norte da Europa	0,009 ±0,001
Coreia	0,005 ±0,002
Paquistão	0,003 ±0,001
Biaka/Mbuti	0,002 ±0,000
Bengali/Punjabi	0,002 ±0,001
Khoisan	0,001 ±0,000
Papua/Bougainville	0,001 ±0,000
Sudeste da Ásia	0,001 ±0,000
Mongol	0,001 ±0,000
Sul da China	0,001 ±0,000

Fonte: Resultados originais da pesquisa

De acordo com os resultados do Modelo1, em média, os indivíduos possuem maior similaridade genética com populações dos grupos Basco/Ibéria, Albânia/Itália/Sardenha, Europa Ocidental, Esan/Iorubá, Karitiana/Surui/Wichi, Oriente Médio/Norte da África e Japão. Ao considerar os grupos continentais, a média de ancestralidade global foi de 0,775 europeia, 0,104 africana, 0,074 nativo americana, 0,041 leste asiática, 0,005 sul asiática, 0,001 oceânica.

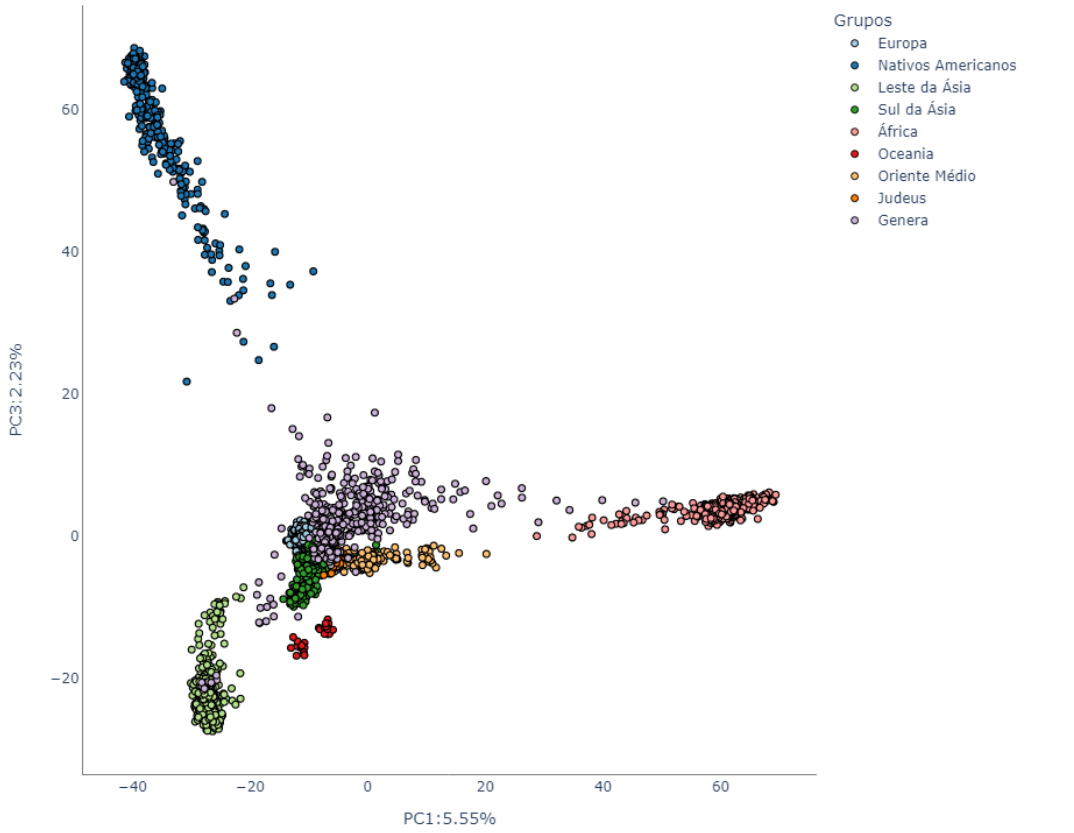
As Figuras 16, 17 e 18 mostram o resultado da PCA, no qual os indivíduos de São Paulo foram projetados no espaço construído previamente com o conjunto de referência. Há indivíduos distribuídos entre os grupos de origem africana, nativa americana, europeia e asiática do conjunto de referência, exemplificando a ideia de continuidade da ancestralidade genética.

Figura 16 – PCA com indivíduos de São Paulo (Genera): PC1 x PC2



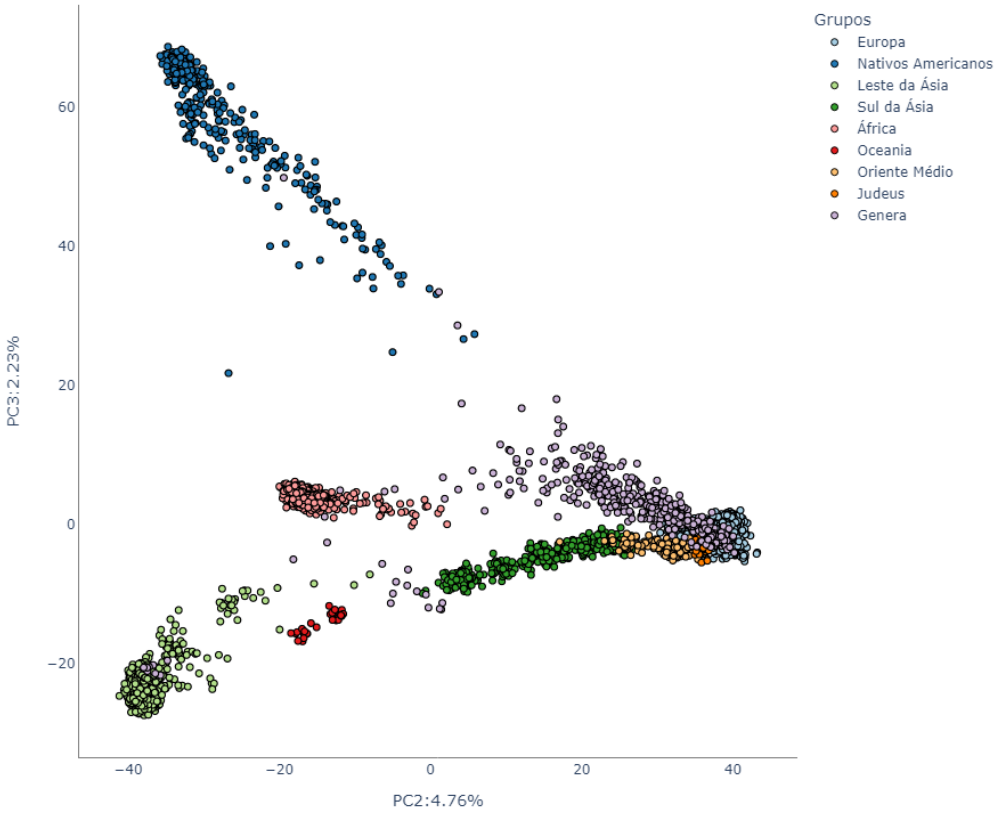
Fonte: Resultados originais da pesquisa

Figura 17 – PCA com indivíduos de São Paulo (Genera): PC1 x PC3



Fonte: Resultados originais da pesquisa

Figura 18 – PCA com indivíduos de São Paulo (Genera): PC2 x PC3



Fonte: Resultados originais da pesquisa

4 DISCUSSÃO

A estrutura triangular observada nos gráficos 7 e 16 com os componentes principais 1 e 2 está bastante similar ao que costuma ser observado na literatura, como em Budiarto *et al.* (2020), Gaspar e Breen (2019) e Koenig *et al.* (2023). A PCA mostrou que os grupos populacionais africanos, leste asiáticos e europeus ocupam os vértices, enquanto os indivíduos miscigenados são posicionados entre eles. Supondo que a ancestralidade continental seja uma das principais fontes de variabilidade, a PCA tende a dispor esses grupos a uma distância considerável uns dos outros. Esse padrão é útil como uma verificação de integridade dos dados, proporcionando uma representação visual que confirma uma distribuição esperada dos grupos populacionais (DIAZ-PAPKOVICH; ANDERSON-TROCMÉ; GRAVEL, 2021).

Os valores obtidos para a variabilidade dos dois primeiros componentes principais parecem baixos, mas é o que se costuma observar para esse tipo de dado, estando um pouco maiores do que os obtidos no trabalho de Gaspar e Breen (2019). O fato de o projeto atual ter utilizado mais indivíduos e um conjunto de SNPs diferentes pode explicar a pequena variação.

