UNIVERSIDADE FEDERAL DE MINAS GERAIS

Instituto de Ciências Biológicas

Programa Interunidades de Pós-graduação em Bioinformática

Lucas Faria Costa

ANCESTRALIDADE GENÔMICA E DINÂMICA DA MISCIGENAÇÃO:
estimativas e aspectos computacionais

Belo Horizonte

2023

Lucas Faria Costa

ANCESTRALIDADE GENÔMICA E DINÂMICA DA MISCIGENAÇÃO:
estimativas e aspectos computacionais

Dissertação apresentada ao Programa Interunidades de Pós Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Bioinformática

Orientador: Prof. Dr. Eduardo Martin Tarazona-Santos

Belo Horizonte
2023

043 Costa, Lucas Faria.
Ancestralidade genômica e dinâmica da miscigenação: estimativas e aspectos computacionais [manuscrito] / Lucas Faria Costa. – 2023.
76 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Eduardo Martin Tarazona-Santos.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Genética Populacional. 3. Miscigenação. 4. Ancestralidade comum. 5. Marcadores Genéticos. I. Tarazona-Santos, Eduardo Martin. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

LUCAS FARIA COSTA

Às quatorze horas do dia **25 de setembro de 2023**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Ancestralidade Genômica e Dinâmica da Miscigenação: Estimativas e Aspectos Computacionais"**, requisito para obtenção do grau de Mestre em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Eduardo Martin Tarazona Santos**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor(a)/Pesquisador(a)	**Instituição**	**Indicação**
Dr. Eduardo Martin Tarazona Santos	UFMG	**Aprovado**
Dr. Francisco Pereira Lobo	UFMG	**Aprovado**
Dr. Fabio Luiz Buranelo Toral	UFMG	**Aprovado**

Pelas indicações, o candidato foi considerado: **Aprovado**

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 25 de setembro de 2023.

Documento assinado eletronicamente por **Eduardo Martin Tarazona Santos**, **Professor do Magistério Superior**, em 26/09/2023, às 08:57, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fabio Luiz Buranelo Toral**, **Professor do Magistério Superior**, em 26/09/2023, às 09:24, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

AGRADECIMENTOS

Agradeço aos meus pais, Berenice e Rodrigo, por me darem a estrutura e apoio para chegar até a pós graduação. Minha mãe por sempre me auxiliar e me dar todo o suporte durante todo o meu período de estudo. Ao meu pai por ter me mostrado o mundo da ciência e questionamento. A minha irmã Manu pelas caminhadas na UFMG. A minha companheira Clara, que nos momentos mais difíceis esteve ao meu lado me doando seu tempo e atenção.

Agradeço também a todos os membros do Laboratório de Diversidade Genética Humana por terem me acolhido desde a Iniciação Científica e me ajudado sempre que precisei. Ao meu Orientador Eduardo, obrigado por sempre estar sempre presente e tornar a jornada mais leve. A Camila, pelo suporte durante os momentos mais complicados. A Carol pela paciência incalculável. Ao meu amigo Rafael pelos longos anos de amizade, camaradagem e por ter sido a ponte com o LDGH. Ao meu amigo Thiago por ter sempre me ajudado, pela confiança e paciência. Agradeço a Ferdi, Hanaisa, Bruno e Ricardo por me auxiliarem durante a produção dessa dissertação. Agradeço ao meu grande amigo Pena pelos mais de onze anos de incentivo na ciência.

Agradeço, em especial, a CAPES por permitir que eu me dedicasse exclusivamente ao estudo e aprimoramento científico. E obrigado PPG Bioinformática por manterem a excelência do programa.

RESUMO

Por conta da colonização europeia e o tráfico de pessoas escravizadas, a miscigenação nas Américas é uma característica marcante de suas populações que vai além da historicidade. Esse processo ficou também marcado nos genomas dos indivíduos americanos de forma que podemos recuperar informações sobre a história da miscigenação através do uso de dados genéticos. A ancestralidade genética pode ser inferida em três níveis: **populacional** (enfoque clássico), quando se infere a proporção de contribuição de cada população parental na formação da população miscigenada; **individual**, quando se infere a proporção de contribuição de cada população parental no genoma de um indivíduo, e **local ou cromossômica**, possível a partir da disponibilidade recente de dados de alta densidade genômica, quando se infere a ancestralidade de cada fragmento de cada cromossomo de um indivíduo miscigenado. Estatisticamente, se espera que o tamanho do fragmento de ancestralidade seja inversamente proporcional ao tempo em que o evento de miscigenação o inseriu naquela população: quanto mais antigo o evento, menor será o fragmento (e a recíproca verdadeira). Partindo desse conceito, nosso grupo desenvolveu um método baseado em Computação Bayesiana Aproximada para inferir as distribuições *a posteriori* de 9 parâmetros que representam as contribuições das populações ancestrais europeia, africana e nativa americana em 3 pulsos que ocorreriam ao longo dos 500 anos de miscigenação nas Américas (Kehdy *et al.* 2015). Esse projeto tem como objetivo inferir e analisar a ancestralidade de indivíduos latino-americanos no contexto de projetos em colaboração juntamente com a validação do método desenvolvido pelo nosso grupo que busca inferir a dinâmica do processo de miscigenação através dos estudos genéticos populacionais remontando a história genética de como as interações entre as populações parentais ocorreram e sua validação.

Palavras-chave: Bioinformática; Genética de populações; miscigenação; dinâmica de miscigenação; ancestralidade; marcadores genéticos; modelagem; populações latinas; pipelines bioinfomáticos; estruturação populacional; simulações.

ABSTRACT

Due to European colonization and the trafficking of enslaved people, admixture in the Americas is a defining characteristic of its populations that extends beyond historical context. This process is also imprinted in the genomes of American individuals, allowing us to retrieve information about the history of admixture through genetic data. Genetic ancestry can be inferred at three levels: population-level (classical approach), where the proportion of contribution from each parental population to the admixed population is estimated; individual-level, where the proportion of each parental population's contribution to an individual's genome is inferred; and local or chromosomal-level, made possible by the recent availability of high-density genomic data, where the ancestry of each fragment of each chromosome in an admixed individual can be determined. Statistically, the size of an ancestry segment is expected to be inversely proportional to the time elapsed since the admixture event introduced it into the population: the older the event, the smaller the fragment (and vice versa). Based on this concept, our group developed a method using Approximate Bayesian Computation to infer the posterior distributions of nine parameters representing the contributions of European, African, and Native American ancestral populations in three admixture pulses that occurred over 500 years of admixture in the Americas (Kehdy et al. 2015). This project aims to infer and analyze the ancestry of Latin American individuals in the context of collaborative projects, along with validating the method developed by our group, which seeks to infer the dynamics of the admixture process through population genetics studies, reconstructing the genetic history of how interactions between parental populations occurred and its validation

Key words: Bioinformatics; Population genetics; Admixture; admixture dynamics; ancestry; genetic markers; modeling; Latin populations; bioinformatics pipelines; population structuring; simulations.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES E SIGLAS

ABC	Approximate Bayesian Computation (Computação Bayesiana Aproximada)
ACB	African Caribbean in Barbados (Afro-caribenhos em Barbados)
AFR	Africanos
ASW	Ancestralidade Africana no Sudoeste dos Estados Unidos da América
AWS	Amazon Web Services
CDX	Chinese Dai in Xishuangbanna, China (Chinese Dai em Xishuangbanna)
CEU	Northern Europeans from Utah (Europeus do norte de Utah)
CHB	Han Chinese in Beijing, China (Han Chines em Pequim)
CHS	Han Chinese South (Han Chinês do Sul)
CLM	Colombian in Medellín (Colombianos em Medellín)
CV	Cross-Validation (Validação Cruzada)
DBSNP	Single Nucleotide Polymorphism Database
DNA	Ácido desoxirribonucleico
EM	Expectation-Maximization (Expectativa de Maximização)
ESN	Esan in Nigeria (Esan na Nigéria)
EUR	Europeus
FIOCRUZ RJ	Fundação Oswaldo Cruz - Rio de Janeiro
GBR	British From England and Scotland (Britânicos da Inglaterra e Escócia)
GWAS	Genome-wide association studies (Estudo de associação genômica ampla)
GWD	Gambian in Western Division (Gâmbia na Divisão Ocidental)

HGDP	Human Genome Diversity Project (Projeto de Diversidade do Genoma Humano)
HPD	High Posteriori Density (Densidade a posteriori máxima)
HWE	Hardy-Weinberg Equilibrium (Equilíbrio de Hardy-Weinberg)
IBD	Identical by descent (Idêntico por descendência)
IBS	Iberian Populations in Spain (Ibéricos na Espanha)
IMC	Índice de Massa Corporal
JPT	Japanese in Tokyo, Japan (Japonês em Tóquio)
LD	Linkage Disequilibrium (Desequilíbrio de Ligação)
LDGH	Laboratório de Diversidade Genética Humana
LOCA	Local Ancestry (Ancestralidade Local)
LU3	Low Uniform Admixture from 3 Populations (Miscigenação Baixa e Uniforme de 3 Populações)
LWK	Luhya in Webuye, Kenya (Luhya em Webuye Quênia)
MCMC	Monte Carlo Markov Chain (Cadeia de Markov Monte Carlo)
MSL	Mende in Sierra Leone (Mende em Serra Leoa)
MXL	Mexican Ancestry in Los Angeles (Ancestralidade Mexicana em Los Angeles)
NAT	Nativo-Americanos
NAToRA	Network Algorithm to Relatedness Analysis (Algoritmo de rede para Análise de Parentesco)
NGS	Next Generation Sequencing (Sequenciamento de nova geração)
PA1	Predominant Ancient Admixture from 1 Population (Miscigenação Predominantemente Antiga de uma população)
PCA	Principal Component Analysis (Análise de Componente Principais)
PEL	Peruvian in Lima (Peruanos em Lima)
PI1	Predominant Intermediate Admixture from 1 Population (Miscigenação Predominantemente Intermediária de uma população)
POD	Pseudo Observed Data (Dados Pseudo-Observados)
PR1	Predominant Intermediate Admixture from 1 Population (Miscigenação Predominantemente Recente de 1 População)
PRS	Polygenic risk scores (Escore de Risco Poligênico)
PUR	Porto Rico
RAM	Random-access memory (Memória de acesso aleatório)
REAP	Relatedness Estimation in Admixed Populations (Estimação de Parentesco em Populações Miscigenadas)

SNP	Single Nucleotide Polymorphism (Polimorfismo de Nucleotídeo Único)
SNV	Single Nucleotide Variant (Variante de Nucleotídeo Único)
TSI	Toscani in Italia (Toscanos na Itália)
YRI	Yoruba in Ibadan (Yoruba em Ibadan)

SUMÁRIO

CAPÍTULO 1. ANCESTRALIDADE GENÔMICA EM POPULAÇÕES DA AMÉRICA LATINA

1.1 Introdução

Definimos populações miscigenadas aquelas que são fruto da reprodução entre indivíduos de duas ou mais populações parentais isoladas e diferenciadas entre elas. Os latino-americanos formam uma população de origem tri-híbrida, isto é, seu material genético, majoritariamente, advém da miscigenação entre Africanos, Europeus e Nativos-Americanos. Dados genéticos permitem inferir as ancestralidades em três níveis: populacional, quando se infere a proporção da contribuição de cada população parental na formação da população mista; individual, quando se infere a proporção da contribuição de cada população parental no genoma de um indivíduo; e local cromossômica (atualmente viável com dados de alta densidade), consiste em descobrir de qual população ancestral deriva cada região cromossômica de cada cromossomo de um indivíduo. Essas regiões contínuas de uma única ancestralidade são chamadas de *tracts*.

Partindo de parâmetros genéticos populacionais clássicos (populações parentais, miscigenadas, frequências alélicas, número efetivo populacional etc) e através de marcadores mais ou menos informativos de ancestralidade que são encontrados no genoma humano, podemos estimar as frações que cada população ancestral contribuiu em um processo de miscigenação em cada população, em cada indivíduo ou cromossomo (Figura 1).

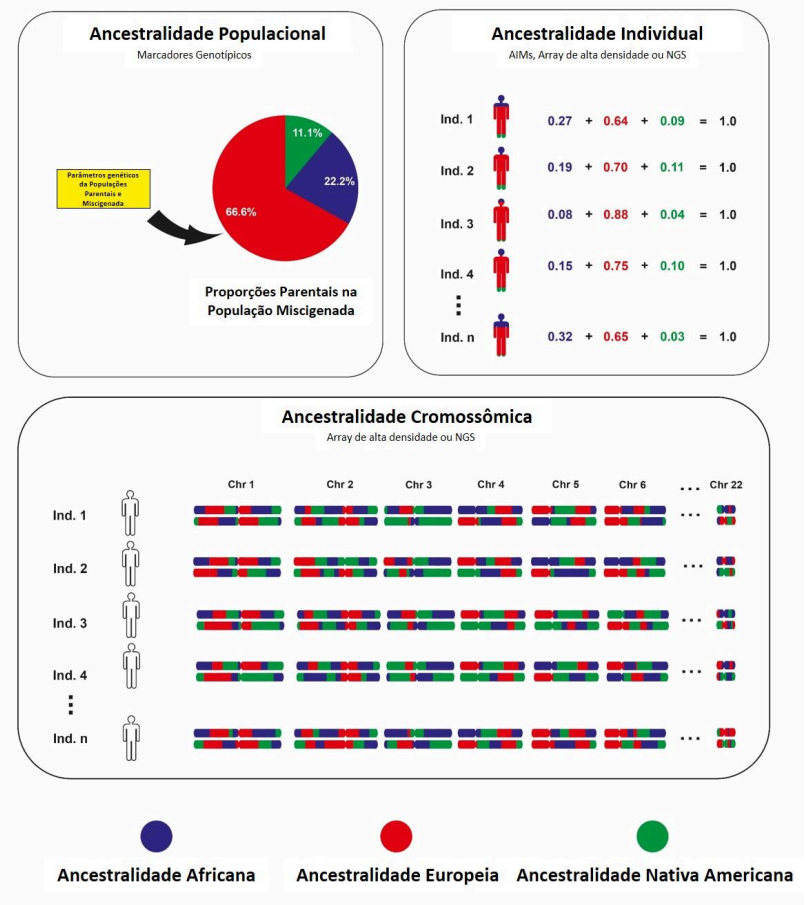

Figura 1. Esquema dos níveis de ancestralidade. A ancestralidade populacional é a fração que cada população parental contribui em uma população miscigenada. A ancestralidade individual estima a fração de cada população parental do indivíduo. E a ancestralidade cromossômica estima a população parental de origem em cada segmento cromossômico de cada indivíduo. As cores representam as populações parentais. Azul - Africana; Vermelho - Europeia ; Verde - Nativo Americana. Fonte: Soares-Souza *et al*., 2018.

Remontar a história de populações miscigenadas como as da América Latina faz-se necessário, pois além de serem negligenciadas em estudos genéticos, assim como as Africanas (Sirugo *et al*. 2019, Popejoy *et al.* 2016), trazem à tona o passado violento e estratificado como observado em Kehdy *et al*. 2015 ao evidenciar geneticamente o conhecido viés sexual, observando diferenças de proporções parentais ao comparar o cromossomo X com autossômicos o que evidencia a assinatura no genoma de acasalamentos entre homens predominantemente europeus com mulheres com uma maior ancestralidade africana ou indígena.

Ademais, o papel do estudo de ancestralidade está associado a um melhor entendimento das bases genéticas de doenças. Por exemplo, Scliar *et al*. 2021, encontraram uma variante (rs114066381-A) que está associada ao aumento do índice de massa corporal (IMC) em mulheres miscigenadas que a possuem, e que se sabe que tem origem africana porque foi

encontrada em fragmentos genômicos (*tracts*) de origem africana. Outro exemplo da relação entre ancestralidade e saúde é no cálculo de *polygenic risk scores* (PRS). O PRS usa informações genômicas para avaliar as chances de uma pessoa ter ou desenvolver uma condição médica específica através de um cálculo estatístico baseado na presença ou ausência de múltiplas variantes genômicas, sem levar em conta fatores ambientais ou outros.

Alguns métodos inferem a ancestralidade a partir de dados genéticos usando genótipos independentes entre si. Esses métodos usam informações de marcadores genéticos para inferir a ancestralidade de um indivíduo ou população. Nesse contexto, o ADMIXTURE (Alexander, Novembre and Lange 2009) (Figura 2). Trata-se de uma análise (no nosso caso, não supervisionada, ou seja, sem identificação das populações parentais como referência) de agrupamento, que infere a proporção de contribuição de cada população parental K no genoma de cada indivíduo, bem como as frequências alélicas de cada uma das K populações, ajustando o equilíbrio de Hardy-Weinberg em cada uma das K populações/*clusters* (Borda *et al.* 2020). O número máximo de populações parentais é dado pelo usuário, para cada *dataset*. Ainda entre os métodos que usam genótipos independentes, a análise de componentes principais (PCA) é uma técnica estatística que reduz a dimensionalidade das informações obtidas através dos genótipos e permite visualizar a estrutura populacional em um espaço de baixa dimensão (Figura 2). O PCA é baseado na decomposição da matriz de covariância dos dados genéticos em seus componentes principais. Esses componentes são vetores que representam as direções de maior variação dos dados genéticos, no qual cada componente principal é uma combinação linear do número de cópias de um alelo em um SNV (0, 1 ou 2) e através disso projeta em eixos ortogonais a variação dos dados genéticos dos indivíduos estudados.

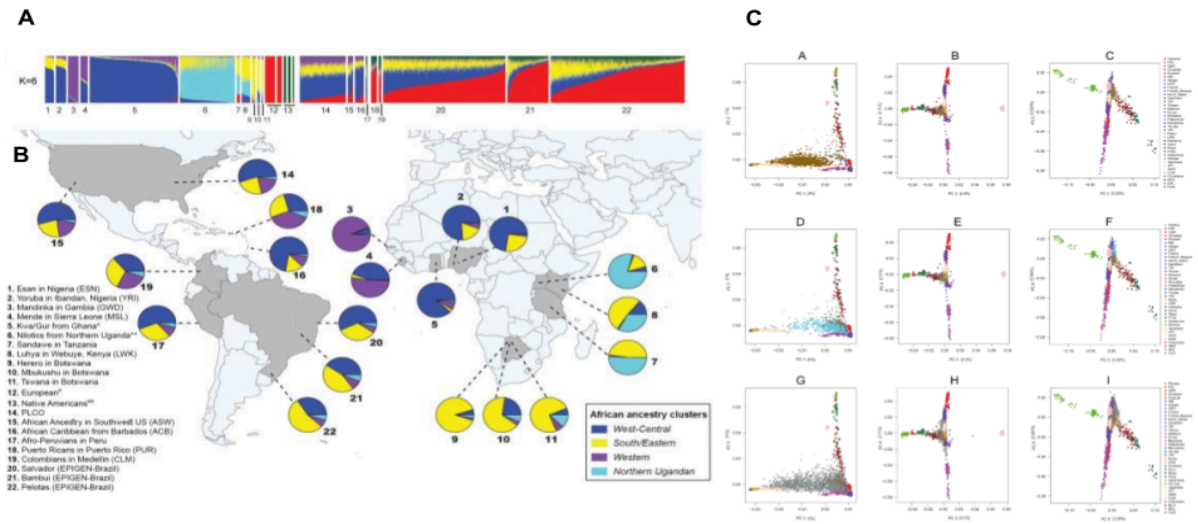

Figura 2. Análise de ancestralidade baseada em dados de genótipos independentes. A Plot de análise de ancestralidade de populações africanas e miscigenadas utilizando o *software* ADMIXTURE (Figura adaptada de Gouveia *et al.* 2020). **B** médias de ancestralidade intracontinental africana por população para populações africanas e miscigenadas (Figura adaptada de Gouveia *et al.* 2020). **C** Plot de Análise de Componentes Principais (PCA) para diversos componentes principais (PC's) através do *software* EIGENSOFT (Patterson *et al.* 2006 e Price *et al.* 2006) de coortes brasileiras do estudo EPIGEN-Brasil (Figura adaptada de Kehdy *et al*. 2015).

Além dos dados de genótipos, é possível utilizar os dados de haplótipos para a inferência de ancestralidade. Métodos baseados em haplótipos utilizam informações genéticas compartilhadas entre indivíduos de uma população e populações de referência (parentais) para estimar a ancestralidade. Através da comparação dos padrões de variação genética, é possível identificar regiões compartilhadas e aplicar modelos estatísticos para quantificar as contribuições ancestrais.