Algumas considerações quanto aos resultados obtidos na PCA são pertinentes. Os dados faltantes presentes no conjunto de dados foram imputados com o valor mais frequente. Apesar de não ser a solução ótima, para a finalidade deste projeto, essa alternativa foi suficiente. Além disso, os componentes extraídos praticamente não são afetados quando sua variância é muito maior do que o ruído causado pelos dados faltantes (EIDELMAN, 2020; PRIVÉ *et al.*, 2020). Por esta razão, no pré-processamento, os dados foram limitados aos SNPs em comum entre o conjunto de referência e os dados da Genera, e os SNPs com mais de 1% de genótipos faltantes foram excluídos. Como a quantidade de dados era muito grande, essa perda de SNPs não foi um problema para os propósitos deste projeto. No entanto, se o objetivo fosse analisar outros conjuntos de dados com baixa sobreposição de SNPs, o indicado seria aplicar imputação genômica (GASPAR; BREEN, 2019).

Embora os PCs possam ser usados para visualização e permitir ter uma ideia da distribuição dos indivíduos, a PCA por si não é capaz de fazer análise de classificação. Se esse for o objetivo, deve-se ou utilizar os PCs como variáveis em

outros modelos, ou usar modelos capazes de reduzir a dimensionalidade dos dados e realizar classificação, como a Análise Discriminante Linear (CHEUNG; GAHAN; MCNEVIN, 2018; GASPAR; BREEN, 2019). Por fim, deve-se ter cuidado ao interpretar os gráficos da PCA, pois as distâncias observadas podem não ser uma equivalência exata das distâncias genéticas entre as populações, e a estrutura geral entre os indivíduos pode variar de acordo com as amostras e variantes dos dados usados na análise. Elhaik (2022), em seu estudo, realizou testes avaliando o uso da PCA e alertou sobre as limitações da PCA e seu uso em pesquisas de ancestralidade, incluindo considerações quanto a interpretação de distância genética, alterações nos resultados da PCA devido a mudanças no número de amostras, escolha das populações de referência, entre outras considerações.

Em relação a análise supervisionada, algumas considerações sobre a validação também são importantes. Neste projeto, foi usado a validação cruzada estratificada para avaliar os modelos, mas outras abordagens mais exaustivas poderiam ter sido usadas. Na abordagem “leave one out” (deixar um fora), uma amostra é reservada para avaliação e o modelo é treinado com o restante dos dados. O processo se repete n vezes, alterando a amostra que é reservada até passar por todas. Outro método seria a validação cruzada aninhada, na qual há um laço interno para otimização dos hiperparâmetros dos modelos e um laço externo para avaliação (CAWLEY; TALBOT, 2010). Para otimização de hiperparâmetros, ao invés de usar uma busca exaustiva de combinações de parâmetros, também seria possível aplicar uma busca aleatória, na qual apenas um conjunto das combinações é sorteado para avaliação.

Também é importante levar em consideração que os genótipos de judeus e coreanos foram simulados a partir das frequências alélicas, o que pode ter causado uma superestimação do desempenho para esses dois grupos, mesmo havendo separação dos dados em treino e validação. Além disso, o desbalanceamento de indivíduos nos grupos populacionais pode ter afetado os resultados. Em um estudo futuro, pode-se tentar amenizar esses dois pontos com a busca e incorporação de novos dados.

Embora tenha sido possível obter bons valores de precisão e sensibilidade dentro do escopo e tempo deste projeto, a assertividade dos resultados de ancestralidade global depende de vários fatores, como a representatividade das populações e escolha dos modelos. Neste projeto foi possível obter dados genéticos

de diferentes populações, no entanto, alguns grupos foram mais difíceis de obter devido à baixa disponibilidade e representatividade, como as populações nativa americanas e africanas. Recentemente, iniciativas de projetos que incluem essas populações podem ajudar amenizar essa dificuldade e possibilitar que pesquisas tenham um maior entendimento sobre a estrutura genômica de populações miscigenadas (DE OLIVEIRA; SECOLIN; LOPES-CENDES, 2023).

Quanto aos resultados obtidos para os indivíduos de São Paulo, embora os resultados possam ter um viés amostral, uma vez que os dados analisados são provenientes de clientes da Genera, os resultados parecem estar dentro do esperado. Comparando com o resultado de uma coorte baseada em censo de 1171 idosos de São Paulo que fazem parte do estudo Saúde, Bem-estar e Envelhecimento (SABE), a média da ancestralidade global do SABE consultado no estudo dos autores Naslavsky *et al.*, (2022) foi de 0,726 europeia, 0,178 africana, 0,067 nativo americana, e 0,028 leste asiática.

Além de ser importante como ferramenta de autoconhecimento, a inferência de ancestralidade genética, tem recebido destaque como uma análise complementar de outros testes genéticos, como o Escore de Risco Poligênico. Estudos têm demonstrado que a acurácia deste teste é impactada à medida que a distância genética entre as populações de treinamento e alvo aumenta (KACHURI *et al.*, 2024). Isso reforça a importância do desenvolvimento de ferramentas para auxiliar no ajuste ou na transferência dos escores de risco entre populações. Nesse contexto, em um próximo estudo, poderia ser interessante avaliar a aplicação de abordagens vistas neste projeto, como previsão em conjunto e/ou explorar outras metodologias, como ancestralidade local para integrar os resultados de GWAS de diferentes populações para otimizar a construção de Escores de Risco Poligênico.

5 CONCLUSÃO

Para atender aos objetivos deste projeto, foram obtidos dados genéticos que representam a diversidade étnica da população de São Paulo, assim como dados de indivíduos dessa população. Os dados foram tratados seguindo práticas comuns de processamento de qualidade, e analisados com ferramentas de Bioinformática. A aplicação da Análise de Componentes Principais permitiu visualizar a estrutura genética dos grupos populacionais, incluindo São Paulo. Além disso, modelos de ancestralidade global baseados em Estimativa de Máxima Verossimilhança e classificação hierárquica foram implementados em Python, os quais obtiveram um bom desempenho na validação cruzada. Os resultados obtidos exemplificam a eficácia da combinação de mais de um modelo na inferência da ancestralidade genética, além da relevância das técnicas de aprendizado de máquina na compreensão da diversidade de populações complexas, possibilitando inferir a ancestralidade de indivíduos da população de São Paulo considerando 23 grupos populacionais.

DECLARAÇÃO

Eu, Raphael Bruno Amemiya, pesquisador responsável pelo presente projeto, declaro que houve um potencial conflito de interesse devido ao vínculo empregatício com a empresa Genera. No entanto, o vínculo empregatício não interferiu na condução do trabalho acadêmico, cujas atividades foram conduzidas de forma independente e imparcial.

REFERÊNCIAS

ABRAMOVS, Nikita; BRASS, Andrew; TASSABEHJI, May. Hardy-Weinberg equilibrium in the large scale genomic sequencing era. Frontiers in genetics, v. 11, p. 516957, 2020.

ALEXANDER, David H.; LANGE, Kenneth. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC bioinformatics, v. 12, p. 1-6, 2011.

ALEXANDER, David H.; NOVEMBRE, John; LANGE, Kenneth. Fast model-based estimation of ancestry in unrelated individuals. Genome research, v. 19, n. 9, p. 1655-1664, 2009.

ANDRADE, Roberta B. et al. Estimating Asian contribution to the Brazilian population: a new application of a validated set of 61 ancestry informative markers. G3: Genes, Genomes, Genetics, v. 8, n. 11, p. 3577-3582, 2018.

BANSAL, Vikas; LIBIGER, Ondrej. Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. BMC bioinformatics, v. 16, p. 1-11, 2015.

BARBOSA, Fernanda B. et al. Ancestry Informative Marker Panel to Estimate Population Stratification Using Genome-wide Human Array. Annals of human genetics, v. 81, n. 6, p. 225-233, 2017.

BEHR, Aaron A. et al. Pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics, v. 32, n. 18, p. 2817-2823, 2016.

BERGSTRÖM, Anders et al. Insights into human genetic variation and population history from 929 diverse genomes. Science, v. 367, n. 6484, p. eaay5012, 2020.

BRISBIN, Abra et al. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Human biology, v. 84, n. 4, p. 343, 2012.

BUDIARTO, Arif et al. Gaussian mixture model implementation for population stratification estimation from genomics data. Procedia Computer Science, v. 179, p. 202-210, 2021.

BYRSKA-BISHOP, Marta et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell, v. 185, n. 18, p. 3426-3440. e19, 2022.