O RFMix (Maples *et al.* 2013) utiliza um modelo de mistura para inferir a ancestralidade local cromossômica de cada haplótipo em relação a um conjunto de populações de referência. A ancestralidade cromossômica local é estimada através das informações dos haplótipos usando um modelo probabilístico que usa variáveis observadas (alelos) de cromossomos parentais e mistos, bem como um mapa de recombinação, para inferir variáveis não observadas (ancestralidade de cada SNV) ao longo dos cromossomos. Formalmente, o RFMix usa uma abordagem discriminativa, o que significa que ele modela P(Y|X), onde Y é a variável não observada de interesse (ancestralidade) e X variáveis observadas que medimos para nos ajudar a inferir Y (ou seja, alelos, haplótipos e um mapa de recombinação). Ao executar o método iterativamente, o RFMix é capaz de aprender com as amostras misturadas para melhorar o desempenho e corrigir automaticamente os erros de faseamento. O *software* CHROMOPAINTER (Lawson *et al.* 2012), por sua vez, utiliza um modelo de coalescência

para simular a história evolutiva das populações e inferir a ancestralidade de cada haplótipo. GLOBETROTTER (Hellenthal *et al*. 2014) é um programa que pode identificar, datar e descrever eventos de mistura que ocorrem na história ancestral de uma determinada população-alvo nos últimos ~4.500 anos (Figura 3). O fineSTRUCTURE (Lawson *et al*. 2012) é um algoritmo rápido e poderoso para identificar a estrutura da população utilizando as saídas do CHROMOPAINTER e utilizando de modelos de inferência bayesiana para inferir a ancestralidade de cada haplótipo em relação a um conjunto (Figura 3).

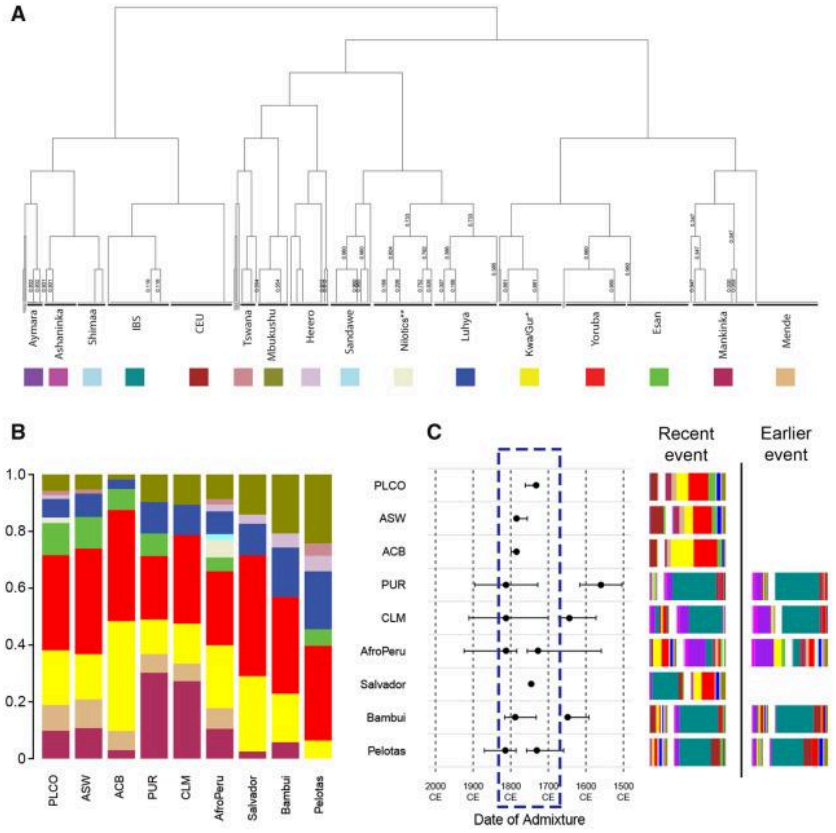

Figura 3. Análise de ancestralidade baseada em dados de haplótipo. A árvore fineSTRUCTURE de indivíduos parentais. **B** Contribuições subcontinentais relativas à ascendência africana total em populações. **C** Inferência GLOBETROTTER. Figuras de Gouveia *et al*. 2020.

A inferência de parentesco é elemento do universo de ancestralidade. Através do *software* RefinedIBD (Browning *et al.* 2013) é possível inferir segmentos idênticos por descendência, IBD, (Figura 4).

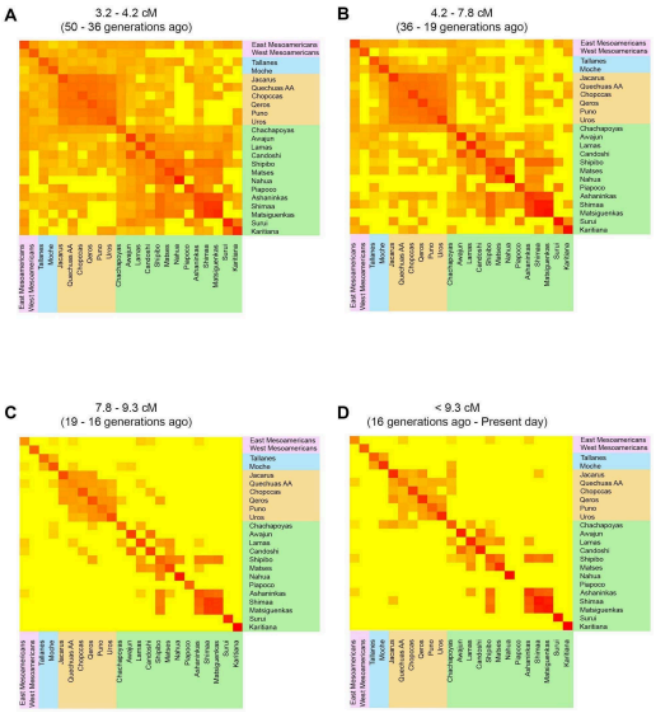

Figura 4. Representação de mapa de calor dos segmentos IBD Cada mapa de calor representa um intervalo de segmentos, em centimorgan (cM), correlacionado com a geração, do ancestral comum mais antigo **A**) e ancestral comum mais recente (**A**) Figura adaptada de Borda *et al.* 2020.

Através dos métodos citados é possível compreender grande parte da estrutura populacional de uma população alvo. Trabalhar de forma interativa entre os resultados e seguir um bom fluxograma de trabalho permite que inferências e observações sejam feitas gerando resultados com boa qualidade, que acabam se complementando (Figura 5).

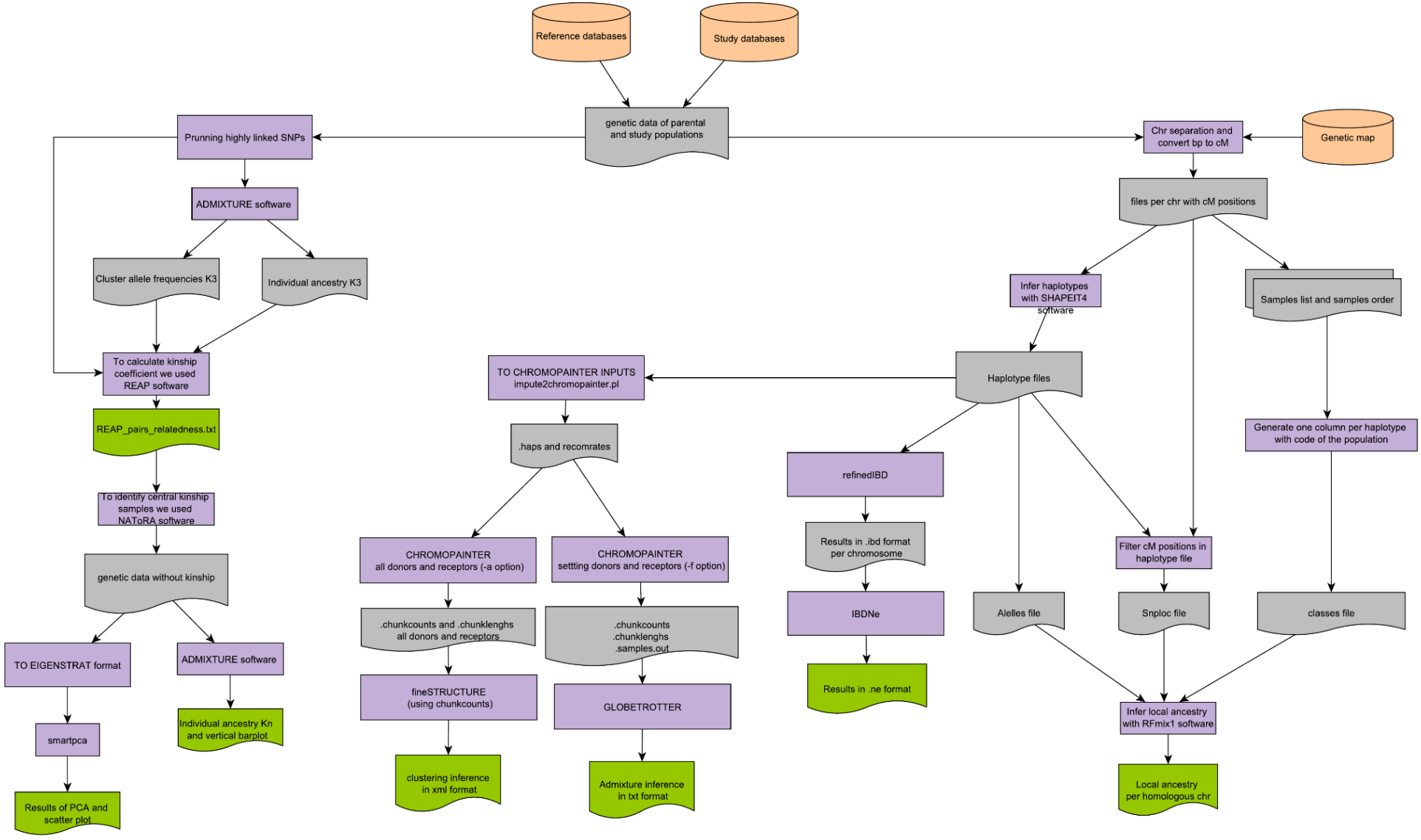
Reference databases
Study databases
genetic data of parental and study populations
Prunning highly linked SNPs
ADMIXTURE software
Cluster allele frequencies K3
Individual ancestry K3
To calculate kinship coefficient we used REAP software
REAP_pairs_relatedness.txt
To identify central kinship samples we used NATcRA software
genetic data without kinship
TO EIGENSTRAT format
smartpca
Results of PCA and scatter plot
ADMIXTURE software
Individual ancestry Kn and vertical barplot
Chr separation and convert bp to cM
Genetic map
files per chr with cM positions
Infer haplotypes with SHAPEIT4 software
Samples list and samples order
Haplotype files
TO CHROMOPAINTER INPUTS impute2chromopainter.pl
.haps and recomrates
CHROMOPAINTER all donors and receptors (-a option)
.chunkcounts and .chunklenghs all donors and receptors
fineSTRUCTURE (using chunkcounts)
clustering inference in xml format
CHROMOPAINTER settting donors and receptors (-f option)
.chunkcounts .chunklenghs .samples.out
GLOBETROTTER
Admixture inference in txt format
refinedIBD
Results in .ibd format per chromosome
IBDNe
Results in .ne format
Alelles file
Filter cM positions in haplotype file
Snploc file
Generate one column per haplotype with code of the population
classes file
Infer local ancestry with RFmix1 software
Local ancestry per homologous chr

Figura 5. Fluxograma para as análises de ancestralidade. Fluxograma para análises de ancestralidade (dados de genótipos e haplótipos) e parentesco (dados de segmentos idênticos por descendência) utilizado pelo LDGH. O acesso ao fluxograma está disponibilizado em https://github.com/ldgh/3A-public**.**

1.1.1 O estudo da ancestralidade no contexto de estudos genéticos em populações latinoamericanas.

Na presente dissertação trabalhamos em três projetos de pesquisa específicos nos quais a inferência da ancestralidade é relevante: (i) O estudo dos padrões de expressão gênica em populações indígenas peruanas dos Andes; (ii) um Estudo de associação genética em larga escala (*Genome-wide Association Study,* GWAS) com COVID-19 da Fundação Oswaldo Cruz e (iii) a iniciativa *Mosaico Translational Genomics.*

1.1.1.1. Os dados de Huánuco, indígenas do Peru, fazem parte de um estudo colaborativo com o Dr. Heinner Guio do centro de pesquisa Inbiomedic, que tem como objetivo estudar o padrão de expressão gênica (transcriptoma) em células periféricas mononucleares do sangue (PBMC) em resposta a estímulos de lipopolissacarídeos de patógenos. Os indivíduos de Huánuco foram genotipadas com o Array Illumina GSA MG v3 na *facility* da Macrogen (Seul, Coréia). Após o processamento do Illumina Genome Studio, obtivemos 737534 SNVs para um total de 44 indivíduos.

1.1.1.2. Os dados FIOCRUZ COVID-19 fazem parte do estudo de GWAS Brasileiro com COVID-19, em parceria com as doutoras Cristiana Couto Garcia e Fernanda de Souza Gomes Kehdy, do Instituto Oswaldo Cruz (IOC/RJ) e Centro de Pesquisa René Rachou (CpRR) (FIOCRUZ), respectivamente. Os dados utilizados são oriundos de 1126 amostras coletadas nas cinco regiões brasileiras através de genotipagem realizada com o array Axiom Human Genotyping SARS-CoV-2 Research Array para 820000 marcadores.

1.1.1.3. Os dados da iniciativa *Mosaico Translational Genomics* são clientes do projeto de extensão no contexto do desenvolvimento do produto *My Mosaic Ancestry*, que leva as análises de ancestralidade diretamente ao consumidor final. Esses dados foram obtidos através da Infinium Global Screening Array (GSA), para 654027 marcadores de um total de 10 indivíduos.

1.2 Objetivos

1.2.1. Organizar e manter os pipelines bioinformáticos do Laboratório de Diversidade Genética Humana para inferir a ancestralidade genômica.

1.2.2. Inferir e analisar a ancestralidade de indivíduos latino-americanos no contexto dos três projetos do LDGH mencionados.

1.3 Metodologia

1.3.1 Controle de Qualidade

Assim como os dados de referência, os dados das populações alvo passaram pelo *software* *MosaiQC* (https://github.com/ldgh/Smart-cleaning-public/), *software* desenvolvido pelo Dr. Thiago Peixoto Leal para dados de array de SNVs em larga escala. *MosaiQC* automatiza o pipeline de controle de qualidade e limpeza de dados usado no Laboratório de Diversidade Genética Humana, desenvolvido por diferentes membros do grupo durante os últimos anos. A limpeza dos dados remove informaçoes de indivíduos de sexo incerto, variantes que apresentam o cromossomo 0 (informação colocada para posições que não foram alinhadas corretamente, não hibridizaram corretamente ou não tem informação), posição genética duplicada, variantes que não tenham sido genotipadas em mais de 10% dos indivíduos, indivíduos com mais de 10% das variantes não genotipadas, variantes que cujos alelos sejam complementares (A/ T ou C/G).

1.3.2 Inferência de Ancestralidade

Para reduzir a dimensionalidade dos dados observados e compreender a disposição dos indivíduos entre si, utilizamos da Análise de Componentes Principais (PCA) a partir dos genótipos individuais. Através do *software* EIGENSOFT (Patterson *et al.* 2006 e Price *et al.* 2006) realizamos PCA utilizando os genótipos de cada indivíduo para evidenciar a distância genética entre os indivíduos de diferentes populações.

Para análise e inferência de ancestralidade de Huánuco, FIOCRUZ COVID-19 e Mosaico, tanto a nível de individual quanto populacional, foi usado o software ADMIXTURE (Alexander, Novembre and Lange 2009). Para garantir a independência dos SNVs, como requerido pelo ADMIXTURE, esse programa foi executado após o filtro de desequilíbrio de ligação que excluiu todos os SNVs com $r^2>0.4$ (referência de r^2). Para o cálculo das proporções parentais dentro das populações miscigenadas foi feita a média da proporção de contribuição de cada população parental dentro de cada população miscigenada. A partir disso, realizamos 14 corridas (independentes) de ADMIXTURE para Huánuco, 8 corridas (independentes) para FIOCRUZ COVID-19 e 2 corridas (independentes) de ADMIXTURE para Mosaico (Tabela 1). Ao executar diversas corridas para um único K podemos checar no

output o *log-likelihood* (verossimilhança logarítmica). O valor logarítmico de um modelo de regressão é uma forma de medir a qualidade do ajuste de um modelo. Quanto maior o valor da verossimilhança logarítmica melhor será aquela corrida para aquele conjunto de dados. Por fim, dentre os melhores resultados selecionados para cada K observamos qual se adequa melhor com nosso modelo. A validação cruzada (CV) é uma técnica estatística usada para avaliar o desempenho e a capacidade de generalização de um modelo. O ADMIXTURE não tenta estimar a evidência do modelo, em vez disso, ele inclui um procedimento de CV que permite ao usuário identificar o valor de K qual modelo tem a melhor precisão preditiva.

Tabela 1. K inicial, K final e números de replicatas do ADMIXTURE para as populações alvo.

Estudo	K inicial	K final	Nº de replicatas
Huánuco	3	12	14
FIOCRUZ COVID-19	3	12	8
Mosaico	3	12	2

As inferências de ancestralidade cromossômica local foram feitas a partir do *software* RFMix (Maples *et al*. 2013). A execução das análises de PCA, ADMIXTURE, RFMix foram realizadas através da ferramenta MosA3ic (Figura 6) desenvolvida pelo Laboratório de Diversidade Genética Humana (LDGH) que automatiza a execução de PCA, ADMIXTURE e RFMix. O pipeline de ancestralidade está organizado no *GitHub* (https://github.com/ldgh/3A-public) do LDGH, pelo qual sou responsável (Figura 6).

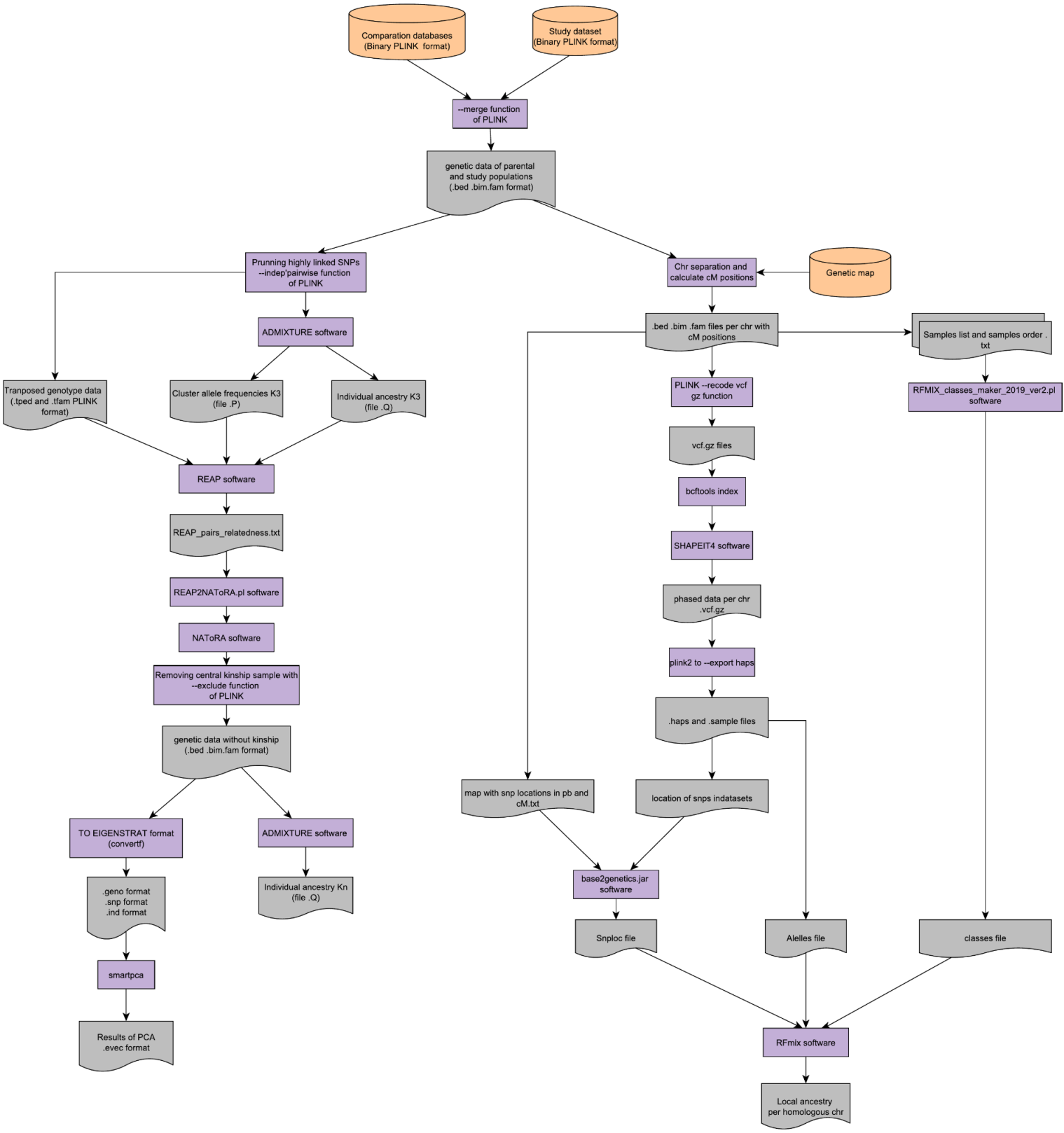

Figura 6. Fluxograma público utilizado para as análises de ancestralidade. Fluxograma baseado na ferramenta MosA3ic, desenvolvida pelo LDGH para automatizar a integração de PCA, ADMIXTURE e RFMix.