CABREROS, Irineo; STOREY, John D. A likelihood-free estimator of population structure bridging admixture models and principal components analysis. Genetics, v. 212, n. 4, p. 1009-1029, 2019.

CAWLEY, Gavin C.; TALBOT, Nicola LC. On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research, v. 11, p. 2079-2107, 2010.

CHANG, Christopher C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience, v. 4, n. 1, p. s13742-015-0047-8, 2015.

CHEUNG, Elaine YY; GAHAN, Michelle Elizabeth; MCNEVIN, Dennis. Prediction of biogeographical ancestry in admixed individuals. Forensic Science International: Genetics, v. 36, p. 104-111, 2018.

DE OLIVEIRA, Thais C.; SECOLIN, Rodrigo; LOPES-CENDES, Iscia. A review of ancestrality and admixture in Latin America and the caribbean focusing on native American and African descendant populations. Frontiers in Genetics, v. 14, p. 1091269, 2023.

DI GAETANO, Cornelia et al. An overview of the genetic structure within the Italian population from genome-wide data. 2012.

DIAZ-PAPKOVICH, Alex; ANDERSON-TROCMÉ, Luke; GRAVEL, Simon. A review of UMAP in population genetics. Journal of Human Genetics, v. 66, n. 1, p. 85-91, 2021.

DUFORET-FREBOURG, Nicolas et al. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. Molecular biology and evolution, v. 33, n. 4, p. 1082-1093, 2016.

DURAND, Eric Y. et al. Ancestry composition: a novel, efficient pipeline for ancestry deconvolution. biorxiv, p. 010512, 2014.

EIDELMAN, Alexis. Python data science handbook by jake VANDERPLAS (2016). Statistique et Société, v. 8, n. 2, p. 45-47, 2020.

ELHAIK, Eran. Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Scientific Reports, v. 12, n. 1, p. 14683, 2022.

FÁVERO, Luiz Paulo; BELFIORE, Patrícia. Data science for business and decision making. Academic Press, 2019.

FRUDAKIS, Tony. Molecular photofitting: predicting ancestry and phenotype using DNA. Elsevier, 2010.

GASPAR, Héléna A.; BREEN, Gerome. Probabilistic ancestry maps: a method to assess and visualize population substructures in genetics. BMC bioinformatics, v. 20, p. 1-11, 2019.

GÉRON, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.", 2022.

GIOLO, Suely R. et al. Brazilian urban population genetic structure reveals a high degree of admixture. European Journal of Human Genetics, v. 20, n. 1, p. 111-116, 2012.

GNECCHI-RUSCONE, Guido Alberto et al. Dissecting the pre-Columbian genomic ancestry of Native Americans along the Andes–Amazonia divide. Molecular biology and evolution, v. 36, n. 6, p. 1254-1269, 2019.

HAGBERG, Aric; SWART, Pieter; S CHULT, Daniel. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

Hail Team. Hail 0.2. Disponível em: https://github.com/hail-is/hail

HALKO, Nathan; MARTINSSON, Per-Gunnar; TROPP, Joel A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review, v. 53, n. 2, p. 217-288, 2011.

HARISMENDY, Olivier et al. Evaluating and sharing global genetic ancestry in biomedical datasets. Journal of the American Medical Informatics Association, v. 26, n. 5, p. 457-461, 2019.

HELLENTHAL, Garrett et al. A genetic atlas of human admixture history. science, v. 343, n. 6172, p. 747-751, 2014.

HOLLOWAY, John W.; PRESCOTT, Susan L. The origins of allergic disease. In: Middleton's allergy essentials. Elsevier, 2017. p. 29-50.

GANNETT, Lisa. Biogeographical ancestry and race. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, v. 47, p. 173-184, 2014.

JORDE, Lynn B.; BAMSHAD, Michael J. Genetic ancestry testing: what is it and why is it important?. Jama, v. 323, n. 11, p. 1089-1090, 2020.

JUNG, Kwang Su et al. KRGDB: the large-scale variant database of 1722 Koreans based on whole genome sequencing. Database, v. 2020, p. baz146, 2020.

KACHURI, Linda et al. Principles and methods for transferring polygenic risk scores across global populations. Nature Reviews Genetics, v. 25, n. 1, p. 8-25, 2024.

KARCZEWSKI, Konrad J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, v. 581, n. 7809, p. 434-443, 2020.

KOENIG, Zan et al. A harmonized public resource of deeply sequenced diverse human genomes. bioRxiv, 2023.

KUMAR, Arvind et al. Xgmix: Local-ancestry inference with stacked xgboost. BioRxiv, p. 2020.04. 21.053876, 2020.

LEUTENEGGER, Anne-Louise et al. Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us?. European Journal of Human Genetics, v. 19, n. 5, p. 583-587, 2011.

MIENYE, Ibomoiye Domor; SUN, Yanxia. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access, v. 10, p. 99129-99149, 2022.

MYCHALECKYJ, Josyf C. et al. Genome-wide analysis in Brazilians reveals highly differentiated Native American genome regions. Molecular Biology and Evolution, v. 34, n. 3, p. 559-574, 2017.

NASLAVSKY, Michel S. et al. Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil. Nature communications, v. 13, n. 1, p. 1004, 2022.

OLSON, Randal S. et al. Data-driven advice for applying machine learning to bioinformatics problems. In: Pacific Symposium on Biocomputing 2018: Proceedings of the Pacific Symposium. 2018. p. 192-203.

PADHUKASAHASRAM, Badri. Inferring ancestry from population genomic data and its applications. Frontiers in genetics, v. 5, p. 98635, 2014.

PAL, Nikhil R. et al. A possibilistic fuzzy c-means clustering algorithm. IEEE transactions on fuzzy systems, v. 13, n. 4, p. 517-530, 2005.

PASANIUC, Bogdan et al. Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. Bioinformatics, v. 29, n. 11, p. 1407-1415, 2013.

PASCHOU, Peristera et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS genetics, v. 3, n. 9, p. e160, 2007.

PEARMAN, William S.; URBAN, Lara; ALEXANDER, Alana. Commonly used Hardy–Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data. Molecular Ecology Resources, v. 22, n. 7, p. 2599-2613, 2022.

PEDREGOSA, Fabian et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, v. 12, p. 2825-2830, 2011.

PENA, Sergio DJ; SANTOS, Fabrício R.; TARAZONA-SANTOS, Eduardo. Genetic admixture in Brazil. In: American Journal of Medical Genetics Part C: Seminars in Medical Genetics. Hoboken, USA: John Wiley & Sons, Inc., 2020. p. 928-938.

PENG, Bo; KIMMEL, Marek. simuPOP: a forward-time population genetics simulation environment. Bioinformatics, v. 21, n. 18, p. 3686-3687, 2005.

PEREIRA, Fabiana dos Santos Carolino Firmo et al. A systematic literature review on the European, African and Amerindian genetic ancestry components on Brazilian health outcomes. Scientific Reports, v. 9, n. 1, p. 8874, 2019.

PHILLIPS, Andelka M. Only a click away—DTC genetics for ancestry, health, love… and more: A view of the business and regulatory landscape. Applied & translational genomics, v. 8, p. 16-22, 2016.

PRIVÉ, Florian et al. Efficient toolkit implementing best practices for principal component analysis of population genetic data. Bioinformatics, v. 36, n. 16, p. 4449-4457, 2020.

PURCELL, Shaun; CHANG, Christopher. PLINK 1.9. Disponível em www.cog-genomics.org/plink/1.9/. Acesso em: 10 jan. 2024.

RASCHKA, Sebastian; LIU, Yuxi Hayden; MIRJALILI, Vahid. Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python. Packt Publishing Ltd, 2022.

RAVEANE, Alessandro et al. Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. Science Advances, v. 5, n. 9, p. eaaw3492, 2019.

ROTIMI, Charles N.; JORDE, Lynn B. Ancestry and disease in the age of genomic medicine. New England Journal of Medicine, v. 363, n. 16, p. 1551-1558, 2010.

ROYAL, Charmaine D. et al. Inferring genetic ancestry: opportunities, challenges, and implications. The American Journal of Human Genetics, v. 86, n. 5, p. 661-673, 2010.

SAGI, Omer; ROKACH, Lior. Ensemble learning: A survey. Wiley interdisciplinary reviews: data mining and knowledge discovery, v. 8, n. 4, p. e1249, 2018.

SCHAEFER, Nathan K.; SHAPIRO, Beth; GREEN, Richard E. AD-LIBS: inferring ancestry across hybrid genomes using low-coverage sequence data. BMC bioinformatics, v. 18, p. 1-22, 2017.