Para a visualização dos dados de PCA e ADMIXTURE, automatizamos a geração dos resultados através de um *pipeline* que utiliza *scripts* em Python 3.10, com a colaboração do aluno de Iniciação Científica Luca Gama (Figura 7).

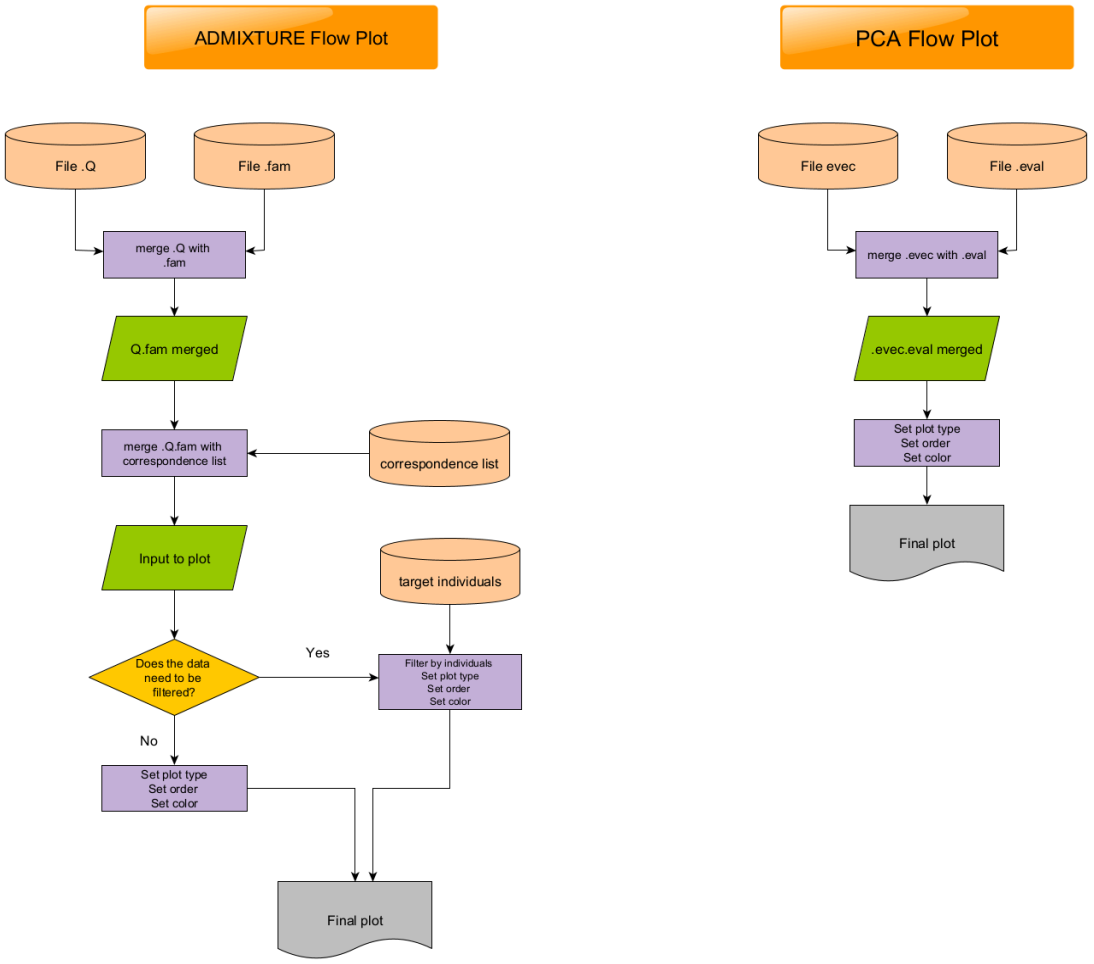

Figura 7. Fluxograma utilizado para os plots de ancestralidade. Fluxograma desenvolvido para o plot personalizado de dados de ancestralidade oriundos do ADMIXTURE e PCA.

1.3.3 Estimativa de parentesco

A fim de compreender a estrutura familiar dos indivíduos, utilizamos *o kinship coefficient* (Φij), que é quantificação da proporção de compartilhamento genético entre um par de indivíduos. Para estimar esse coeficiente, aplicamos o método implementado no *software* REAP (*Relatedness Estimation in Admixed Populations*), que é mais apropriado para indivíduos miscigenados (Thornton *et al.* 2012). As estimativas de *kinship* (Φij) por REAP estão condicionadas às proporções estimadas de ancestralidade individual (IAP, estimada com ADMIXTURE) para K=3 populações parentais e as 3 frequências alélicas da população parental por SNV. Para identificar indivíduos aparentados a serem excluídos para análises de

ancestralidade, usamos NAToRA, um método de poda de parentesco para minimizar a perda de conjunto de dados em análises genéticas e ômicas (Leal *et al.* 2022), usando o corte de parentesco de Φij>0.06. Ao remover os indivíduos relacionados, o NAToRA permite que as análises de admixture sejam mais precisas, pois pressupõe a independência e a aleatoriedade das amostras de forma que isso reduza a chance de superestimar ou subestimar as proporções ancestrais em relação às amostras não relacionadas.

1.3.4 Problemas computacionais na análise de ancestralidade

Desde abril de 2023, sou responsável pelo servidor do nosso grupo de pesquisa, uma máquina DELL PowerEdge T820 (4 processadores Octa-Core Intel(R) Xeon(R) CPU E7- 4820 2.00GHz, 128 GB de memória RAM, com armazenamento de 3T e STORAGES NAS somando 54T), no qual foram realizadas análises desse projeto.

A escalabilidade do tamanho dos dados acompanha diretamente os problemas computacionais enfrentados. Análises feitas, como o ADMIXTURE, demandam alto poder de processamento e acarretam no tempo das inferências que fazemos. Por mais que haja flexibilidade no número de *threads* e memória RAM utilizada, há uma relação inversamente proporcional na qual um baixo número de *threads* ou RAM utilizadas aumentam drasticamente o tempo gasto nas análises. Para exemplificar isso, podemos definir alguns cenários que descrevem diferentes momentos que enfrentamos:

- Uma única análise de ADMIXTURE realizada em nosso servidor, utilizando 10 threads, para 500 indivíduos e aproximadamente 160000 SNVs de K=3 a K=12 levou cerca de 40 horas; o tempo total escala proporcionalmente com a quantidade de replicadas a serem feitas
- Software como RFMix demanda grandes cargas de memória RAM que muitas vezes extrapolam a nossa capacidade computacional. Para tornar viável essa análise localmente é necessário a separação de blocos de indivíduos para execução do método, aumentando consideravelmente o tempo de execução.

A computação em nuvem tem apresenta grandes vantagens como: permite cenários no qual uma mesma análise pode ser realizada simultaneamente (o que otimiza o tempo), a não preocupação com a perda de dados, disponibilidade de grandes quantidade de armazenamento frio e memória RAM. Porém há a questão da portabilidade dos nossos dados, métodos e boas práticas para esse novo horizonte de computação, questão na qual já estamos trabalhando

juntamente ao projeto que temos com a *Amazon Web Service* (AWS) para computação em nuvem de alta performance.

1.3.5 Populações de Referência e Análises

Para as análises de ADMIXTURE, RFMix e PCA, as populações de referência para cada grupo alvo (Huánuco, FIOCRUZ COVID-19 e Mosaico) foram selecionadas de forma diferente. Isso ocorre devido às particularidades e objetivos dos estudos, como por exemplo, o uso de arranjos diferentes com SNVs diferentes.

Para as análises de PCA de Huánuco, ajustamos as populações de referência, mantendo apenas os Nativos Americanos presentes na Tabela 2, pois os resultados não eram informativos quando analisados juntamente a populações europeias e africanas. As análises foram realizadas para 393 indivíduos não aparentados e 125134 SNVs não relacionados.

Para as análises de ancestralidade individual e populacional de Huánuco, executamos ADMIXTURE com K=4 (41692 SNVs) para 2951 indivíduos a fim de inferir ancestralidade nativa americana, europeia, africana e asiática, para as seguintes populações de referências, respectivamente: Surui, Maya, Karitiana, Pima, Aimaras, Ashaninkas, Awajun, Chachapoyas, Chopccas, Lamas, Matses, Matsiguenkas, Moche, Nahua, Qeros, Quechuas, Shimaa, Tallanes, Uros, Candoshi, Jacarus, Shipibo; CEU, GBR, *French* (França), Sardinian, Orcadian, Bergamo, Tuscan, Basque, IBS, TSI; ACB, ASW, Botswana, ESN, GWD, Mandenka, Yoruba, Namibia, *South Africa* (África do Sul), *Kenya* (Quênia), LWK, MSL, *Gahna* (Gana), Tanzânia, YRI; CDX, CHB, CHS, JPT.

Como os resultados de ADMIXTURE com K=4 nos informaram que a ascendência africana e asiática eram pouco significante em nosso conjunto de dados (0.009 e 0.01, respectivamente), retiramos indivíduos africanos e asiáticos para as análises de ADMIXTURE, pois dessa forma poderia ser observada a diferente separação que ocorre entre os Nativos Americanos dentro dos dados de Huánuco e a miscigenação com europeus. Em seguida, podamos nosso conjunto de dados retirando os indivíduos relacionados (informações obtidas através do *software* NAToRA) e executamos novamente as replicatas para os K's definidos (Tabela 1), considerando apenas indivíduos não relacionados e 162957 SNVs não vinculados. Por se tratar de um estudo colaborativo, todos os indivíduos foram analisados, mesmo que retirados no primeiro filtro de parentesco. Para isso, ao identificar os aparentados, realizamos corridas

independentes para cada indivíduo que foi removido, mantendo as mesmas populações parentais e agregando os resultados na visualização final.

Tabela 2. Populações utilizadas para as análises de ancestralidade individual e populacional de Huánuco.

Origem Continental	População	Fonte	Número de Indivíduos (Autossômicos)
Europeus*	IBS	1000 Genomes	107
Nativos Americanos	Tallanes	Borda et al. 2020	35
	Moche	Borda et al. 2020	30
	Jacarus	Borda et al. 2020	17
	Quechuas	Borda et al. 2020	24
	Chopccas	Borda et al. 2020	17
	Qeros	Borda et al. 2020	12
	Aimaras	Borda et al. 2020	16
	Uros	Borda et al. 2020	16
	Ashaninka	Borda et al. 2020	36
	Matsiguenka_A	Borda et al. 2020	3
	Matsiguenka_B	Borda et al. 2020	23
	Matses	Borda et al. 2020	11
	Nahua	Borda et al. 2020	2
	Shipibo	Borda et al. 2020	16
	Candoshi	Borda et al. 2020	16
	Awajun	Borda et al. 2020	23
	Lamas	Borda et al. 2020	21
	Chachapoyas	Borda et al. 2020	40
Miscigenados	Huánuco	Não publicado	35+9**
Total:	-	-	509

*Retirados na análise de PCA. ** 35 indivíduos não aparentados e 9 indivíduos aparentados.

Para a inferência de ancestralidade cromossômica local o RFMix utilizou 208523 SNVs e a análise foi realizada com as opções PopPhased para habilitar a correção de fase e cinco iterações do algoritmo de maximização de expectativa (EM). Todas as outras configurações foram usadas como padrão.

Tabela 3. Populações utilizadas para as análises de ancestralidade local dos dados de Huánuco.

Origem Continental	População	Fonte	Número de Indivíduos (Autossômicos)
Africanos	YRY	1000 Genomes	108
	LWK	1000 Genomes	99
Europeus	IBS	1000 Genomes	107
	TSI	1000 Genomes	107
Nativos Americanos	Tallanes	Borda et al. 2020	35
	Moche	Borda et al. 2020	30
	Jacarus	Borda et al. 2020	17
	Quechuas	Borda et al. 2020	24
	Chopccas	Borda et al. 2020	17
	Qeros	Borda et al. 2020	12
	Aimaras	Borda et al. 2020	16
	Uros	Borda et al. 2020	16
	Ashaninka	Borda et al. 2020	36
	Matsiguenka_A	Borda et al. 2020	3
	Matsiguenka_B	Borda et al. 2020	23
	Matses	Borda et al. 2020	11
	Nahua	Borda et al. 2020	2
	Shipibo	Borda et al. 2020	16
	Candoshi	Borda et al. 2020	16
	Awajun	Borda et al. 2020	23

	Lamas	Borda et al. 2020	21
	Chachapoyas	Borda et al. 2020	40
	Huánuco	não publicado	35
Total:	-	-	814

Para as análises de PCA de FIOCRUZ COVID-19, podamos nosso conjunto de dados retirando os indivíduos relacionados (informações obtidas através do *software* NAToRA) considerando 177994 SNVs não relacionados e realizamos as corridas para as populações de referência presentes na Tabela 4. Além disso, calculamos o PCA uma segunda vez apenas entre os próprios indivíduos (1045) da coorte para um total de 515757 SNVs. Ambas abordagens foram feitas para indivíduos não relacionados.

Para inferir a ancestralidade individual e populacional dos indivíduos FIOCRUZ COVID-19, usamos um conjunto adicional de 1600 indivíduos de diferentes populações que contribuíram para a população brasileira (Kehdy *et al.* 2015, Gouveia *et al.* 2015). Da mesma maneira, indivíduos aparentados foram retirados na primeira análise conjunta e analisados separadamente com as mesmas populações de referência (Tabela 4). As análises foram realizadas para um total de 178024 SNVs não relacionados.

Através das informações compartilhadas com nosso grupo, foi possível fazer o agrupamento dos indivíduos por região (Norte, Nordeste, Centro-Oeste, Sul e Sudeste) e por casos e controle (hospitalizado e não hospitalizado, respectivamente). Dessa maneira, foi possível a análise por região e caso-controle (Seção 1.4.3).

Tabela 4. Populações utilizadas para as análises de componente principal, ancestralidade individual e populacional dos dados FIOCRUZ COVID-19.

Origem Continental	**População**	**Fonte**	**Número de Indivíduos (Autossômicos)**
Africanos	GWD	1000 Genomes	112
	MSL	1000 Genomes	78
	YRI	1000 Genomes	108
	ESN	1000 Genomes	95
	LWK	1000 Genomes	88

Europeus	IBS	1000 Genomes	107
	GBR	1000 Genomes	91
	CEU	1000 Genomes	51
	TSI	1000 Genomes	107
Nativos Americanos	Nahua	Borda et al. 2020	2
	Moche	Borda et al. 2020	30
	Chachapoyas	Borda et al. 2020	40
	Tallanes	Borda et al. 2020	35
	Ashaninkas	Borda et al. 2020	36
	Lamas	Borda et al. 2020	21
	Aimaras	Borda et al. 2020	16
	Awajun	Borda et al. 2020	23
	Candoshi	Borda et al. 2020	16
	Uros	Borda et al. 2020	16
	Shipibo	Borda et al. 2020	16
	Matsiguenka_B	Borda et al. 2020	23
	Quechuas	Borda et al. 2020	24
	Jacarus	Borda et al. 2020	17
	Matses	Borda et al. 2020	11
	Matsiguenka_A	Borda et al. 2020	3
	Chopccas	Borda et al. 2020	17
	Qeros	Borda et al. 2020	12
Asiáticos	CDX	1000 Genomes	91
	CHB	1000 Genomes	103
	CHS	1000 Genomes	107
	JPT	1000 Genomes	104

Miscigenados	FIOCRUZ COVID-19	Não Publicado	1045+42**
Total	-	-	2687

**1045 Indivíduos não aparentados e 42 indivíduos aparentados.

Para as análises de ancestralidade cromossômica local utilizamos as mesmas populações de referência e suas respectivas quantidades de indivíduos da Tabela 4, porém, na população do estudo (FIOCRUZ COVID-19) não retiramos os indivíduos aparentados, o que nos permitiu passar de 1045 para 1087 indivíduos (passando o total de 2645 para 2687 indivíduos). Essa abordagem foi realizada, pois indivíduos aparentados não interferem na inferência de ancestralidade cromossômica local realizada pelo RFMix assim como no ADMIXTURE. Para esses resultados, o RFMix utilizou 205421 SNVs.

Para os dados da Mosaico, a análise de PCA utilizou 404997 SNVs não relacionados (Tabela 5). Para as análises de ancestralidade individual e populacional da Mosaico, executamos o ADMIXTURE. Através das informações obtidas pelo *software* NAToRA houve a necessidade da retirada de indivíduos devido ao parentesco. As análises foram realizadas para 404997 SNVs não vinculados para 1471 indivíduos não aparentados (Tabela 5).

Tabela 5. Populações utilizadas para as análises de ADMIXTURE, PCA e RFMix da Mosaico.

Origem Continental	População	Fonte	Número de Indivíduos (Autossômicos)
Africanos	ESN	1000 Genomes	99
	YRI	1000 Genomes	108
	GWD	1000 Genomes	113
	MSL	1000 Genomes	85
	LWK	1000 Genomes	99
Europeus	TSI	1000 Genomes	107
	IBS	1000 Genomes	107
	GBR	1000 Genomes	91
	CEU	1000 Genomes	99

Nativos Americanos	Pima	HGDP	11
	Maya	HGDP	19
	Surui	HGDP	6
	Karitiana	HGDP	8
Miscigenados	ACB	1000 Genomes	96
	ASW	1000 Genomes	62
	CLM	1000 Genomes	94
	MXL	1000 Genomes	64
	PEL	1000 Genomes	85
	PUR	1000 Genomes	104
Mosaico_miscigenados	Mosaico_Admixed	Mosaico Translational Genomics - Não Publicado	4
Mosaico_2022	Mosaico_Admixed	Esse estudo	10
Total:	-	-	1471

1.4 Resultados e discussão

1.4.1 Huánuco, Peru.

Através das análises de componentes principais é possível observar a correlação entre os dados genéticos e a geolocalização. Por se tratar de uma população peruana próxima aos Andes era esperado que suas componentes genéticas ao serem reduzidas em dimensões menores elas estariam mais próximas das populações andinas quando comparadas às do norte e sul amazônicos (Figura 8).

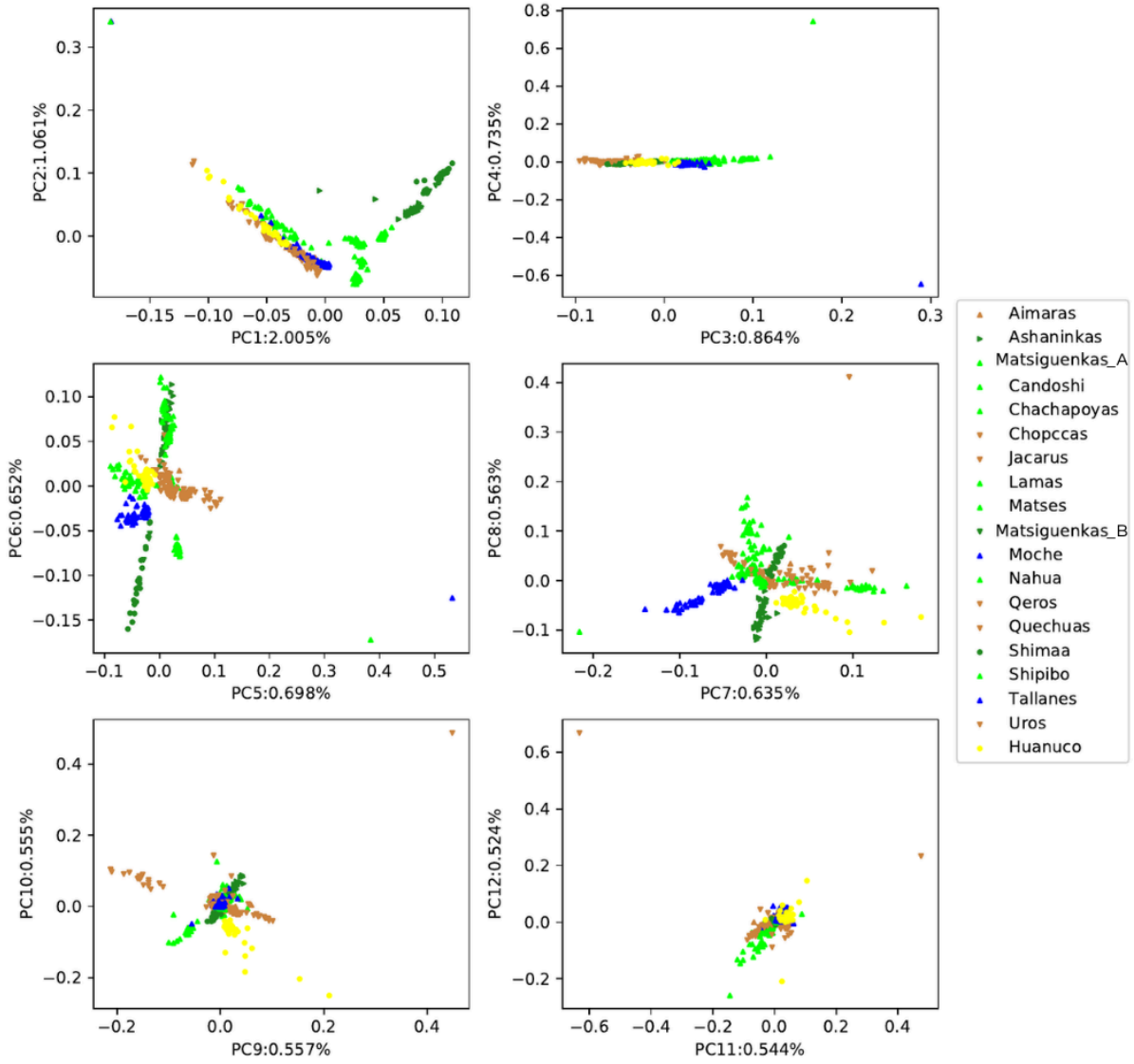

Figura 8. **Análise de Componentes Principais - Huánuco.** Gráfico gerado a partir dos dados obtidos do *software* EIGENSOFT (Patterson *et al.* 2006 e Price *et al.* 2006) para os cálculos de PCA de Huánuco. PC1 e PC2 explicam 2.005% e 1.061% da variação genética, respectivamente, evidenciando a proximidade dos Huánucos da populações andinas. Análise realizada para 125134 SNVs não relacionados.

Em geral, a maioria dos indivíduos Huánuco estudados não têm parentesco (Figura 9). Consistentemente com as informações fornecidas pelos participantes do estudo, as análises genéticas identificaram apenas cinco pares de indivíduos com $0.296 > \Phi ij \geq 0.204$ (parentesco de primeiro grau) e três pares de indivíduos com $0.158 > \Phi ij \geq 0.092$ (parentesco de segundo grau).

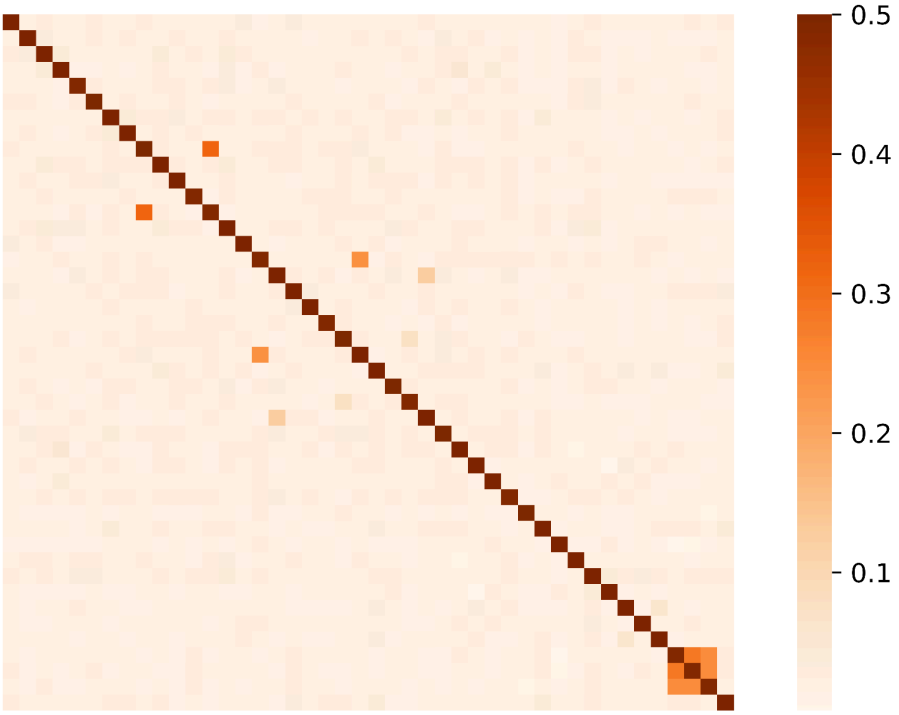

Figura 9. **Kinship - Huánuco**. Mapa de calor representando a matriz de coeficientes de *kinship*, estimada usando REAP (*Relatedness estimation in admixed populations*, Thornton *et al.* 2012) entre os indivíduos de Huánuco. A diagonal representa os coeficientes de parentesco entre os mesmos indivíduos, igual a 0.50.

As análises de ADMIXTURE para K=4 (Tabela 6, Figura 11) revelam que os indivíduos Huánuco são predominantemente Nativos Americanos (≈88%). A partir da divisão subcontinental, observamos uma ancestralidade mais Andina (≈73%) que Amazônica (≈15%) e com baixa ascendência europeia (≈11%).

Tabela 6. Médias das proporções de ancestralidade de Huánuco para K=4 inferidas por ADMIXTURE.

Origem Continental	População	Andino	Norte Amazônico	Sul Amazônico	Europeu
Europa	IBS	0.00061	0.00093	0.0007	0.99777
Baixo Amazonas	Nahua	0.08305	0.49763	0.41931	0.00002
	Matses	0.07227	0.52943	0.39827	0.00002
	Shipibo	0.12906	0.44903	0.39161	0.0303
Yunga na Amazônia	Matsigenka_A	0.00314	0.02202	0.97481	0.00002
	Matsiguenka_B	0.00001	0.06097	0.939	0.00002
	Ashaninkas	0.01308	0.26923	0.71199	0.00569

	Lamas	0.10624	0.65684	0.22366	0.01326
	Candoshi	0.04878	0.82164	0.12956	0.00002
	Awajun	0.00001	0.99389	0.00241	0.00369
	Chachapoyas	0.50443	0.33356	0.03034	0.13167
Costa Andina	Tallanes	0.51898	0.46249	0.01055	0.00797
	Moche	0.60069	0.34193	0.0031	0.05429
Andes Árido	Quechuas	0.80759	0.06612	0.01312	0.11317
	Jacarus	0.91606	0.00202	0.00353	0.07839
	Aimaras	0.88577	0.05517	0.03734	0.02172
	Uros	0.95096	0.01237	0.03357	0.0031
	Chopccas	0.94126	0.04121	0.00964	0.00789
	Qeros	0.96664	0.0114	0.02194	0.00002
Miscigenados	Huanuco	0.7271	0.1489	0.0103	0.1137

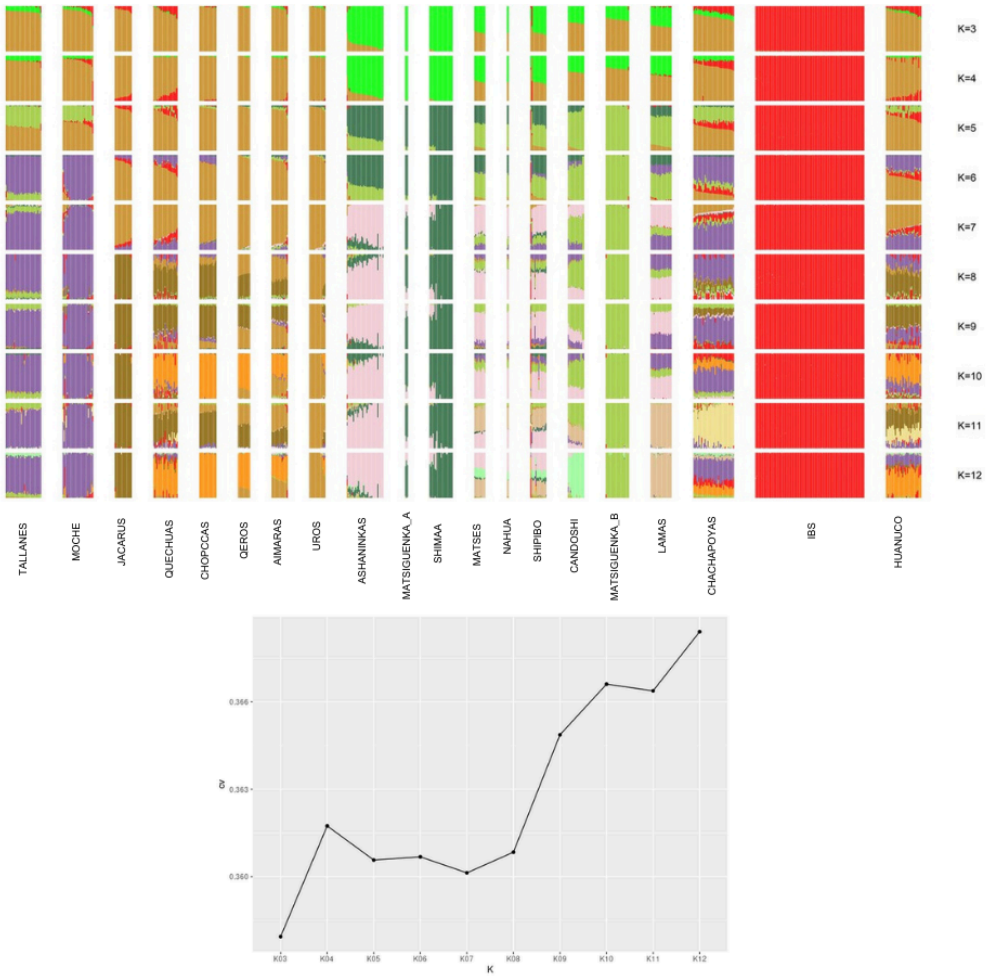

Figura 10. **Admixture Huánuco.** Gráfico de barras vertical das análises de ADMIXTURE com diferentes números de grupos ancestrais (K) para os Huánuco, populações indígenas peruanas de Borda *et al.* 2020 e Ibéricos (IBS). As análises são baseadas em 162957 SNVs. Nos barplots cada indivíduo é representado por uma barra vertical e cada cor corresponde à proporção de ancestralidade de um *cluster* específico. Na parte inferior temos o gráfico de validação cruzada que juntamente aos conhecimentos prévios da população do estudo, observamos que para K's > 7 teremos resultados com menor confiança.

Com base nas análises realizadas, foram observados resultados de ancestralidade nativa que são consistentes com os achados apresentados em Borda *et al.* 2020 (Figura 11). Essa congruência reforça a diversidade genética das populações nativas da América do Sul e destaca a necessidade de estudos adicionais para aprofundar a compreensão da história evolutiva dessas populações.

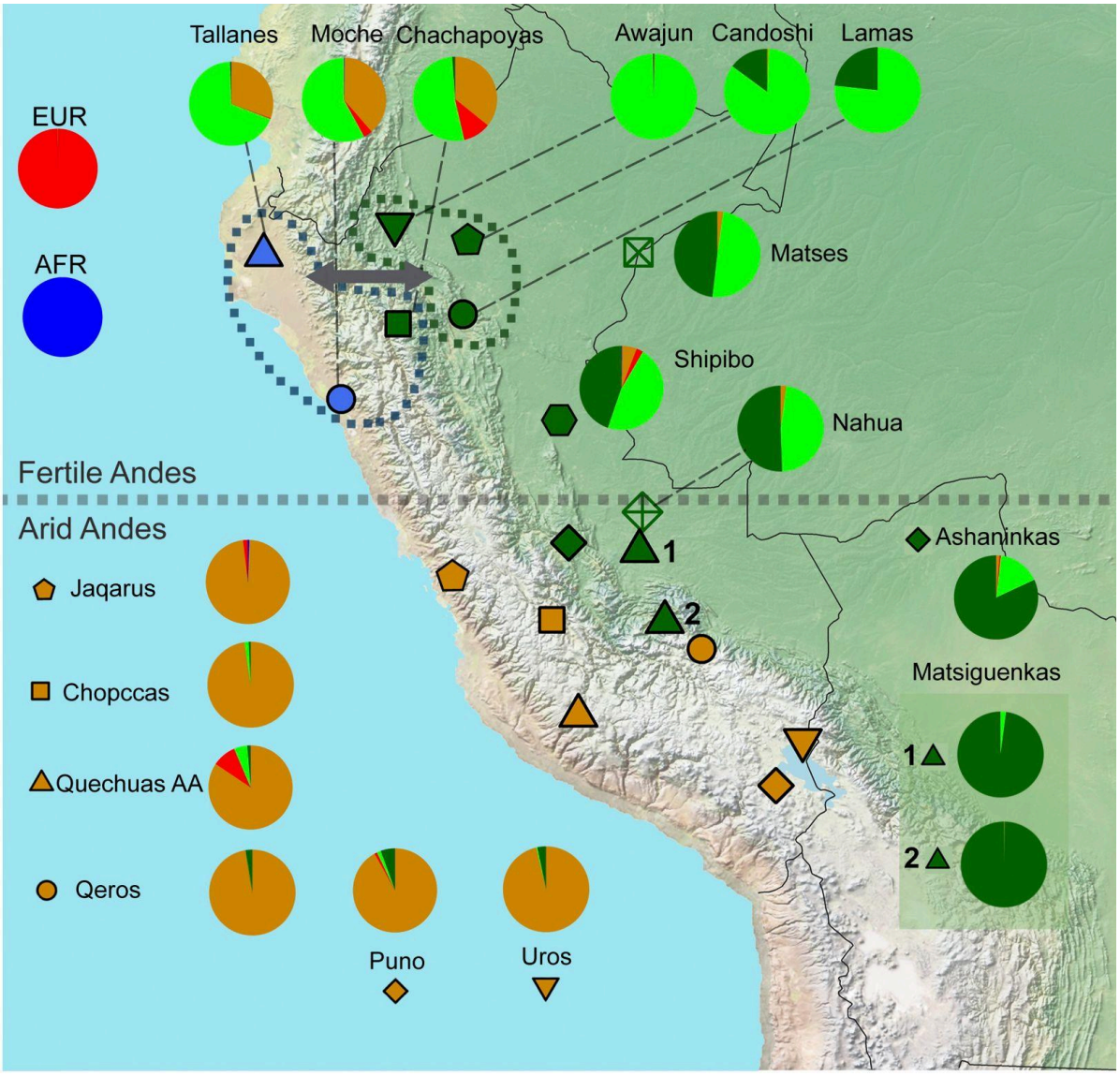

Figura 11. **Ancestralidade subcontinental nativo americana.** Distribuição geográfica e estrutura genética para 18 populações nativas da costa, Andes e Amazônia inferidas a partir do *software* ADMIXTURE (K = 5, correspondente ao menor erro de validação cruzada). Os gráficos de pizza mostram as porcentagens médias sobre indivíduos para os cinco grupos de ADMIXTURE em cada população. A cor Vermelho representa ancestralidade Europeia, Azul ancestralidade Africana, Verde claro ancestralidade nativo americana no Norte Amazônico, Verde Escuro ancestralidade nativa americana do Sul Amazônico e em Ocre temos representada a ancestralidade nativa americana no Andes. Figura adaptada de Borda *et al.* 2020.

A população de Huánuco possui pouca variabilidade de ancestralidade quando tentamos observar contribuição parental que não seja nativo americana. Através dos dados de ancestralidade local, podemos observar que não há uma grande diferença entres os indivíduos mais e menos nativos (Figura 12A e 12B respectivamente).

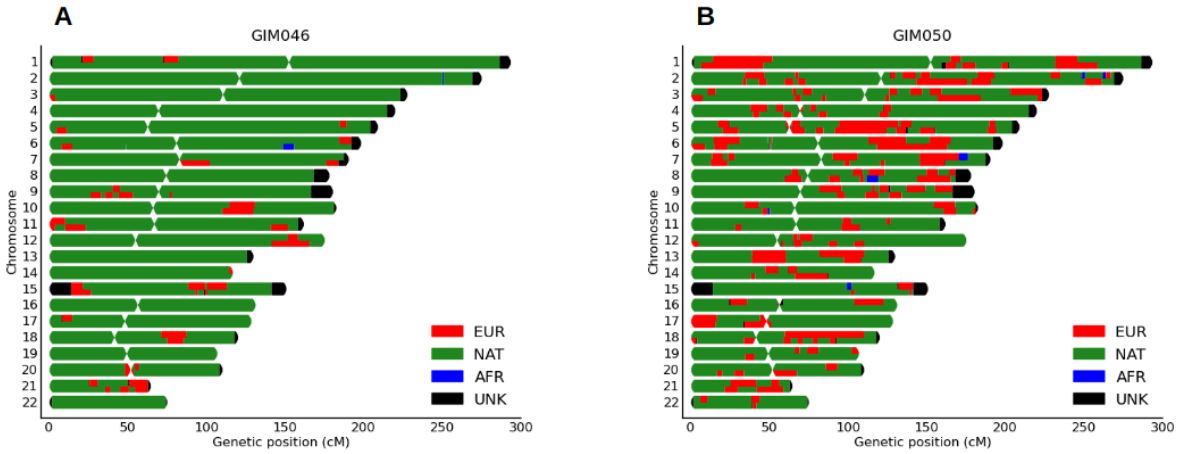

Figura 12. Representação de ancestralidade local - Huánuco. Gráfico da ancestralidade local cromossômica de dois indivíduos gerados a partir dos dados de RFMix a partir das populações parenterais descritas na Tabela 2. **A** indivíduo com maior componente nativo americana, ≈96%. **B** indivíduo com menor componente nativo americana, ≈75%.

1.4.2 GWAS COVID-19 da FIOCRUZ

As análises de PCA evidenciam a estruturação populacional presente nos nossos dados. Uma vez que tivemos acesso às informações de correspondência indivíduo-região, foi possível observar os agrupamentos que ocorrem para os nossos indivíduos, podemos atribuir tais resultados ao alto grau de miscigenação da população estudada (Figura 13).