SCHLEBUSCH, Carina M. et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science, v. 338, n. 6105, p. 374-379, 2012.

SHRINER, Daniel. Overview of admixture mapping. Current protocols in human genetics, v. 94, n. 1, p. 1.23. 1-1.23. 8, 2017.

SHRINGARPURE, Suyash; XING, Eric P. Effects of sample selection bias on the accuracy of population structure and ancestry inference. G3: Genes, Genomes, Genetics, v. 4, n. 5, p. 901-911, 2014.

SILLA, Carlos N.; FREITAS, Alex A. A survey of hierarchical classification across different application domains. Data mining and knowledge discovery, v. 22, p. 31-72, 2011.

SOUZA, Aracele Maria de et al. A systematic scoping review of the genetic ancestry of the Brazilian population. Genetics and Molecular Biology, v. 42, p. 495-508, 2019.
SUDMANT, Peter H. et al. An integrated map of structural variation in 2,504 human genomes. Nature, v. 526, n. 7571, p. 75-81, 2015.

TAN, Taotao; ATKINSON, Elizabeth G. Strategies for the genomic analysis of admixed populations. Annual review of biomedical data science, v. 6, p. 105-127, 2023.

THORNTON, Timothy A.; BERMEJO, Justo Lorenzo. Local and global ancestry inference and applications to genetic association analysis for admixed populations. Genetic epidemiology, v. 38, n. S1, p. S5-S12, 2014.

URNIKYTE, Alina et al. Patterns of genetic structure and adaptive positive selection in the Lithuanian population from high-density SNP data. Scientific reports, v. 9, n. 1, p. 9163, 2019.

VAROQUAUX, Gael; GRISEL, Olivier. Joblib: running python function as pipeline jobs. packages. python. org/joblib, 2009.

VIRTANEN, Pauli et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods, v. 17, n. 3, p. 261-272, 2020.

WATHEN, Michael J. et al. LEI: a novel allele frequency-based feature selection method for multi-ancestry admixed populations. Scientific reports, v. 9, n. 1, p. 11103, 2019.

WOLLSTEIN, Andreas; LAO, Oscar. Detecting individual ancestry in the human genome. Investigative genetics, v. 6, p. 1-12, 2015.

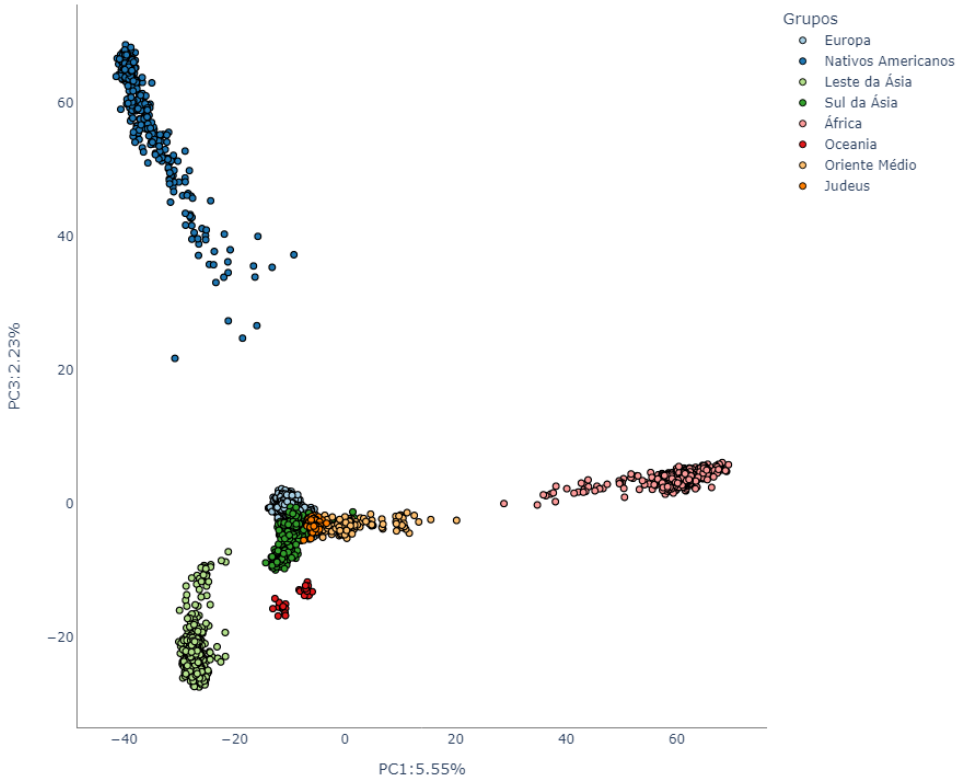
YUAN, Kai et al. Models, methods and tools for ancestry inference and admixture analysis. Quantitative Biology, v. 5, n. 3, p. 236-250, 2017.

ZHAO, Shilin et al. Strategies for processing and quality control of Illumina genotyping arrays. Briefings in bioinformatics, v. 19, n. 5, p. 765-775, 2018.

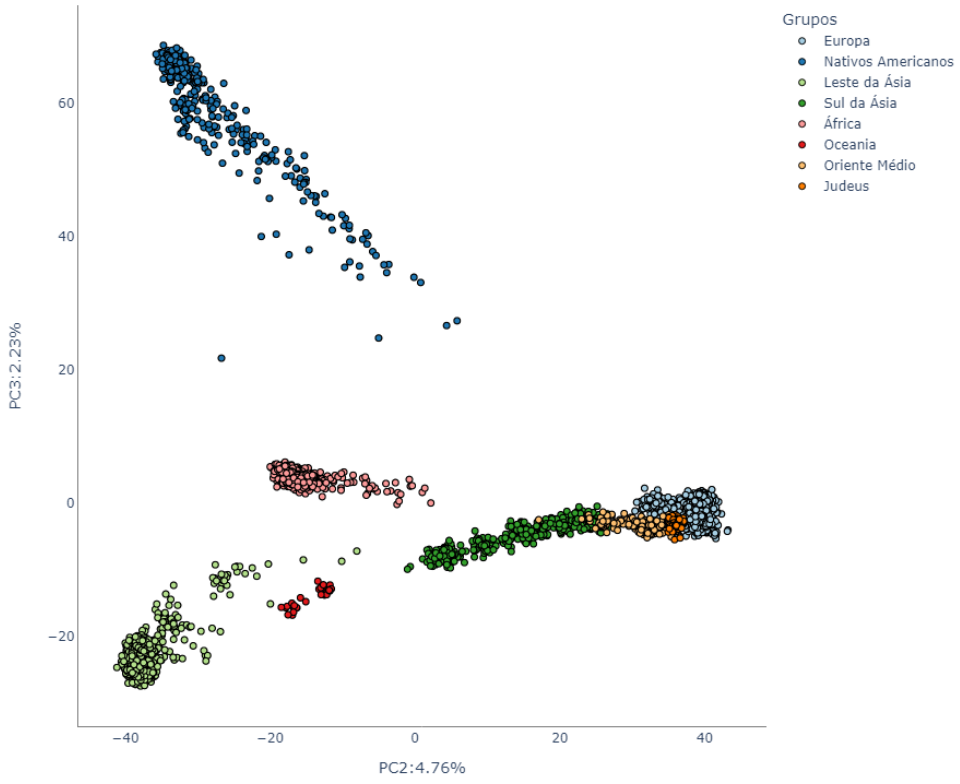
ZHENG, Xiuwen et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics, v. 28, n. 24, p. 3326-3328, 2012.

APÊNDICE

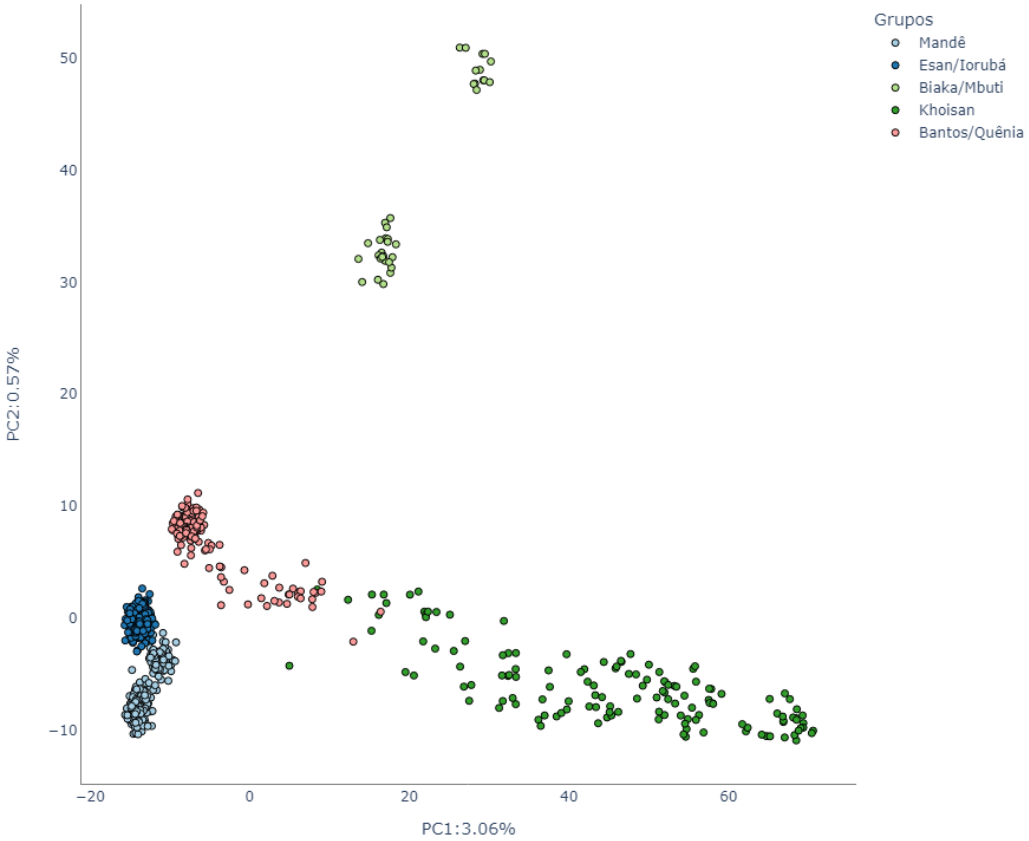
APÊNDICE A – PCA do conjunto de referência, PC1 x PC3



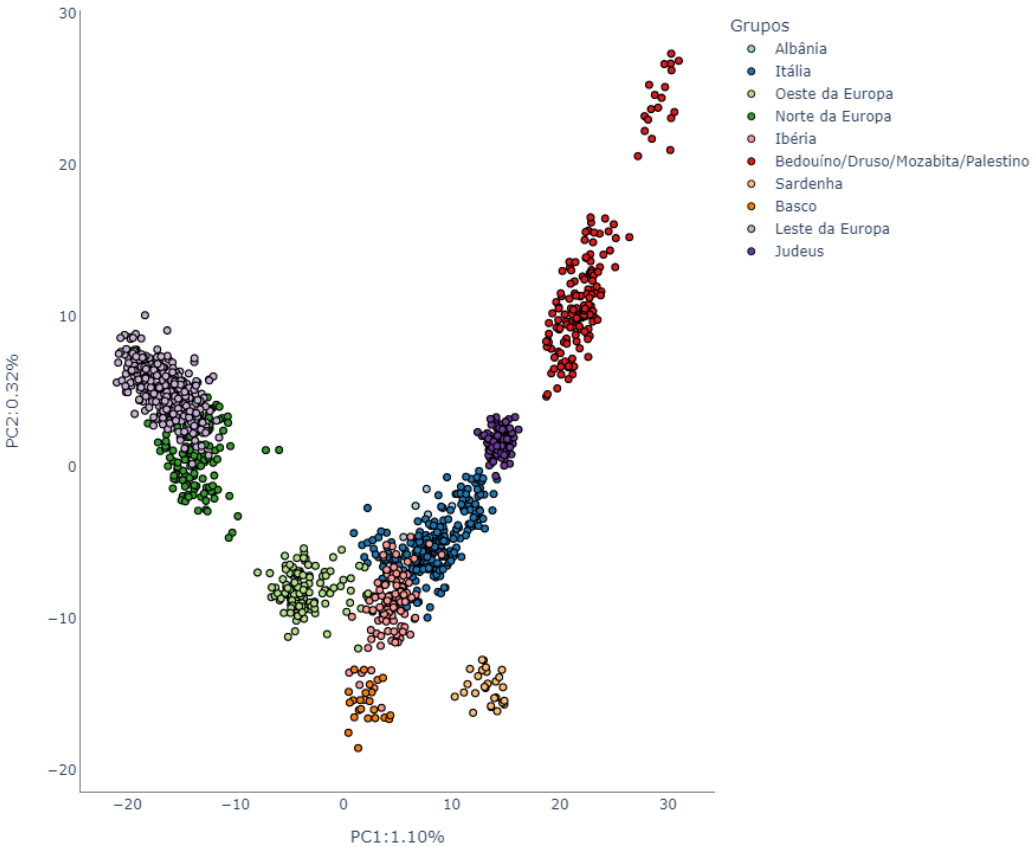
APÊNDICE B – PCA do conjunto de referência, PC2 x PC3



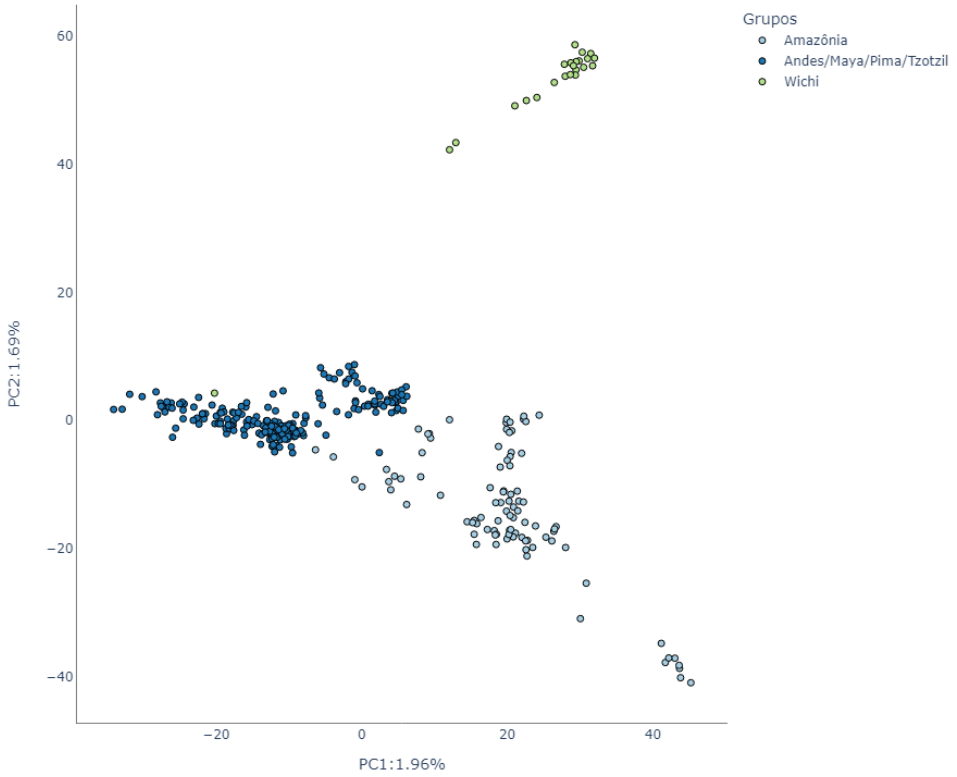
APÊNDICE C – PCA do conjunto de referência (África)