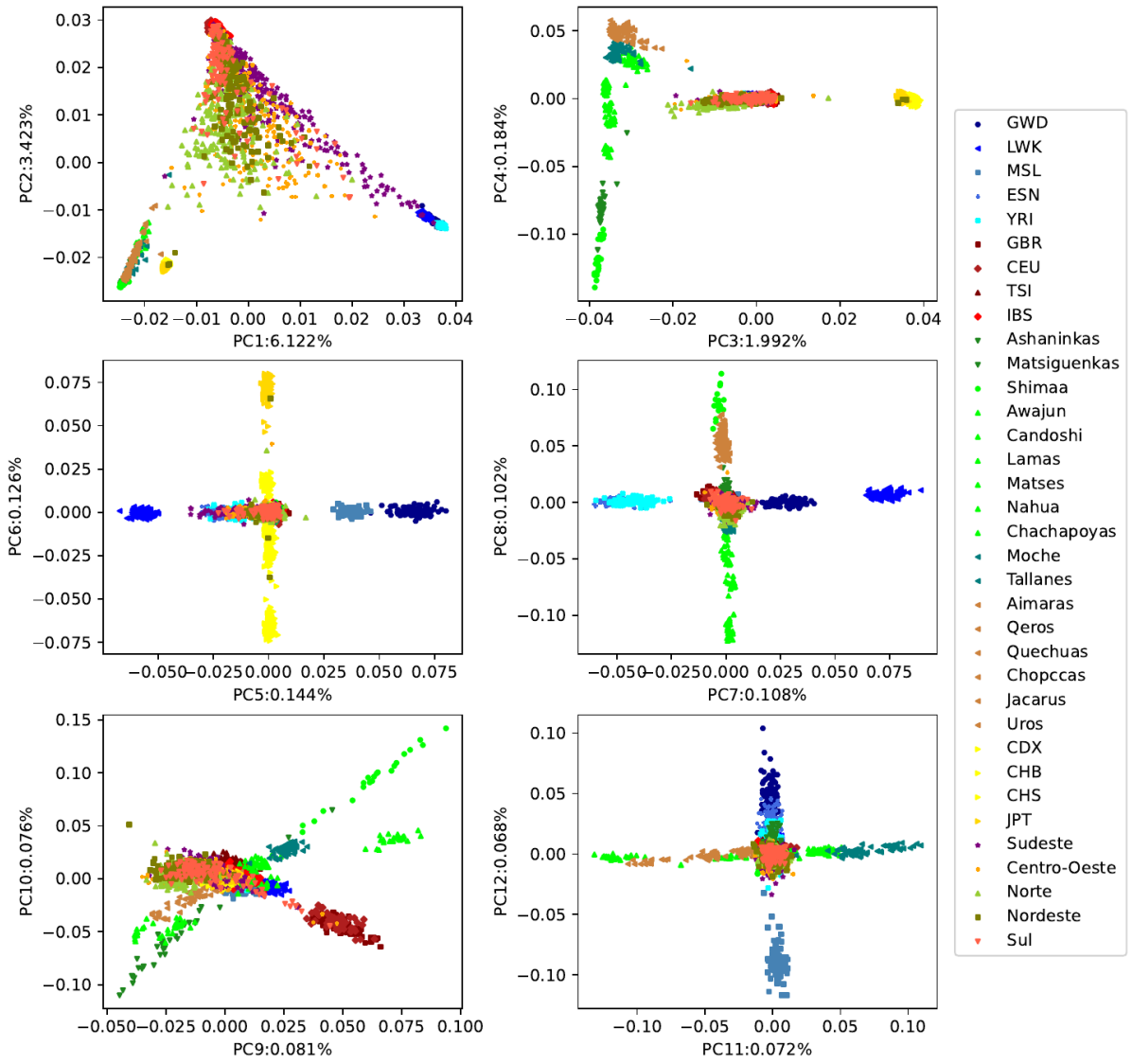

Figura 13. Análise de componente Principal - FIOCRUZ COVID-19. Gráfico gerado a partir dos dados obtidos do *software* EIGENSOFT (Patterson *et al*. 2006, Price *et al*. 2006) para os cálculos de PCA de FIOCRUZ COVID-19 separados por região. PC1 e PC2 explicam 6.122% e 3.423% da variação genética.

Em geral, a maioria dos indivíduos da amostras da FIOCRUZ COVID-19 estudados não têm parentesco (Figura 14). Através dos resultados de parentescos, de 1126 indivíduos, foi possível observar replicatas e aparentados (39 indivíduos e 42 indivíduos, respectivamente). Consistentemente com as informações fornecidas pelos participantes do estudo, as análises genéticas identificaram apenas 33 pares de indivíduos com $0.296 > \Phi ij \geq 0.204$ (parentesco de primeiro grau) e 9 pares de indivíduos com $0.158 > \Phi ij \geq 0.092$ (parentesco de segundo grau).

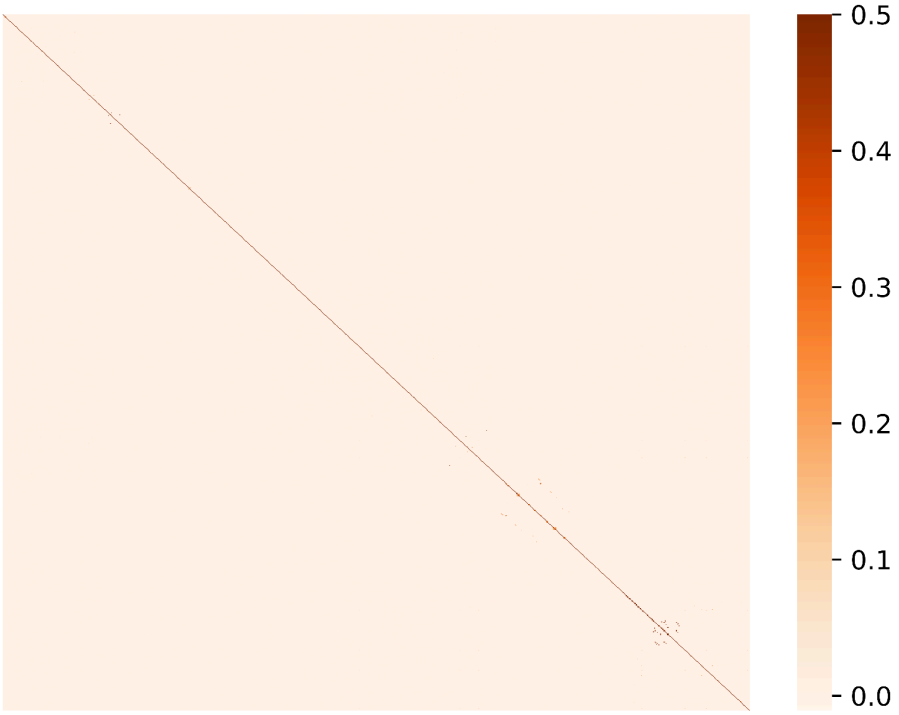

Figura 14. **Kinship - FIOCRUZ COVID-19**. Mapa de calor representando a matriz de coeficientes de *kinship*, estimada usando REAP (*Relatedness estimation in admixed populations*, Thornton *et al.* 2012). A diagonal representa os coeficientes de parentesco entre os mesmos indivíduos, igual a 0.50.

As análises de ADMIXTURE para K=4 (Tabela 7, Figura 15) revelam que os indivíduos da FIOCRUZ COVID-19 são predominantemente Europeus (≈66%) com uma menor ascendência africana e nativo americana (≈18% e ≈14% respectivamente). A contribuição asiática é pouco significativa (≈0.95%). Corroborando com a história demográfica de ancestralidade do nosso país (Kehdy *et al.* 2015), foi possível observar uma maior componente nativo americana nos indivíduos do Norte e Centro-Oeste (≈25% e ≈17%). Da mesma maneira, Sul e Sudeste apresentaram uma componente europeia maior que as demais ≈76% e ≈69%, respectivamente.

Tabela 7. Médias das proporções de ancestralidade para K =4 para FIOCRUZ COVID-19

Origem continental	Pop	Europeu	Nativo Americano	Asiático	Africano
África	YRI	0.00010	0.00001	0.00003	0.99985
	LWK	0.03145	0.00036	0.01007	0.95813
	MSL	0.00010	0.00008	0.00004	0.99977
	ESN	0.00001	0.00001	0.00001	0.99997

	GWD	0.01259	0.00035	0.00125	0.98581
Europa	CEU	0.99105	0.00840	0.00054	0.00001
	GBR	0.99333	0.00658	0.00007	0.00001
	TSI	0.99631	0.00030	0.00307	0.00031
	IBS	0.98917	0.00044	0.00094	0.00945
América Latina	Aimaras	0.03122	0.96463	0.00315	0.00099
	Ashaninkas	0.00576	0.99422	0.00001	0.00001
	Matsiguenkas_B	0.00421	0.99572	0.00006	0.00001
	Candoshi	0.00001	0.99983	0.00015	0.00001
	Lamas	0.01892	0.98024	0.00073	0.00010
	Matses	0.00001	0.99997	0.00001	0.00001
	Matsiguenkas_A	0.00001	0.99997	0.00001	0.00001
	Nahua	0.00001	0.99997	0.00001	0.00001
	Qeros	0.00762	0.99165	0.00072	0.00001
	Quechuas	0.13041	0.86328	0.00398	0.00233
	Shimaa	0.00024	0.99899	0.00001	0.00076
	Shipibo	0.02799	0.96909	0.00018	0.00274
	Tallanes	0.01138	0.98308	0.00301	0.00253
	Uros	0.01248	0.98610	0.00141	0.00001
	Chopccas	0.02997	0.96763	0.00220	0.00019
	Moche	0.06278	0.92617	0.00616	0.00489
	Chachapoyas	0.13223	0.85650	0.00470	0.00657
	Jacarus	0.10097	0.89254	0.00088	0.00560
Ásia	CDX	0.00015	0.00008	0.99976	0.00001
	CHB	0.00355	0.00806	0.98837	0.00001
	CHS	0.00010	0.00033	0.99956	0.00001

	JPT	0.00001	0.01316	0.98682	0.00001
FIOCRUZ COVID-19	Nordeste	0.67465	0.14645	0.02459	0.15431
	Norte	0.58132	0.25294	0.00821	0.15753
	Centro-Oeste	0.59181	0.16549	0.00690	0.23580
	Sudeste	0.69007	0.06604	0.00427	0.23962
	Sul	0.76796	0.10824	0.00371	0.12009
Origem continental	**Pop**	**Europeu**	**Nativo Americano**	**Asiático**	**Africano**

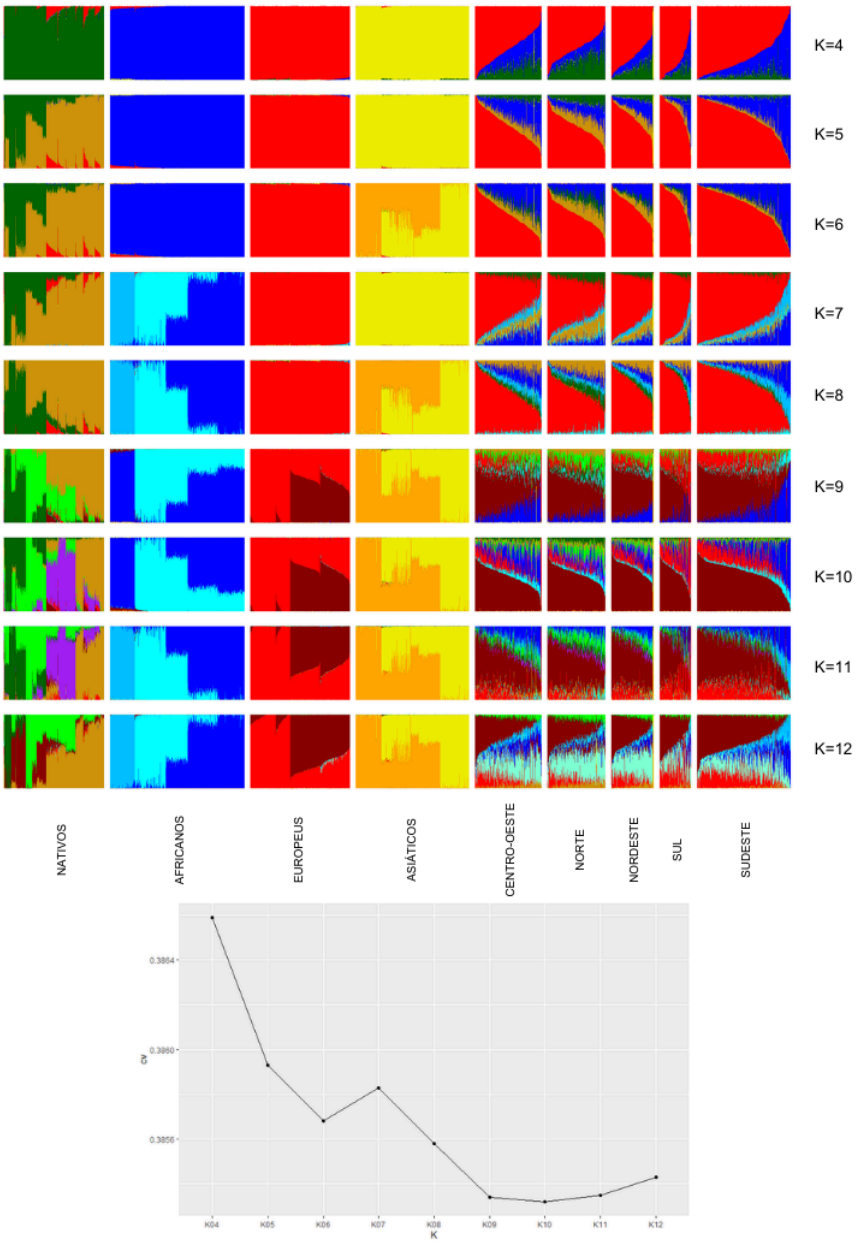

Figura 15. **Admixture FIOCRUZ COVID-19.** Gráfico de barras vertical das análises de ADMIXTURE com diferentes números de grupos ancestrais (K) para o indivíduos do estudo da FIOCRUZ COVID-19 e outras populações (Tabela 4). Os indivíduos do estudo foram separados por região e caso-controle. Abaixo o gráfico de validação cruzada indica que para valores de K>6 as inferências perdem poder estatístico para nosso estudo quando comparado com K's menores.

Os dados de fenótipo da FIOCRUZ COVID-19 nos forneceram a informação de CASO (Hospitalizado) e CONTROLE (Não hospitalizado) de cada indivíduo. Através disso, foi possível observar uma diferença de proporção de ancestralidade africana entre casos e controles (22% e 16% respectivamente). Essa diferença pode indicar uma correlação entre ancestralidade e hospitalização que poderá ser verificada durante os estudos de associação genômica juntamente com os resultados obtidos da ancestralidade local.

1.4.3 Mosaico

Assim como nos resultados anteriores, a análise de componentes principais demonstrou a correlação entre os dados genéticos e a geolocalização (Figura 16). Por se tratar de uma população miscigenada, era esperado que suas componentes genéticas ao serem reduzidas em dimensões menores estariam mais dispersas para PC's de maiores valores (PC1 e PC2). Apesar da análise de PCA ter sido feita com outras populações miscigenadas (como, por exemplo, PUR, MXL, CLM e PEL), as componentes PC3 e PC4 separam os indivíduos da Mosaico dos demais miscigenados. Isso pode estar relacionado com o fato de que o processo de formação dessas populações ocorreu de forma diferente, indicando a diversidade presente na américa latina.

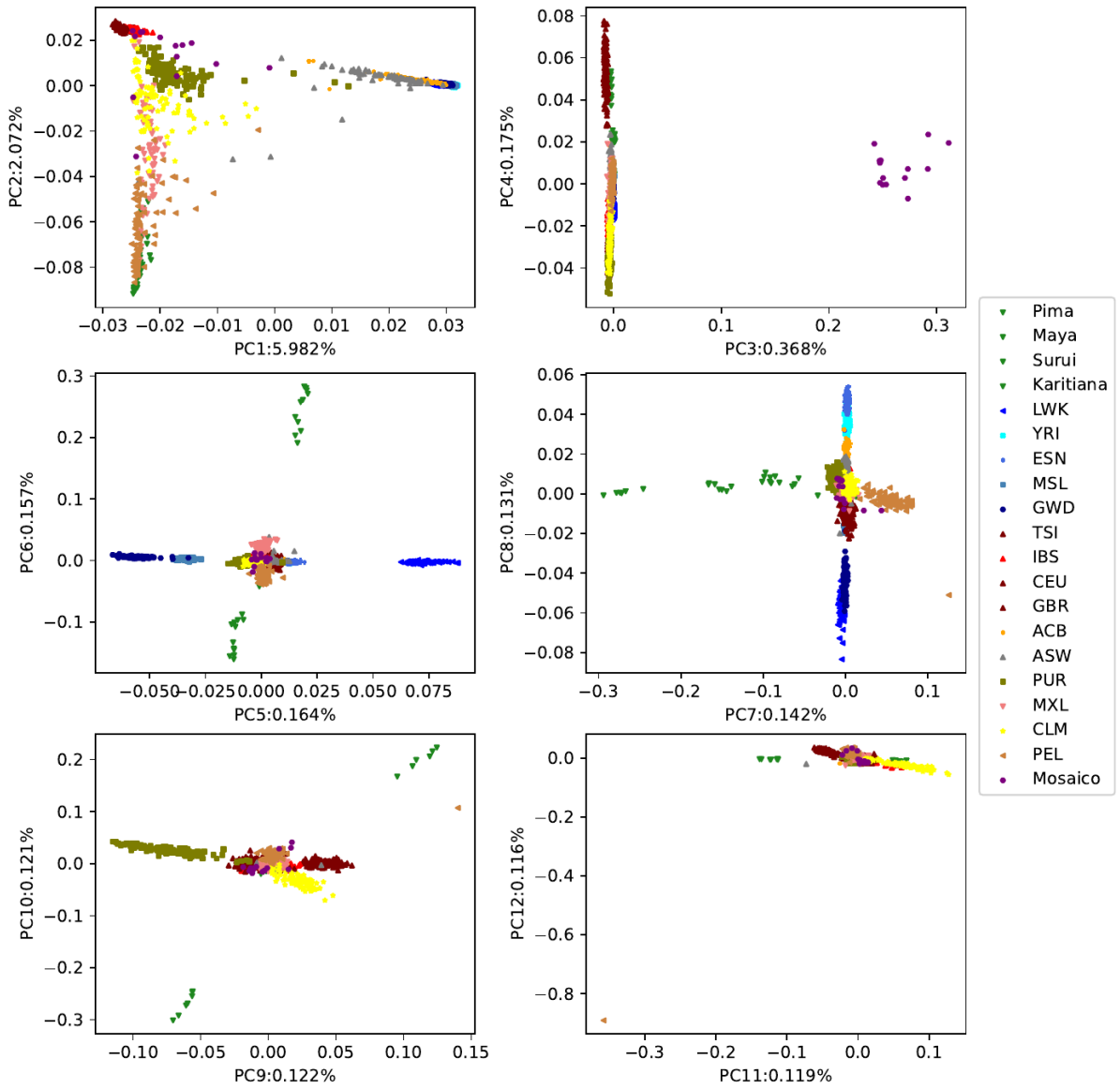

Figura 16. Análise de componente Principal - Mosaico Gráfico gerado a partir dos dados obtidos do *software* EIGENSOFT (Patterson *et al.* 2006 e Price *et al.* 2006) para os cálculos de PCA da Mosaico. Análise feita para indivíduos não aparentados e sem SNVs relacionados.

Em geral, a maioria dos indivíduos das amostras da Mosaico estudados não têm parentesco (Figura 17). Foi possível, através dos resultados de parentescos, observar que não havia duplicatas. Consistentemente com as informações fornecidas pelos participantes do estudo, as análises genéticas identificaram apenas uma família de pai, mãe e um(a) filho(a). Os demais indivíduos não apresentaram parentesco significativo. Os indivíduos aparentados retirados (sugeridos pelo *software* NAToRA) foram analisados separadamente e adicionados aos resultados finais de ADMIXTURE.

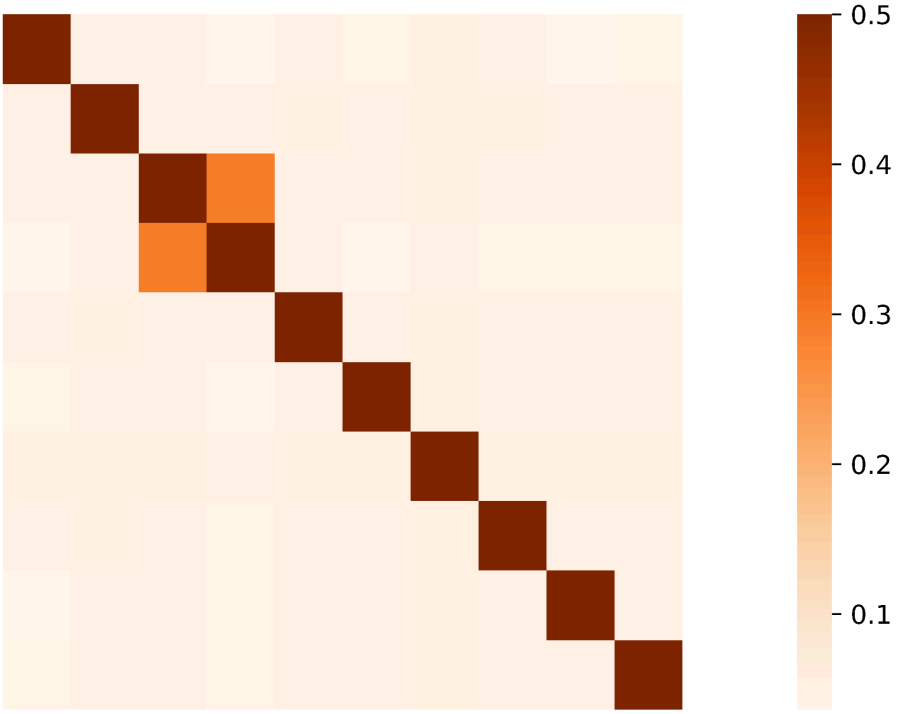

Figura 17. **Kinship - Mosaico.** Mapa de calor representando a matriz de coeficientes de *kinship*, estimada usando REAP (*Relatedness estimation in admixed populations*, Thornton *et al.* 2012). A diagonal representa os coeficientes de parentesco entre os mesmos indivíduos, igual a 0.50.