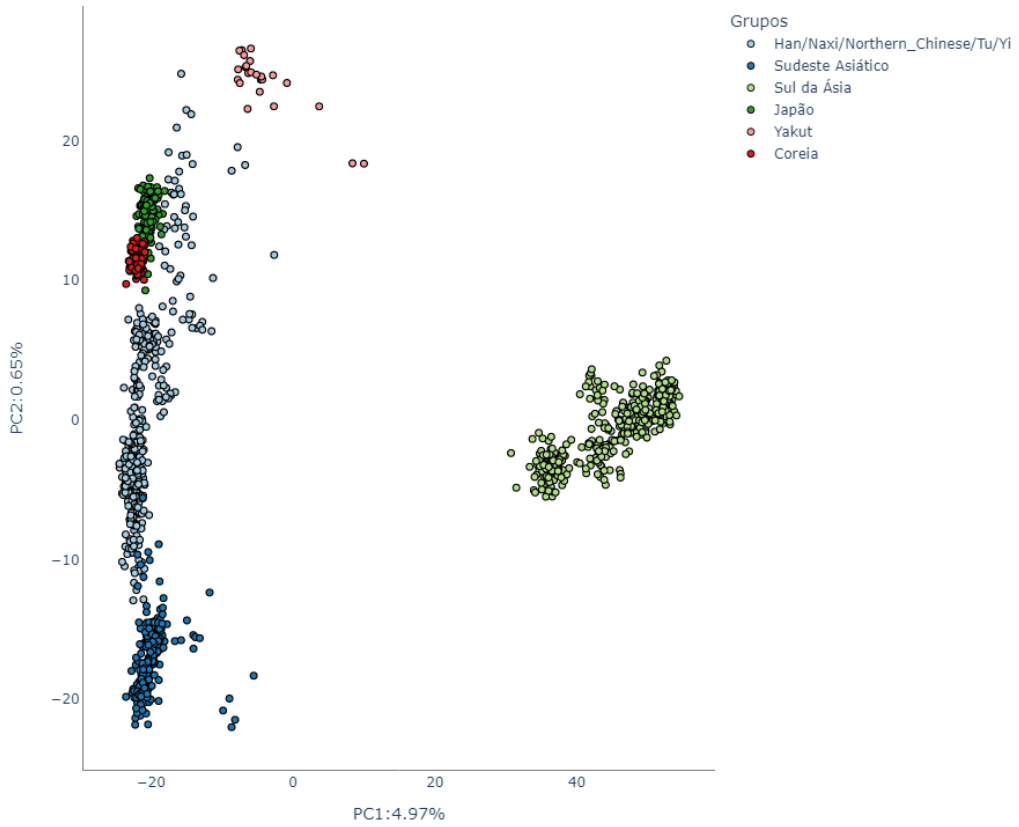
APÊNDICE D – PCA do conjunto de referência (Europa)



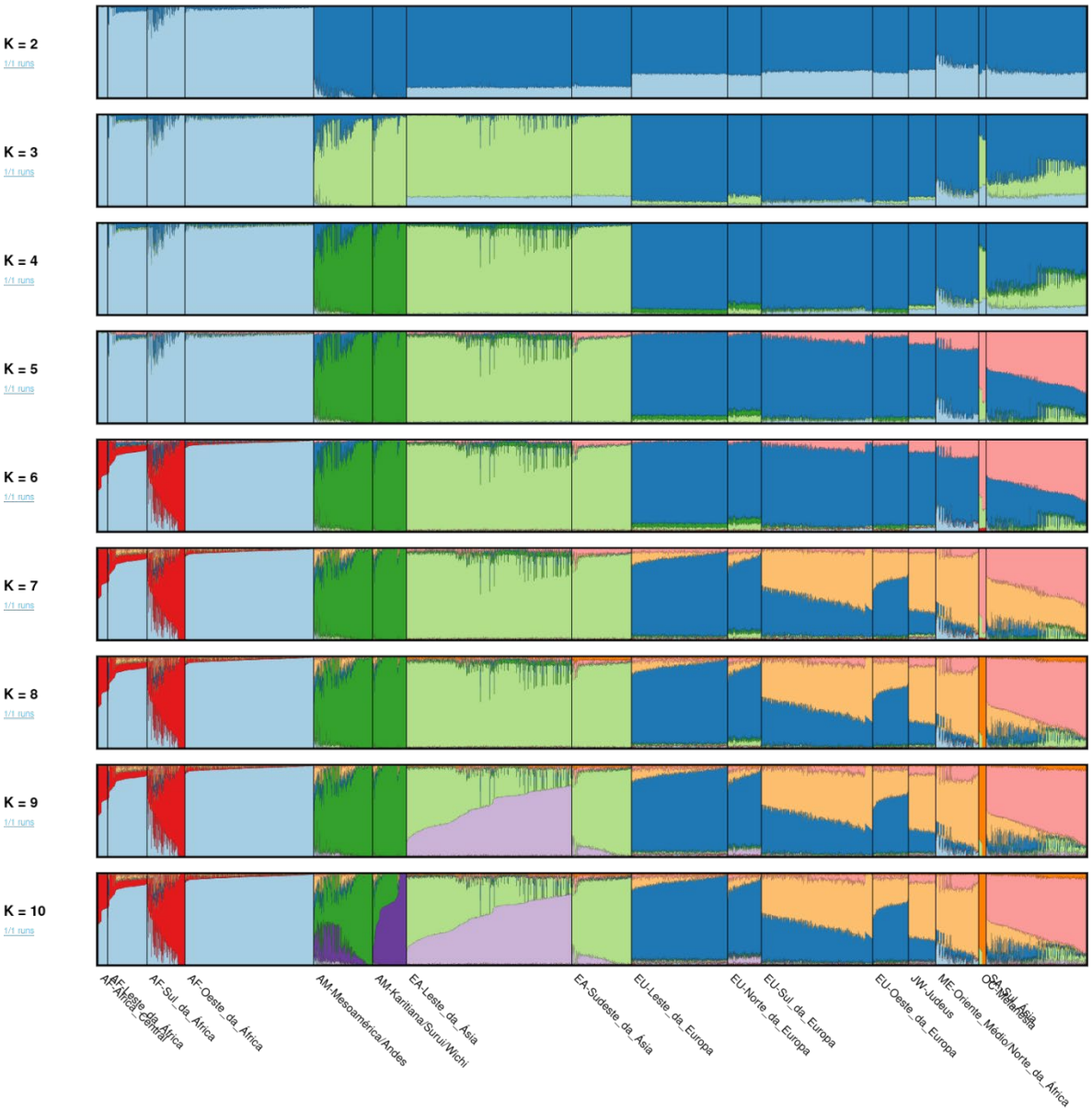
APÊNDICE E – PCA do conjunto de referência (Nativos Americanos)



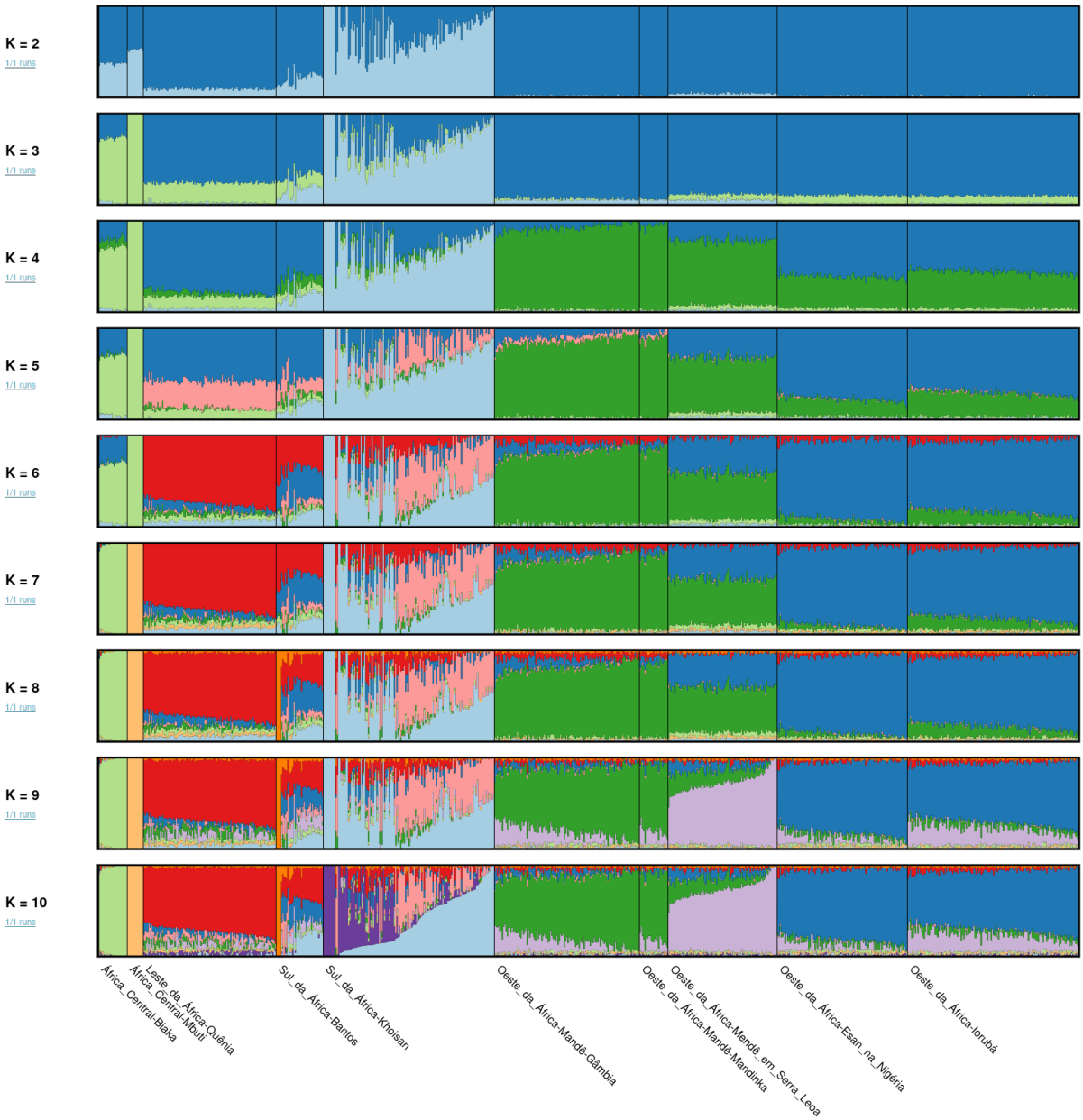
APÊNDICE F – PCA do conjunto de referência (Ásia)