As análises de ADMIXTURE para K=3 (Tabela 8, Figura 18) revelam que os indivíduos da Mosaico são predominantemente Europeus (≈77%) com uma menor ascendência africana e nativo americana (≈14% e ≈9% respectivamente).

Tabela 8. Médias das proporções de ancestralidade para K=3 para Mosaico

Origem Continental	População	Africano	Europeu	Nativo Americano
África	ESN	0.99998	0.00001	0.00001
	YRI	0.99998	0.00001	0.00001
	GWD	0.995455	0.004461	0.000084
	MSL	0.99998	0.00001	0.00001
	LWK	0.973573	0.02184	0.004588
Europa	TSI	0.016568	0.983377	0.000055
	IBS	0.021033	0.978763	0.000204
	GBR	0.000042	0.997737	0.002221

	CEU	0.000415	0.997585	0.001999
Nativos América Latina	Pima	0.000568	0.000514	0.998918
	Maya	0.019571	0.062191	0.918238
	Surui	0.000011	0.00001	0.999979
	Karitiana	0.00001	0.00001	0.99998
Miscigenados América Latina	ACB	0.889354	0.108269	0.002377
	ASW	0.756783	0.202677	0.04054
	CLM	0.090228	0.63368	0.276092
	MXL	0.053147	0.443023	0.50383
	PEL	0.033355	0.169258	0.797387
	PUR	0.149743	0.707561	0.142696
América Latina	Mosaico	0.138492	0.76853	0.092978

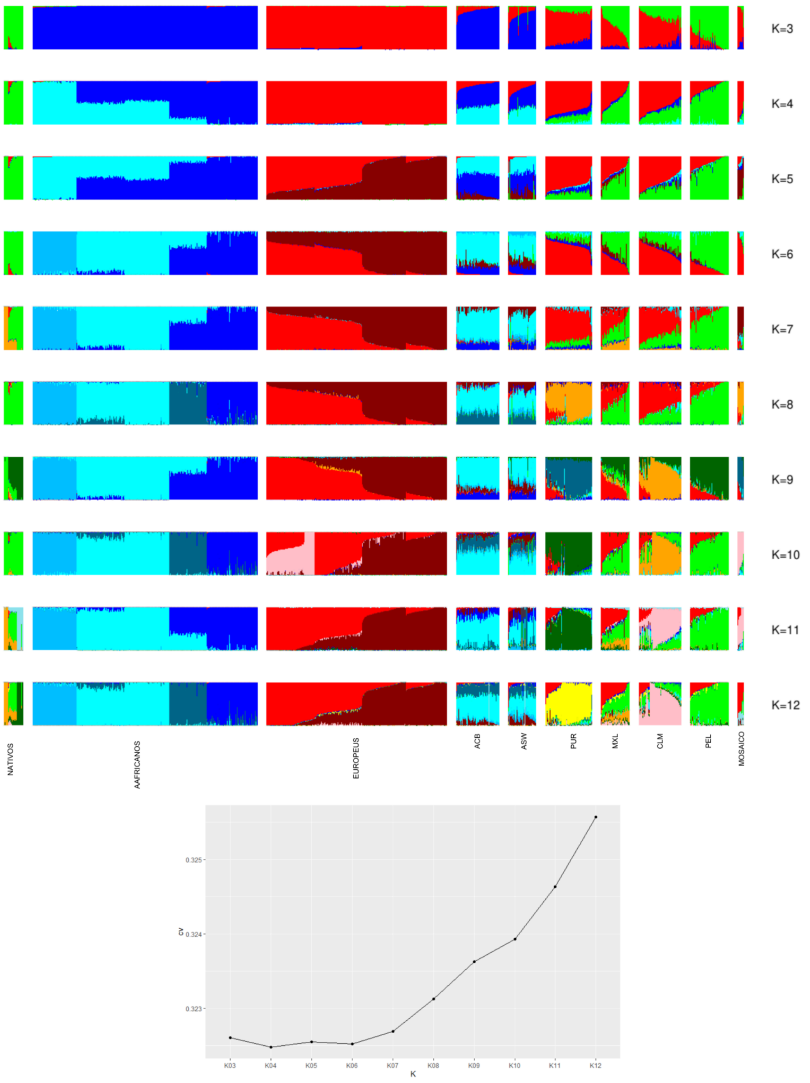

Figura 18. **Admixture - Mosaico.** Gráfico de barras vertical das análises de ADMIXTURE com diferentes números de grupos ancestrais (K) para os indivíduos do estudo da Mosaico e outras populações (Tabela 5) para 404997 SNVs em que não estão em LD. Abaixo o gráfico de erro validação cruzada que aponta um maior erro para correlações feitas para valores de K > 6.

Os indivíduos da Mosaico possuem alta variabilidade de ancestralidade quando comparados entre si. Através dos dados de ancestralidade local, podemos observar que há uma grande diferença entres os indivíduos com mais componentes nativas e o com mais componente africana (Figura 19A e 19B respectivamente).

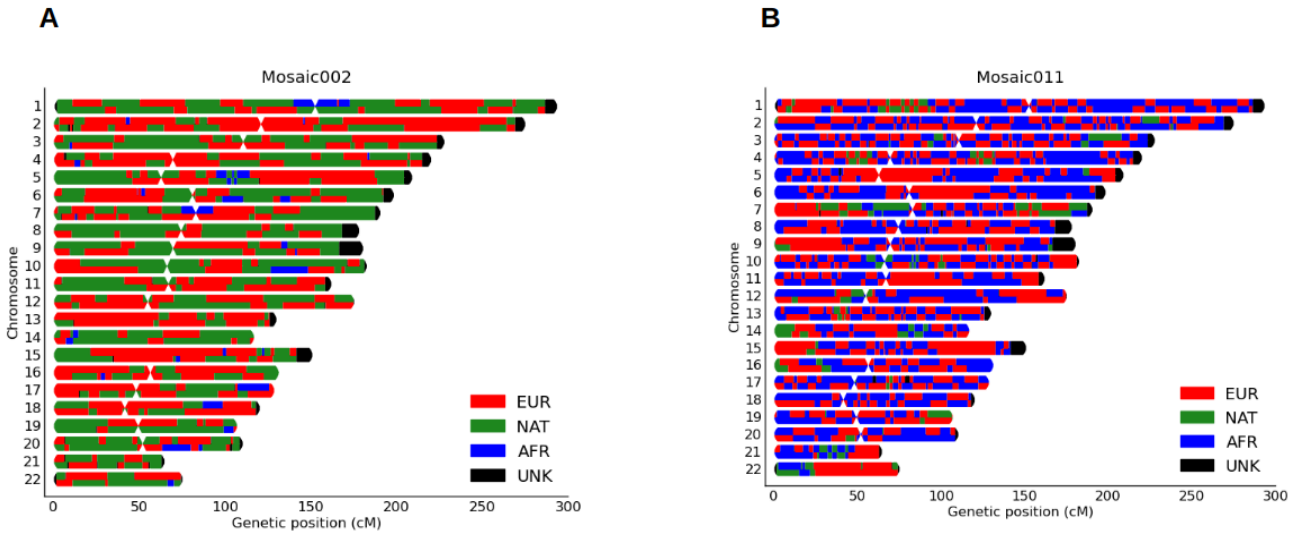

Figura 19. Gráficos de ancestralidade local - Mosaico. A indivíduo com mais componentes nativas, **B** indivíduo com mais componentes africanas.

1.4.4 Discussão

Os desafios enfrentados e a importância da constante manutenção e atualização dos pipelines e metodologias empregados em análises genéticas populacionais são aspectos essenciais no contexto bioinformático atual. A automação dos processos, por meio do domínio avançado de linguagens de programação, desempenha um papel fundamental na implementação de novas abordagens que possam validar as hipóteses estabelecidas. Uma vez que as análises são realizadas localmente, as limitações enfrentadas para as corridas estão atreladas ao poder computacional. Assim, à medida que as novas ferramentas de inferência de ancestralidade demandarem mais poder computacional, estaremos limitados às metodologias ultrapassadas. Por exemplo, o software NgsAdmix que incorpora dados de sequenciamento completo via Sequenciamento de Nova Geração, porém demandando maior poder computacional. Deste modo, como próximos passos, faremos o aporte dos nossos pipelines antigos e atualizados para as análises de computação em nuvem, para que possamos diminuir o tempo e aumentar a confiança dos nossos resultados.

Os resultados obtidos a partir das análises de ancestralidade realizadas em colaboração com os parceiros do LDGH corroboram com a ideia da diversidade genética entre as populações nativas americanas. Torna-se cada vez mais imperativo um acompanhamento específico para cada grupo étnico nativo da América, especialmente em um mundo globalizado, onde esses indivíduos podem estar expostos a riscos. Além disso, os estudos que buscam entender a complexidade das populações miscigenadas na América Latina desempenham um papel

crucial na construção de uma base de dados populacionais sólida e soberana. Isso permitirá a realização de estudos de associação genômica mais precisos, utilizando como referência não apenas dados de europeus ou norte-americanos, mas também dados das populações nativas da América Latina.

A baixa densidade de SNVs que obtivemos não é suficiente para uma análise precisa da ancestralidade cromossômica local. Como perspectiva, planejamos realizar uma reavaliação, incorporando novas populações ancestrais como referência e utilizando a técnica de imputação de dados para substancialmente aumentar nosso banco de dados de variantes das populações alvo.

1.4.5 Conclusão

A população de Huánuco mostrou-se como uma população mais isolada devido ao fato de sua baixa variabilidade de ancestralidade, o que sugere poucos eventos de miscigenação. Já as populações da FIOCRUZ COVID-19 e Mosaico, por se tratarem de populações de indivíduos exclusivamente brasileiras, demonstraram alta variabilidade genética de ancestralidade em seus indivíduos, o que já era esperado, pois trata-se de uma população na qual ocorrem diversos eventos de miscigenação.

CAPÍTULO 2. MISCIGENAÇÃO: FORMALIZAÇÃO E SIMULAÇÕES

2.1 Introdução

Uma vez com as populações parentais observadas através da estruturação populacional (via PCA) e ancestralidade populacional (via ADMIXTURE), como visto ao longo do Capítulo 1, podemos realizar a inferência da ancestralidade cromossômica local com maior precisão. A partir dela é possível inferir a dinâmica de miscigenação, já que os tamanhos e a quantidade de fragmentos de ancestralidade contínua, *tracts*, estão relacionados ao momento em que a miscigenação ocorreu. Quanto menor for o *tract*, mais antigo foi o evento de miscigenação. Isso ocorre devido ao *crossing-over*, que com o passar das gerações acaba fragmentando os *tracts*, tornando-os cada vez menores, formando assim um mosaico de ancestralidades (Figura 20).

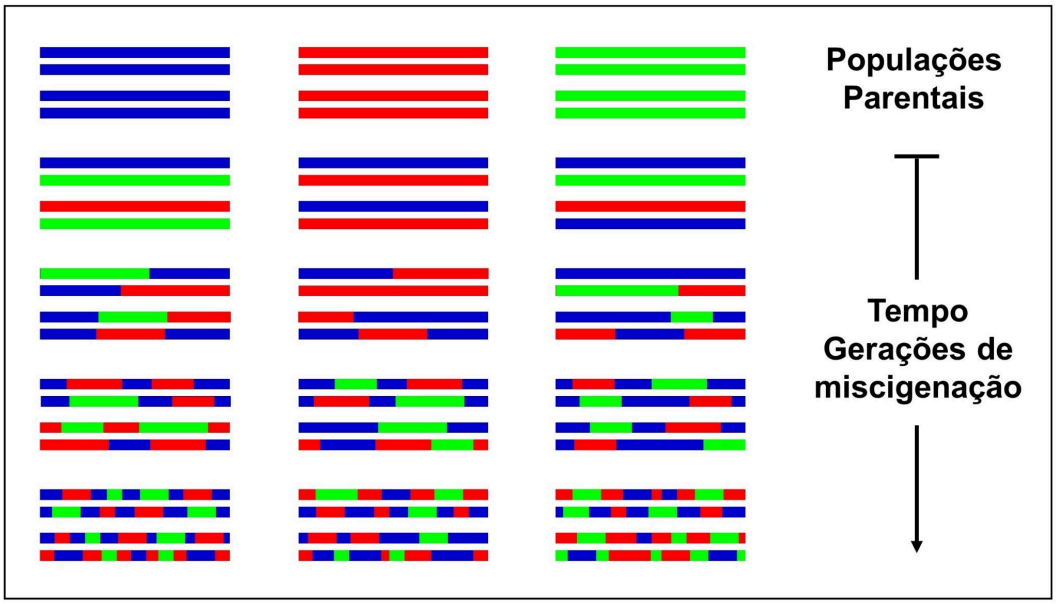

Figura 20. Padrão esquemático dos fragmentos cromossômicos de ancestralidade ao longo das gerações em uma população miscigenada. Cada cor representa uma origem ancestral distinta. Devido ao processo de recombinação, os cromossomos que na primeira geração eram completos de uma única ancestralidade se quebram e formam pequenos fragmentos de ancestralidades diferentes. Fonte: Tarazona-Santos *et al.* 2016.

Com base na distribuição de *tracts* de ancestralidade da dados reais (similares a Figura 12 e Figura 19), Kehdy *et al*. 2015 propuseram uma metodologia fundamentada no modelo bayesiano para a inferência da dinâmica de miscigenação de uma população miscigenada pós colonização das Américas nos últimos 500 anos (20 gerações de 25 anos cada) a partir das distribuições de fragmentos das diferentes ancestralidades. Essa metodologia faz uso do processo estocástico implementado por Liang e Nielsen 2014 através do algoritmo *multipulses*. Essa abordagem visa inferir os parâmetros que definem a história evolutiva das populações alvo, comparando sua distribuição observada de *tracts* com distribuições de segmentos de ancestralidade gerados por simulações de cenários de miscigenação. O modelo histórico-demográfico de Kehdy *et al*. 2015 possui nove parâmetros a estimar: três correspondentes a ancestralidades a cada três pulsos de miscigenação que ocorreram em momentos diferentes. Cada pulso possui três possíveis proporções de imigrantes (m) para cada população parental: EUR que corresponde à população Europeia, AFR que correspondente à Africana e NAT que corresponde à Nativo Americana. Dessa forma, para cada pulso, teremos um proporção de migração de cada população parental (Figura 21).

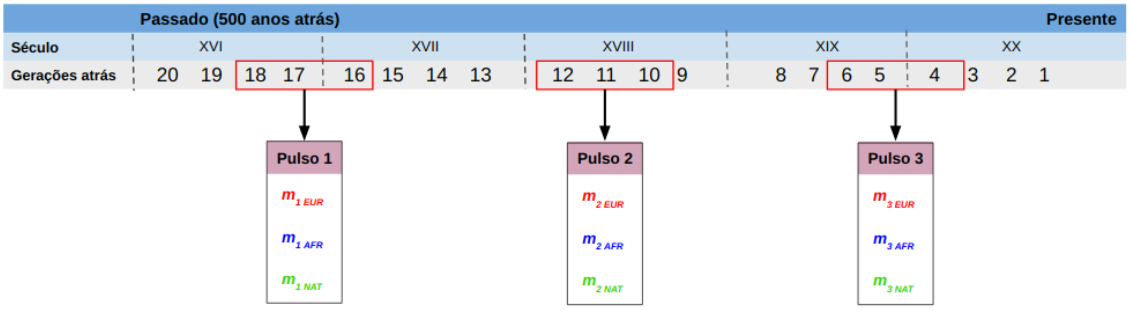

Figura 21. Modelo demográfico utilizado por Kehdy *et al.* 2015. Definição dos pulsos de migração ao longo do tempo e dos parâmetros associados a cada pulso na aplicação do método nos cromossomos autossômicos. Figura adaptada de Kehdy *et al.* 2015.

Dessa forma, ao aplicarmos esta metodologia a dados reais, como os presentes no Capítulo 1 deste texto, podemos descrever, estatisticamente, a dinâmica de miscigenação de uma população. Porém a estratégia adotada por Kehdy *et al*. 2015 mesmo solucionando a questão proposta possui limitações e necessita de validação estatística.

2.2 Objetivos

2.2.1 Avaliar a capacidade do modelo de Liang-Nielsen produzir distribuições de tamanhos de *tracts* diferentes a partir de cenários de miscigenação diferentes.

2.2.2 Dedução da fórmula teórica, $M_{a,g}$, para inferência da ancestralidade populacional a partir de cenários teóricos de miscigenação

2.2.3 Avaliar a dinâmica da ancestralidade populacional ($M_{a,g}$) em função dos diferentes cenários de miscigenação a partir da fórmula teórica.

2.2.4 Verificar se as proporções de ancestralidade populacional final estimadas a partir dos *tracts* simulados pelo modelo de Liang e Nielsen são similares àquelas modeladas com a fórmula $M_{a,g}$

2.3 Liang-Nielsen e Cenários

Em Liang e Nielsen 2014, foram apresentados vários modelos para analisar a distribuição de comprimentos de *tracts* de ancestralidade contínua em populações que passaram por eventos de miscigenação ao longo das gerações. Ao realizarem simulações para testar a precisão e a eficiência de cada modelo em diferentes cenários de mistura de populações, foi observado que outros métodos propostos anteriormente são precisos e eficientes em diferentes cenários de miscigenação, mas não para eventos recentes de miscigenação. Dessa forma é proposto um modelo estocástico, baseado em intervalos diádicos (intervalos do tipo

$I_{j,k} = [2^{-j}k, 2^{-j}(k+1))$ sendo $j, k \in Z$), para gerar *tracts* de ancestralidade em populações com miscigenação recente.

Para caracterizar (descrever suas propriedades e comportamentos em termos de probabilidades e incertezas) o processo estocástico, pelo qual segmentos de material genético ancestral são recombinados a fim de formar o genoma de um indivíduo de interesse (*proband*), foi proposto um modelo de processo de cópia de ancestralidade *(ancestor-copying process*), que representa a linha de descendência do genoma do *proband* como uma função da posição genômica. Dessa forma, ao definir os pais de um indivíduo, a sua ancestralidade em uma posição específica no genoma é então determinada pelas escolhas dos pais esquerdo e direito no passado na árvore genealógica (Figura 22).

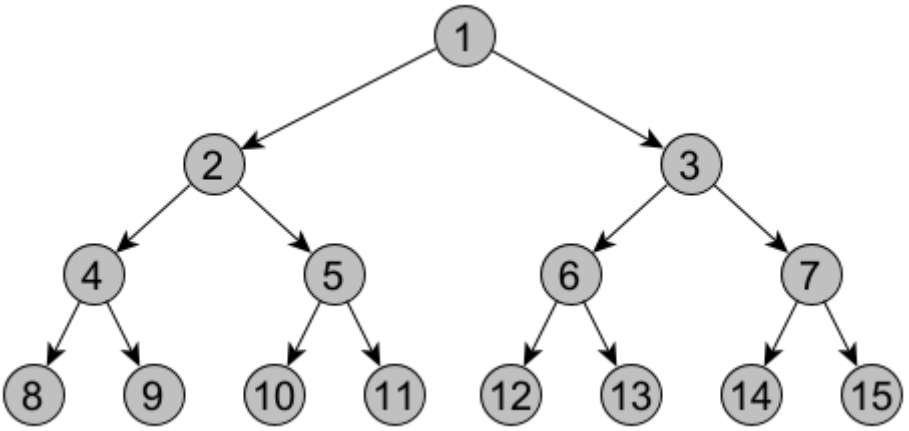

Figura 22. Árvore Binária perfeita Modelo de representação em grafos de uma árvore binária perfeita. Chamamos de Pais a esquerda e Pais a direita os nós cujos valores estão contidos nos seguintes conjuntos, respectivamente: {2,4,5,8,9,10,11} e {3,6,7,12,13,14,15}.

O modelo proposto por Liang e Nielsen 2014 (simulador) representa a estrutura genealógica de uma amostra em uma árvore binária perfeita (*perfect binary tree mode*l), Figura 22, em que a história genealógica pode ser representada em termos de intervalos diádicos e aproximar as distribuições de comprimentos de *tracts* de ancestralidade em populações com miscigenação recente (Figura 23).