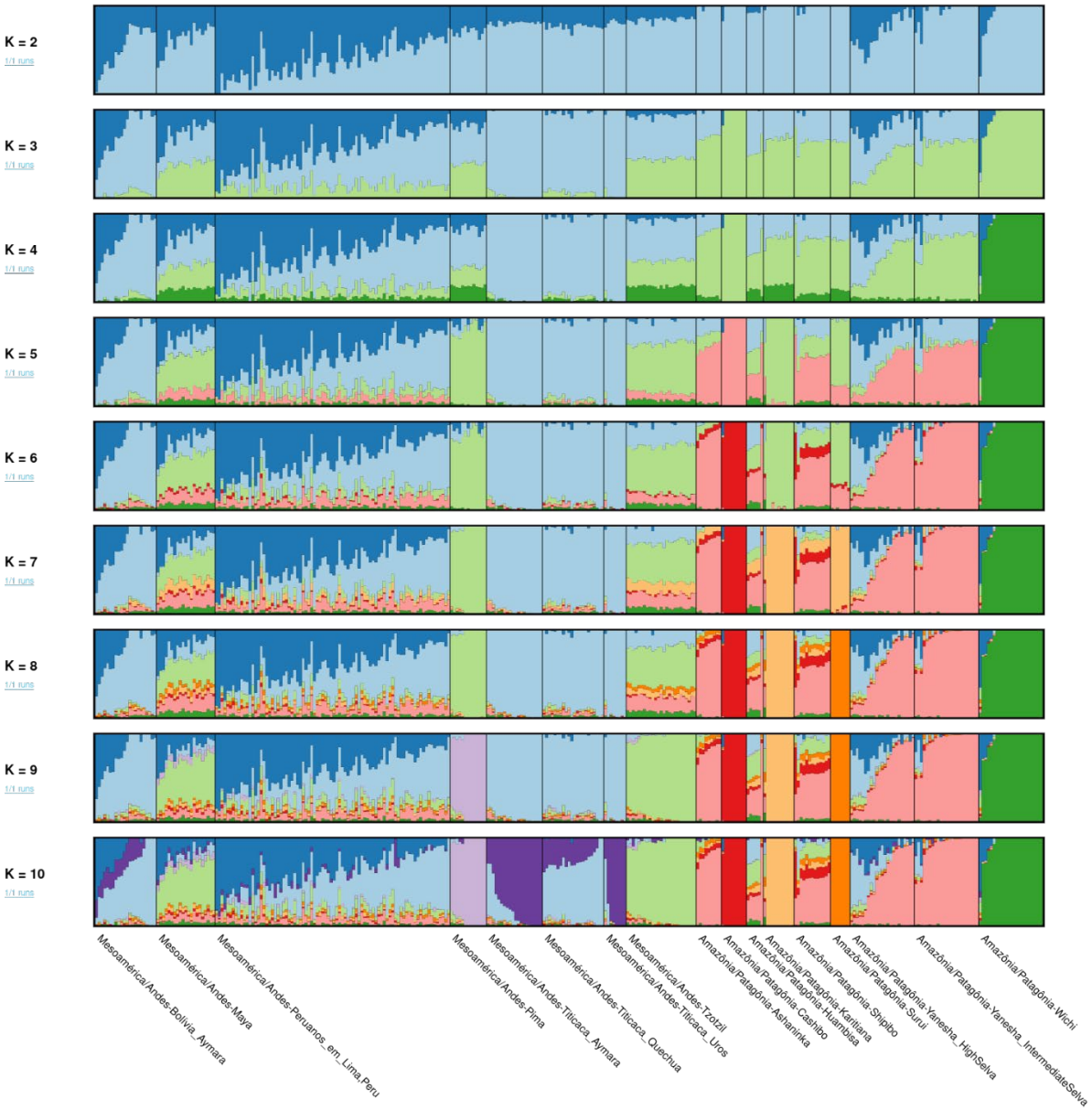
APÊNDICE G – Admixture com todos os grupos populacionais



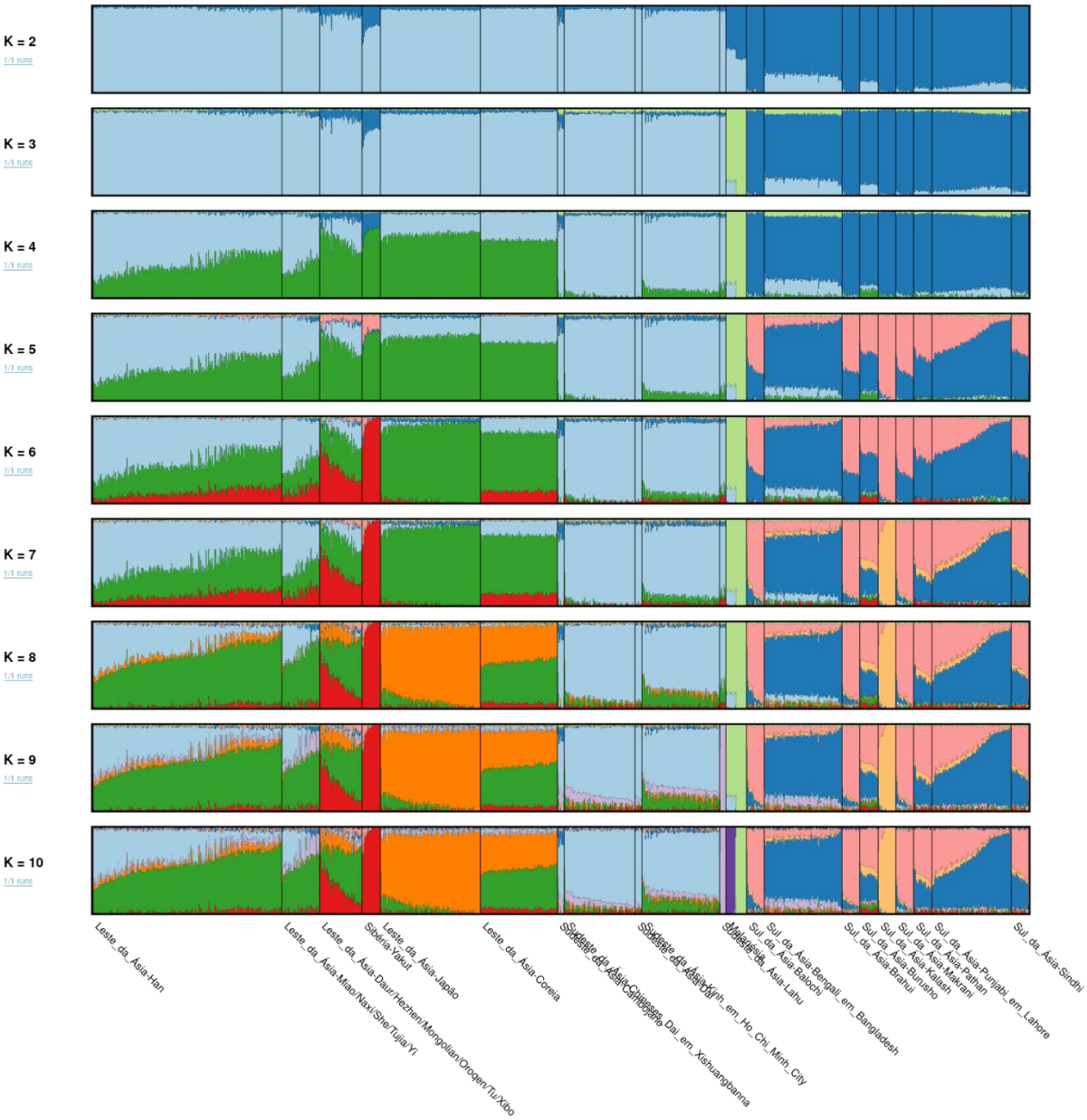
APÊNDICE H – Admixture com os grupos da África