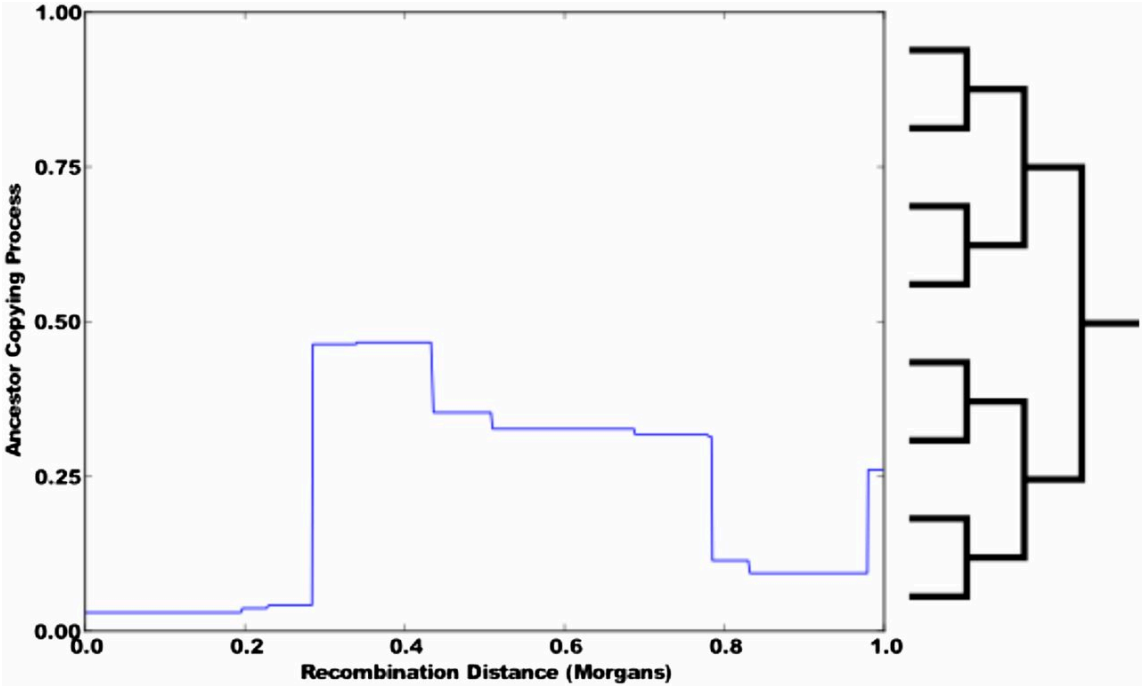

Figura 23. Processo de cópia de ancestralidade. Uma realização do processo de cópia ancestral. Nesse caso, o processo fica no intervalo $[1, \frac{1}{2})$, indicando que esse comprimento do cromossomo foi herdado inteiramente do pai esquerdo do *proband*. O processo salta entre $[1, \frac{1}{4})$ e $[\frac{1}{4}, \frac{1}{2})$ três vezes, indicando que cada avô esquerdo contribuiu com dois blocos para o *proband.* O *pedigree*, até os oito bisavós do *proband,* é mostrado à direita. Cada ancestral foi colocado em seu intervalo diádico correspondente. Figura de Liang e Nielsen 2014.

Para compreender o funcionamento do simulador de Liang-Nielsen, baseado em árvores binárias perfeitas, tome uma árvore binária perfeita de profundidade 4 e 15 nós (Figura 24A):

- nível 1 possui 1 nó (a raiz) e possui um pai/mãe
- nível 2 com 2 nós e cada nó possui um pai/mãe
- nível 3 com 4 nós e cada nó possui um pai/mãe
- nível 4 com 8 nós e nenhum deles possui pai/mãe

Para exemplificar o funcionamento do simulador de Liang e Nielsen 2014 (chamaremos de Liang-Nielsen), tome um cenário de fluxo migracional da seguinte forma: durante geração 1 ocorre um pulso migracional de 0.3 da população Vermelha, na geração 2 teremos um pulso de 0.1 da geração Verde e por fim um fluxo migracional de 1 da população Azul na geração 3 (Figura 24B). Seguiremos o mesmo sentido de temporalidade de gerações apresentados na Figura 21: geração 1 mais recente e geração 3 a mais antiga.

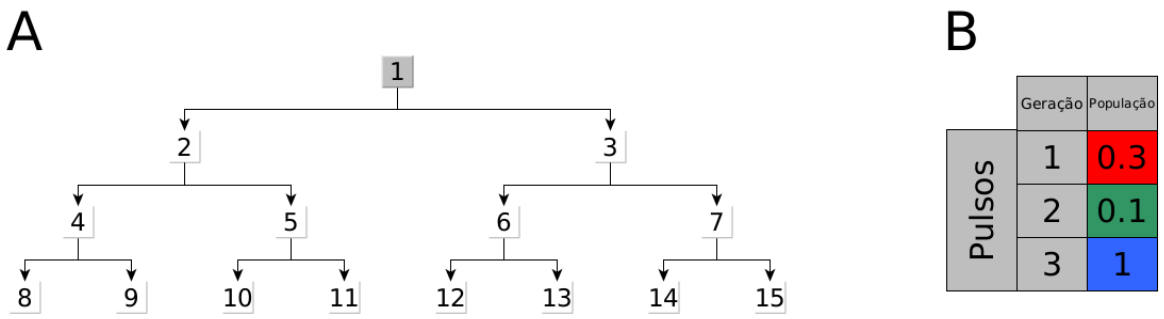

Figura 24. Árvore binária perfeita e pulsos. O modelo de árvore binária perfeita. **B** fluxo migracional das populações Vermelho, Verde e Azul durante 3 gerações distintas. O indivíduo 1 (cor cinza) é chamado de miscigenado.

Para o exemplo (Figura 25), escolhemos um cromossomo hipotético de tamanho de 5 Morgans. Tomamos o pai/mãe à esquerda da raiz (Figura 25A) e geramos um número aleatório, U(n), pertencente a uma distribuição uniforme U de intervalo (0,1). Caso o valor do pulso migracional, V(p) (Figura 24B), da geração seja menor que U(n), esse pai/mãe assumirá o papel de uma população parental e todos os nós abaixo daquele nível serão apagados. Caso U(n)>V(p), aquele pai/mãe assume papel de não-parental e continuaremos descendo a árvore, tomando o avô/avó à esquerda e repetindo o processo descrito para os níveis mais altos da árvore até encontrar um indivíduo parental. Ao encontrarmos um nó tido como parental, paramos o processo, e a partir de uma Exp(T), onde T é a geração em que o pulso analisado ocorreu, geramos o primeiro tamanho de *tract* de ancestralidade para aquela população parental. Voltamos ao nó raiz e tomamos o pai/mãe à direita. Caso U(n)<V(p), aquele pai/mãe é definido como parental e os nós abaixo dele são apagados, caso U(n)>V(p), definimos esse pai/mãe como não-parental e continuamos descendo a árvore tomando o avô/avó a esquerda e repetindo o processo para os níveis mais altos até encontrar um indivíduo parental. Ao encontrar uma parental que parte do pai/mãe à direita, paramos o processo, geramos uma Exp(T) e teremos o valor do *tract* de ancestralidade daquela população que teve seu pulso migracional. Somamos esse tamanho de *tract* ao primeiro que foi encontrado no pai/mãe esquerdo teremos o tamanho total até aquele momento. Finalizando a corrida pai/mãe esquerdo e pai/mãe direito, voltamos ao nó e somamos ao tamanho total de tracts o primeiro tamanho encontrado no parental inicial e podemos voltar a descida da árvore repetindo o processo sempre que encontramos um novo indivíduos parental.

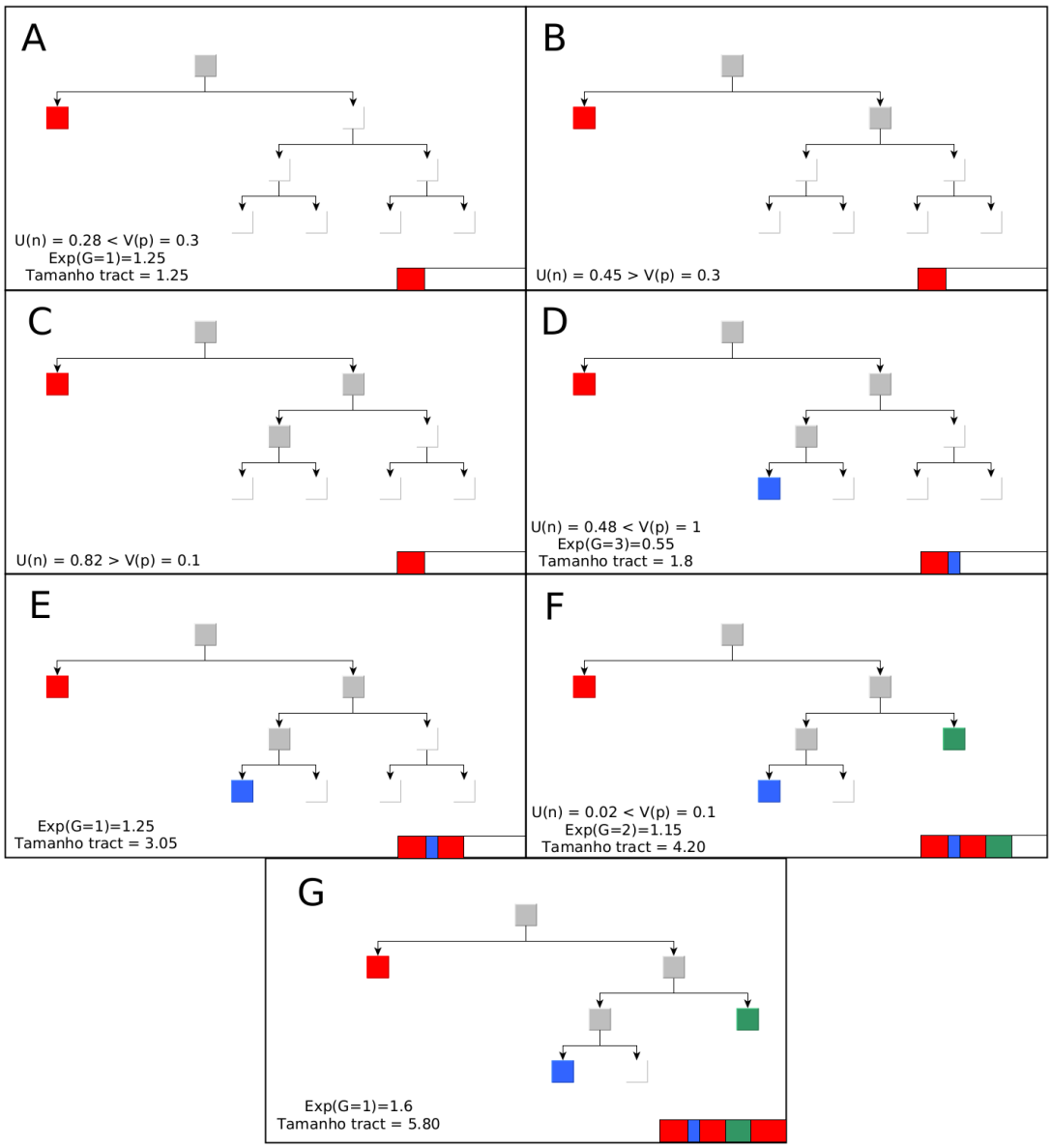

Figura 25. Exemplo da execução do simulador de Liang e Nielsen 2014 para árvores binárias perfeitas para um cromossomo hipotético de 5 Morgans. Para a árvore da Figura 24A, U(n) será a o valor aleatório de uma distribuição uniforme de intervalo (0,1) e V(p) o valor do pulso referente a Figura 24B. **A** Começamos a descida da árvore para o pai/mãe a esquerda e temos U(n) < P(v), logo esse pai/mãe torna-se parental para Vermelho (Figura 25B), calculamos Exp(T) para encontrar o valor do *tract* da ancestralidade Vermelho.Todos os nós abaixo deles são apagados. **B** Voltamos o nó raiz e começamos a descida para pai/mãe a direita, onde temos que U(n) > P(v) o que nos força a continuar descendo a árvore para o avô/avó a esquerda. **C** Novamente, temos que U(n) > P(v), forçando a descida. **D** Para o bisavô/bisavó a esquerda temos que U(n) < P(v) tornando esse indivíduo parental para Azul, calculamos Exp(T) para encontrar o valor do *tract* da ancestralidade Azul, acumulando com o valor de tract encontrado anteriormente. **E** Voltamos ao nó raiz, calculamos Exp(T) para encontrar o valor do *tract* da ancestralidade Vermelho e acumulando ao resultado passado. **F** Retomamos a descida da árvore agora pelo lado do pai/mãe a direita e avô/avó a direita, temos que U(n) < P(v) o que define esse indivíduo como parental para população Verde, calculamos Exp(T) para encontrar o valor do *tract* da ancestralidade Verde e somamos ao tamanho total do cromossomo até o momento. Assim, os nós abaixo deste são apagados. **G** Mais uma vez voltamos ao nó raiz, calculamos Exp(T) para encontrar o valor do *tract* da ancestralidade Vermelho e acumulando ao resultado passado. Como o tamanho do cromossomo hipotético é de 5 Morgans, neste momento paramos o processo pois já atingimos o valor máximo.

Para aplicação do simulador de Liang-Nielsen, que considera pulsos migracionais, definimos um modelo hipotético no qual três populações P, Q e R (baseando-se no modelo Latino Americano Tri-híbrido) coexistem e interagem entre si ao longo de 20 gerações (Figura 26). Neste contexto, definimos a ocorrência de três pulsos migratórios, cada um compreendendo intervalos de três gerações, durante os quais cada população ancestral contribui de acordo com seu respectivo grau de participação. Cada contribuição de um pulso migratório (m) é expressa como um valor variando de 0 a 1, assegurando, portanto, que a soma das contribuições das populações durante um determinado pulso não exceda o limite máximo de 1. Assim, $m_{i,g}$ é a proporção de imigrantes de ancestralidade i (P, Q ou R) chegados na geração g.

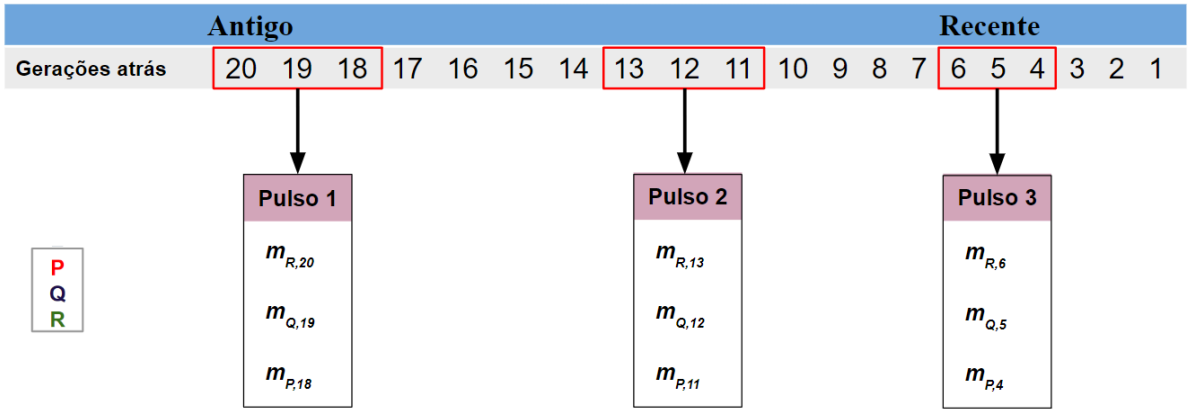

Figura 26. **Gerações e Pulsos**. Para este modelo, utilizamos três populações (P, Q e R) que coexistem e interagem entre si ao longo de 20 gerações, sendo a geração 20 a mais antiga e a geração 1 a mais recente, baseando-se no modelo demográfico apresentado por *Kehdy et al. 2015*. Três pulsos migracionais ocorrem, no qual cada um deles acontece em um bloco de três gerações. As gerações escolhidas foram: 4, 5 e 6 (Pulso Recente); 11, 12 e 13 (Pulso Intermediário); 18, 19 e 20 (Pulso Antigo). Ao longo de cada pulso, as populações P, Q e R contribuem com um fluxo migracional *m* em uma geração específica.

Tais gerações foram escolhidas devido a história demográfica de migração para o brasil ao longo dos últimos 500 anos. Para observar o comportamento das simulações, definimos sete cenários em que os pulsos migratórios se diferem entre si pelo momento em que o pulso ocorre e pela proporção de contribuição de cada população (Tabela 9). A População parental Q é fixada como aquela que terá o peso do fluxo migracional maior que as demais durante o pulso (Tabela 9, em azul). Os valores 0.8, 0.4 e 0.1 significam, respectivamente, migração intensa, intermediário e pequena. A geração 21 será aquela na qual apenas a fundadora existirá com peso 1. Para esse modelo escolhemos a P como a fundadora, portanto todo fluxo migratório será 0 para as populações parentais nessa geração.

Tabela 9. Cenários definidos para serem calculados na fórmula

	Pop. Fundadora Geração 21	Pulso 1			Pulso 2			Pulso 3		
População	**P**	**R**	**Q**	**P**	**R**	**Q**	**P**	**R**	**Q**	**P**
Cenário 1 (PA1)	1	0.1	**0.8**	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Cenário 2 (PI1)	1	0.1	0.1	0.1	0.1	**0.8**	0.1	0.1	0.1	0.1
Cenário 3 (PR1)	1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	**0.8**	0.1
Cenário 4 (PA1)	1	0.1	**0.4**	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Cenário 5 (PI1)	1	0.1	0.1	0.1	0.1	**0.4**	0.1	0.1	0.1	0.1
Cenário 6 (PR1)	1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	**0.4**	0.1
Cenário 7 (LU3)	1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

As siglas PA1, PI1, PR1 e LU3 simbolizam, em ordem, *Predominant Ancient Admixture from 1 Population* (miscigenação predominantemente antiga de uma população), *Predominant Intermediate Admixture from 1 Population* (miscigenação predominantemente intermediária de uma população), *Predominant Recent Admixture from 1 Population* (miscigenação predominantemente recente de uma população) e *Low Uniform Admixture from 3 Populations* (miscigenação baixa e uniforme de 3 populações). O Pulso 1 compreende as gerações 20,19 e 18. O Pulso 2 compreende as gerações 13,12 e 11. O Pulso 3 compreende as gerações 6,5 e 4.

Cada cenário foi simulado 100 vezes, porém o simulador leva em consideração o tamanho (em Morgans) de cada cromossomo. Assim, simulamos para cada um dos 22 cromossomos autossômicos os mesmos cenários, o que nos deixou com um total de 15.400 dados simulados observados. Para isso foi necessária a automatização da execução das simulações (Figura 27).

Automatização de Simulações

Inputs

```
INPUT1: Cenários de migração
INPUT2: Valores de recombinação por cromossomo
INPUT3: Numero de Simuçaões
INPUT4: Nome para output
INPUT5: Simulador de Liang e Nielsen 2014
```

```
#função para executar comandos do sistema operacional
FUNÇÃO executa(x):
    IMPRIME x com uma nova linha ("\n")
    EXECUTAR comando do sistema 'x'


#cria pasta baseado na quantidas de cenários para serem executados onde serão salvas as simulações feitas
de cada cenários
FUNÇÃO criar_pastas(w, z):
    PARA i de 1 ATÉ o tamanho de w:
        a = CONSTRUIR uma string 'mkdir {INPUT4}/CHR{i}'
        executa(a)
        PARA j de 1 ATÉ o tamanho de z:
            b = CONSTRUIR uma string 'mkdir {INPUT4}/CHR{i}/M{j}'
            executa(b)


#lê um arquivo salvando suas linhas como itens de um vetor
FUNÇÃO ler_arquivo(y):
    itens = lista vazia
    ABRIR arquivo 'y' em modo de leitura ('r') como f
    linhas = LER todas as linhas do arquivo 'f'
    PARA cada linha 'i' em linhas:
        ADICIONAR 'i' com espaços em branco à direita removidos à lista 'itens'
    FECHAR arquivo 'f'
    RETORNAR itens

cenarios = ler_arquivo(INPUT1)
cromossomos = ler_arquivo(INPUT2)

recombinacao = lista vazia
PARA cada i em cromossomos:
    ADICIONAR i dividido por espaços à lista 'recombinacao'


valores = lista vazia
PARA i de 1 ATÉ o tamanho de cenarios:
    ADICIONAR i à lista 'valores'

criar_pastas(recombinacao, cenarios)


#executa o simulador de Liang-Nielsen salvando baseando no número de cenários e simulações e salvando nas
pastas referentes
PARA cada i em recombinacao:
    ch = i[0]
    PARA cada j, k EMPARELHADOS com cenarios, valores:
        PARA p de 1 ATÉ INPUT3:
            comando = CONSTRUIR uma string 'time python3 {INPUT5} {j} -c {i[0]} -r {i[1]} >
{INPUT4}/CHR{i[0]}/M{k}/sim{p}.txt'
            executa(comando)
```

Figura 27. Fluxograma do *script* de automatização. Pseudocódigo do processo de automatização da execução do simulador de Liang-Nielsen.