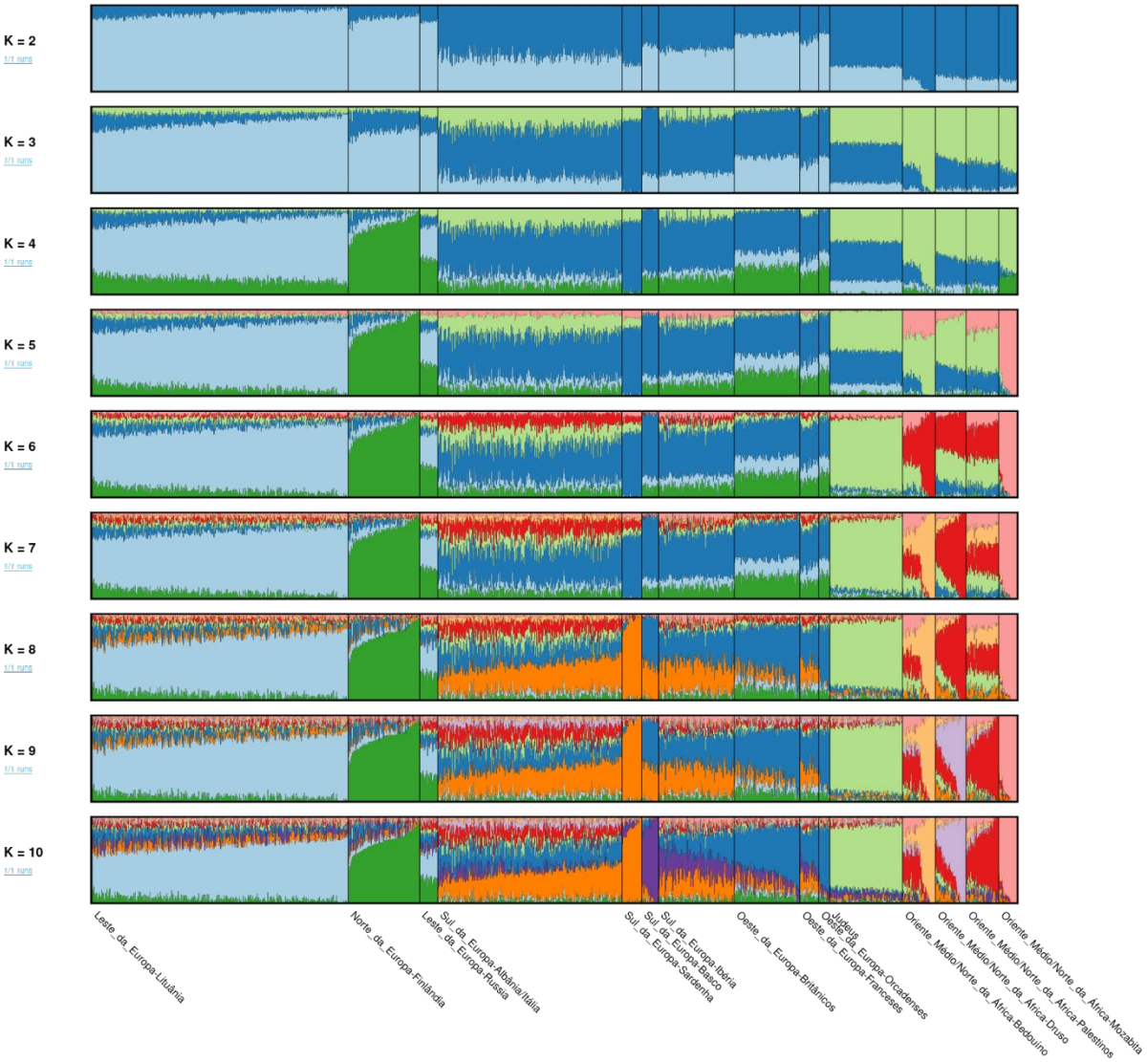
APÊNDICE I – Admixture com os grupos Nativo Americanos



APÊNDICE J – Admixture com os grupos da Ásia



APÊNDICE L – Admixture com os grupos da Europa



APÊNDICE M – RMSE dos modelos supervisionados

(continua)

	Modelos	RMSE (std)
1	VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=200)	0,040 (0,003)
2	VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=300)	0,041 (0,003)
3	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=200) N3 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=200)	0,041 (0,003)
4	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=300) N3 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=300)	0,042 (0,003)
5	VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=100)	0,042 (0,003)
6	RegMLEMix (λ=100)	0,042 (0,003)
7	RegMLEMix (λ=50)	0,043 (0,003)
8	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=100) N3 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=100)	0,043 (0,003)
9	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - RegMLEMix (λ=100) N3 - RegMLEMix (λ=100)	0,044 (0,002)
10	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - RegMLEMix (λ=50) N3 - RegMLEMix (λ=50)	0,044 (0,003)
11	VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=50)	0,045 (0,003)
12	RegMLEMix (λ=200)	0,046 (0,005)
13	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=50) N3 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=50)	0,046 (0,003)
14	RegMLEMix (λ=25)	0,046 (0,003)
15	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - RegMLEMix (λ=200) N3 - RegMLEMix (λ=200)	0,047 (0,005)
16	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=25) N3 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=25)	0,047 (0,003)
17	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - RegMLEMix (λ=25) N3 - RegMLEMix (λ=25)	0,047 (0,003)
18	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - RegMLEMix (λ=300) N3 - RegMLEMix (λ=300)	0,047 (0,004)

APÊNDICE M – RMSE dos modelos supervisionados

(conclusão)

	Modelos	RMSE (std)
19	VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=25)	0,048 (0,003)
20	RegMLEMix (λ=300)	0,050 (0,005)
21	RegMLEMix (λ=0)	0,051 (0,003)
22	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - RegMLEMix (λ=0) N3 - RegMLEMix (λ=0)	0,052 (0,003)

APÊNDICE N – Média dos valores de precisão e sensibilidade dos modelos

(continua)

	Modelos	Precisão (std)	Sensibilidade(std)
1	VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=200)	0,960(0,02)	0,943(0,02)
2	VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=300)	0,954(0,02)	0,933(0,02)
3	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=200) N3 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=200)	0,964(0,02)	0,945(0,02)
4	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=300) N3 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=300)	0,961(0,01)	0,945(0,01)
5	VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=100)	0,966(0,02)	0,946(0,02)
6	RegMLEMix (λ=100)	0,963(0,02)	0,939(0,02)
7	RegMLEMix (λ=50)	0,968(0,02)	0,940(0,02)
8	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=100) N3 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=100)	0,961(0,02)	0,933(0,03)
9	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - RegMLEMix (λ=100) N3 - RegMLEMix (λ=100)	0,956(0,02)	0,925(0,03)
10	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - RegMLEMix (λ=50) N3 - RegMLEMix (λ=50)	0,961(0,02)	0,924(0,03)
11	VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=50)	0,969(0,02)	0,947(0,02)
12	RegMLEMix (λ=200)	0,960(0,02)	0,941(0,03)
13	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=50) N3 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=50)	0,964(0,02)	0,932(0,03)
14	RegMLEMix (λ=25)	0,969(0,02)	0,944(0,02)
15	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - RegMLEMix (λ=200) N3 - RegMLEMix (λ=200)	0,965(0,02)	0,946(0,02)

APÊNDICE N – Média dos valores de precisão e sensibilidade dos modelos

(conclusão)

	Modelos	Precisão (std)	Sensibilidade(std)
16	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=25) N3 - VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=25)	0,965(0,02)	0,937(0,03)
17	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - RegMLEMix (λ=25) N3 - RegMLEMix (λ=25)	0,964(0,02)	0,933(0,03)
18	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - RegMLEMix (λ=300) N3 - RegMLEMix (λ=300)	0,968(0,01)	0,949(0,01)
19	VotingClassifier: RegMLEMix (λ=0) e RegMLEMix (λ=25)	0,971(0,02)	0,949(0,02)
20	RegMLEMix (λ=300)	0,954(0,02)	0,931(0,02)
21	RegMLEMix (λ=0)	0,971(0,02)	0,953(0,02)
22	HierarchicalLocalClassifier: N1 - RegMLEMix (λ=0) N2 - RegMLEMix (λ=0) N3 - RegMLEMix (λ=0)	0,964(0,02)	0,936(0,03)