2.3.1 Fórmula Teórica de Ancestralidade Populacional

Propomos e implementamos uma equação de primeiro grau que, através das proporções das contribuições de cada população parental ao longo de pulsos migracionais ($m_{i,g}$), é capaz de descrever a proporção final, ou proporção final esperada, de cada população parental ($M_{a,g}$) (Figura 28) ao longo de gerações.

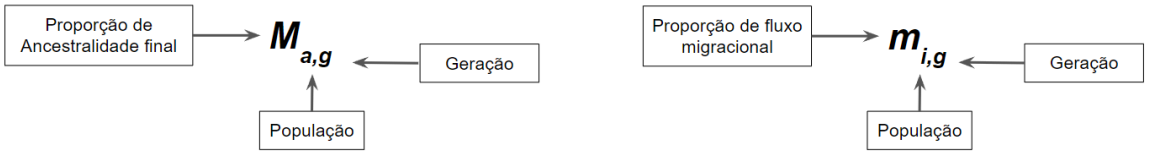

Figura 28. Proporção final esperada e pulso migracional. Para cada população proposta em um modelo, teremos o seu $M_{a,g}$ dado em função do seu $m_{i,g}$.

Seja $M_{a,g}$ a proporção esperada (proporção final) da população ancestral a na geração g e $m_{i,g}$ a proporção de imigrantes de ancestralidade i chegados na geração g, em uma dada população. Quando não há fluxo migracional de nenhuma população, repete-se os $M_{a,g}$ da geração passada. Assim temos:

$$M_{a,g} = \begin{cases} a = i & M_{a,(g+1)} - m_{i,g} \cdot M_{a,(g+1)} + m_{i,g} \\ a \neq i & M_{a,(g+1)} - m_{i,g} \cdot M_{a,(g+1)} \end{cases} \quad (1)$$

$$a, i \in \{Populações\}$$
$$g \in [geração\,recente,\ geração\,antiga]$$
$$M_{a,g} \in \mathbb{R}^+ | \, M_{R,g} + M_{Q,g} + M_{P,g} = 1$$
$$m_{i,g} \in \mathbb{R}^+ \mid m_{i,g} < 1$$

Para exemplificar, tome uma população N que segue um modelo tri-híbrido formada por três populações ancestrais P, Q e R de proporções 0.8, 0.1 e 0.1 na geração 2, respectivamente. Ou seja: $M_{P,2} = 0.8$, $M_{Q,2} = 0.1$ e $M_{R,2} = 0.1$ Sendo a geração 2 mais antiga que a geração 1 (assim como na Figura X), e considerando-se a população Q como a única população que terá o fluxo migracional com peso 0.1 na geração 1 ($m_{Q,1} = 0.1, m_{P,1} = 0$ e $m_{R,1} = 0$), a proporção final das populações ancestrais após o fluxo migracional na geração 1 ($M_{a,1}$) será dada por:

$$M_{P,1} = \{P \neq Q \quad M_{P,2} - m_{Q,1} \cdot M_{P,2}$$
$$M_{P,1} = \{0,8 - (0,1 \cdot .0,8)$$
$$M_{P,1} = 0.72$$

$$M_{Q,1} = \{P = Q \quad M_{Q,2} - m_{Q,1} \cdot M_{Q,2} + m_{Q,1}$$
$$M_{Q,1} = \{0,1 - (0,1 \cdot 0,1) + 0,1$$
$$M_{Q,1} = 0.19$$

$$M_{R,1} = \{R \neq Q \quad M_{R,2} - m_{Q,1} \cdot M_{R,2}$$
$$M_{R,1} = \{0,1 - (0,1 \cdot 0,1)$$
$$M_{R,1} = 0,09$$

$$M_{P,1} + M_{Q,1} + M_{R,1} = 1$$

Vale ressaltar que, para esse modelo, o valor de $m_{i,g}$ deve ser menor do que 1 pelo seguinte fato: caso $m_{i,g} = 1$, as demais populações não coexistiram naquela geração, o que não se adequaria à realidade que estamos tentando descrever.

Para observar o comportamento da Equação (1), utilizamos o mesmo modelo apresentado na Figura 26 e os mesmos cenários da Tabela 9, uma vez que posteriormente iremos comparar os resultados obtidos via Liang-Nielsen e Fórmula Teórica. Como a Fórmula dispensa o tamanho do cromossomo pelo fato de trabalhar apenas com as proporções observadas de fluxo migracional, a execução de cada cenário é feita apenas uma vez (Tabela 10).

Tabela 10. Fluxo e proporção de contribuição por geração para o Cenário 1.

Gerações	**Fluxo Migracional**		
	m_P	m_Q	m_R
21	0	0	0
20	0	0	0.1
19	0	0.8	0
18	0.1	0	0
17	0	0	0
16	0	0	0
15	0	0	0

14	0	0	0
13	0	0	0.1
12	0	0.1	0
11	0.1	0	0
10	0	0	0
9	0	0	0
8	0	0	0
7	0	0	0
6	0	0	0.1
5	0	0.1	0
4	0.1	0	0
3	0	0	0
2	0	0	0
1	0	0	0

Em azul estão os blocos de gerações que compõem um pulso migracional. Sendo as gerações 20, 19 e 18 formando o Pulso 1 (antigo), gerações 13, 12 e 11 compondo o Pulso 2 (intermediário) e as gerações 6, 5 e 4 compondo o Pulso 3 (recente). Durante a geração 21 não ocorre pulso, existindo apenas a população fundadora, P, desse modo o valor de pulso migracional é 0 para P, Q e R.

2.4 Resultados

2.4.1. Comparação das distribuições dos tamanhos dos tracts entre os cenários calculados pelos Modelo de Liang-Nielsen

O simulador de Liang e Nielsen 2014 não nos permite observar as proporções dos M's (proporções finais de P, Q e R) ao longo das gerações de uma maneira simples. Após a criação dos cenários (Tabela 9) e execução do algoritmo, recebemos os tamanho dos *tracts* finais por indivíduo para o cromossomo simulado. O *output* consiste em um arquivo de texto com quatro colunas: População Parental, Tamanho do *Tract* de Ancestralidade, Indivíduo e Cromossomo.

Ao concatenarmos todas as simulações do mesmo cenário e dividirmos os resultados em 100 *bins* e baseado-se nas distribuições dos segmentos de ancestralidade, ao analisarmos os

cenários que possuem o mesmo peso de pulso migracional, é possível observar uma diferença na densidade dos dados, indicando que cenários com o mesmo peso de pulso, mas que ocorrem em momentos diferentes, levam a distribuição de *tracts* diferentes (Figura 29). Além disso, os valores máximo e mínimo desses segmentos de ancestralidade possuem diferenças entre os conjuntos.

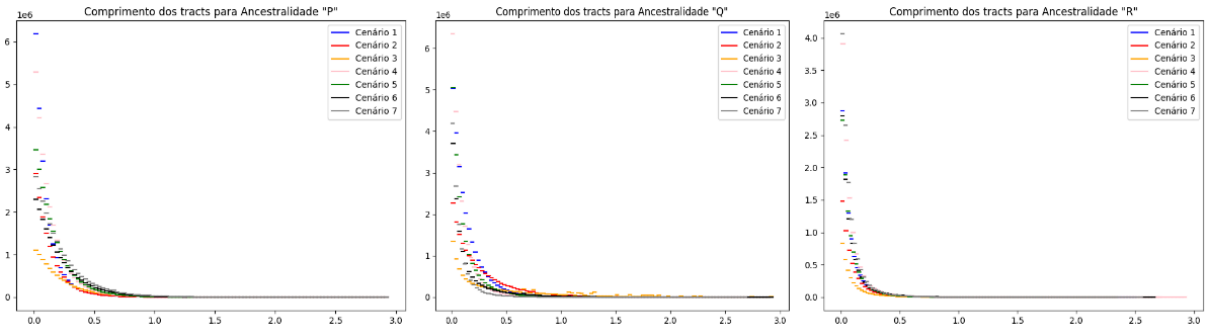

Figura 29. Distribuição dos tamanhos de *tracts* ao longo dos cenários para Q, R e P. Comparação entre a distribuição dos tamanhos dos *tracts* entre cenários em 100 *bins* para cada cenário. No eixo das abscissas está representado o tamanho do segmento de ancestralidade em Morgans enquanto na Ordenada a contagem de ocorrência em escala de *1e6*.

Ao nos concentrarmos na visualização dos últimos *bins* para cada ancestralidade (P, Q e R) podemos observar uma maior diferença na distribuição dos tracts para cada cenário. Além disso, temos que os cenários de fluxo recente possuem segmentos maiores (Figura 30).

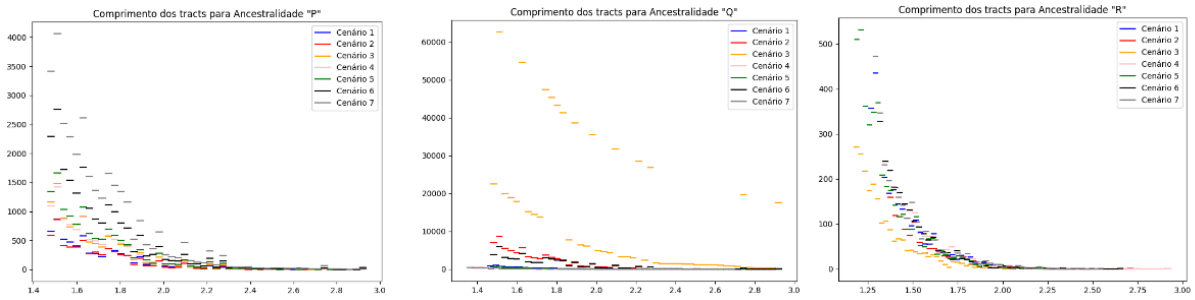

Figura 30. Distribuição dos tamanhos de *tracts* maiores que ao longo dos cenários para Q, R e P. Comparação entre a distribuição dos tamanhos dos *tracts* entre cenários entre os 50 últimos *bins*. No eixo das abscissas está representado o tamanho do segmento de ancestralidade em Morgans enquanto na Ordenada a contagem de ocorrência em escala de *1e6*.

As diferenças entre cada cenário ficam mais evidentes quando observamos as estatísticas sumárias dos valores dos segmentos de ancestralidade para cada Cenário. Os maiores fragmentos de ancestralidade são oriundos dos cenários de miscigenação recentes (Tabela 11) , enquanto os menores pertencentes aos mais antigos.

Tabela 11. Estatísticas sumárias para os tamanhos dos segmentos de ancestralidade, em Morgans, obtidos via simulador de Liang-Nielsen.

Cenários	Pop	Tamanho Médio de	Mediana	Percentil 10	Primeiro Quartil	Terceiro Quartil	Percentil 90	Desvio Padrão	Variância

		Tract							
Cenário 1	P	0.108236	0.067088	0.009932	0.027314	0.139699	0.249106	0.131756	0.017360
	Q	0.147869	0.096559	0.014156	0.039101	0.199917	0.345621	0.160703	0.025825
	R	0.090417	0.051412	0.007382	0.020474	0.111855	0.213642	0.119072	0.014178
Cenário 2	P	0.136524	0.149509	0.013972	0.063197	0.181886	0.309941	0.150804	0.022742
	Q	0.261841	0.173422	0.022540	0.065721	0.367788	0.626155	0.271630	0.073783
	R	0.107205	0.063287	0.009062	0.025197	0.136815	0.254625	0.132814	0.017640
Cenário 3	P	0.214528	0.149509	0.023514	0.063197	0.295660	0.490511	0.214161	0.045865
	Q	0.473021	0.268441	0.021705	0.072284	0.072284	1.247088	0.527911	0.278690
	R	0.111341	0.060955	0.008411	0.023637	0.139824	0.277870	0.142782	0.020387
Cenário 4	P	0.141192	0.092399	0.188638	0.038140	0.038140	0.325595	0.156221	0.024405
	Q	0.104417	0.063693	0.009229	0.025576	0.135203	0.244518	0.126444	0.015988
	R	0.087696	0.050093	0.108560	0.050093	0.020009	0.206178	0.115590	0.013361
Cenário 5	P	0.169918	0.117246	0.018617	0.049919	0.230795	0.385233	0.175153	0.030679
	Q	0.127358	0.073879	0.010063	0.028412	0.166875	0.312186	0.154309	0.023811
	R	0.092420	0.053651	0.007701	0.021399	0.116118	0.218146	0.118590	0.014064
Cenário 6	P	0.213223	0.148695	0.023508	0.063024	0.293736	0.487321	0.212595	0.045197
	Q	0.168994	0.076132	0.009673	0.027733	0.202285	0.456363	0.240852	0.058010
	R	0.097313	0.053860	0.007618	0.021239	0.119817	0.235584	0.128350	0.016474
Cenário 7	P	0.215825	0.149217	0.023553	0.063153	0.296290	0.494671	0.217580	0.047341
	Q	0.089478	0.050186	0.007197	0.019988	0.109304	0.209966	0.121366	0.014730
	R	0.085919	0.049299	0.007113	0.019705	0.106499	0.201480	0.113151	0.012803

2.4.2 Comparação de M e sua dinâmica entre os os cenários via Fórmula

A partir da definição de cada cenário (Tabela 9) construímos as proporções de ancestralidade por geração baseada no pulso de cada cenário, por exemplo o Cenário 1 na Tabela 12. Dessa forma acompanhamos as variações das populações P, Q e R ao longo das gerações.

Tabela 12. Distribuição da proporção das populações por geração do Cenário 1 aplicado a fórmula teórica.

Cenário 1

Gerações	População			Fluxo Migracional (m)		
	P	Q	R	m_P	m_Q	m_R
21	1	0	0	0	0	0
20	0.9	0	0.1	0	0	**0.1**
19	0.18	0.8	0.02	0	**0.8**	0
18	0.262	0.72	0.018	**0.1**	0	0
17	0.262	0.72	0.018	0	0	0
16	0.262	0.72	0.018	0	0	0
15	0.262	0.72	0.018	0	0	0
14	0.262	0.72	0.018	0	0	0
13	0.2358	0.648	0.1162	0	0	**0.1**
12	0.21222	0.6832	0.10458	0	**0.1**	0
11	0.290998	0.61488	0.094122	**0.1**	0	0
10	0.290998	0.61488	0.094122	0	0	0
9	0.290998	0.61488	0.094122	0	0	0
8	0.290998	0.61488	0.094122	0	0	0
7	0.290998	0.61488	0.094122	0	0	0
6	0.2618982	0.553392	0.1847098	0	0	**0.1**
5	0.23570838	0.5980528	0.16623882	0	**0.1**	0
4	0.312137542	0.53824752	0.149614938	**0.1**	0	0
3	0.312137542	0.53824752	0.149614938	0	0	0
2	0.312137542	0.53824752	0.149614938	0	0	0
1	0.312137542	0.53824752	0.149614938	0	0	0

Ao comparar Cenários como o 1, 2 e 3 (Figura 31A, 31B e 31C, respectivamente) podemos observar que pulsos migracionais de mesmo peso resultam em Ms finais (proporção final de ancestralidade)(Tabela 13) diferentes. O mesmo pode ser observado para as comparações de

Cenário 4, 5 e 6. Uma vez que os cenários do tipo PA (miscigenação predominantemente antiga de uma população) possuem um tempo maior de interação das populações P e R após o maior pulso de Q, teremos uma proporção de M final da ancestralidade Q menor quando comparamos cenários do tipo PR (miscigenação predominantemente recente).

Tabela 13. Proporção final (M's finais) para P, Q e R através da fórmula.

	M Final		
	P	**Q**	**R**
Cenário 1	0.312137542	0.53824752	0.149614938
Cenário 2	0.270803242	0.62550882	0.103687938
Cenário 3	0.214103242	0.74520882	0.040687938
Cenário 4	0.484324426	0.34692876	0.168746814
Cenário 5	0.466609726	0.38432646	0.149063814
Cenário 6	0.442309726	0.43562646	0.122063814
Cenário 7	0.613464589	0.20343969	0.183095721

O decrescimento da população parental P acontece mais rapidamente em cenários com pulsos maiores de migração. Eventos migracionais mais recentes possuem uma grande influência no presente, mas ao observarmos o Cenário 1 e Cenário 6 (Figura 31A e 31F) vemos que o peso de um fluxo migracional maior afeta mais as proporções da população parental fundadora do que o momento em que ele ocorreu.

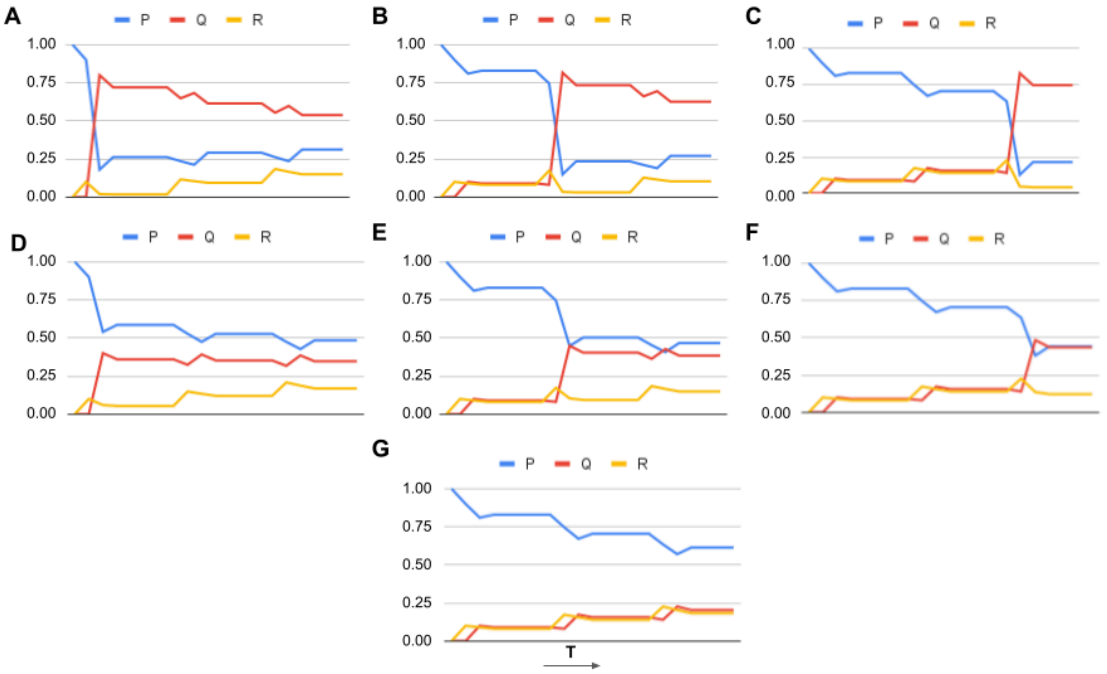

Figura 31. Proporções de ancestralidade ao longo das gerações por cenário via fórmulas. Para cada cenário é possível acompanhar a distribuição das proporções de ancestralidade (como apresentado na Tabela 11). **A** Cenário 1, **B** Cenário 2, **C** Cenário 3, **D** Cenário 4, **E** Cenário 5, **F** Cenário 6, **G** Cenário 7. O eixo horizontal representa a passagem de tempo (gerações) e no eixo vertical as proporções de ancestralidade.

O Cenário 7 (que apresenta todos os pulsos migracionais de mesmo peso), mantém a população fundadora com maior proporção final de M. Além disso, Q e R possuem proporções diferentes (0.20 > 0.18) devido ao fato da ordem em que os pulsos ocorrem. Mudando a ordem em que as populações aparecem (momento em que ocorre o pulso), a desigualdade se inverte.

Os analisarmos as diferenças entre os M finais de Q, observamos que $\Delta Q(Cenário3 Cenário1) > \Delta Q(Cenario6 Cenário4)$, 0.2069613>0.0886977, revelando que um peso maior de fluxo migracional afeta mais as proporções finais de Q do que propriamente dito o momento em que ele ocorreu.

Caso durante o bloco de maior pulso de Q (seja ele PA1, PI1 ou PR1) o peso dos pulsos de P e R fossem 0 e mantidos da mesma forma nos demais blocos de pulsos (0.1), também pudemos observar cenários diferentes produzindo Ms finais diferentes (Tabela 14)(Figura 32).

Tabela 14. Proporção final (Ms finais) para P, Q e R através da fórmula com peso 0 de P e R.

	M Final		
	P	**Q**	**R**
Cenário 1	0.2791882	0.5807628	0.140049

Cenário 2	0.2208682	0.686322	0.0928098
Cenário 3	0.1408682	0.831122	0.0280098
Cenário 4	0.4917646	0.3681864	0.140049
Cenário 5	0.4626046	0.420966	0.1164294
Cenário 6	0.4226046	0.493366	0.0840294

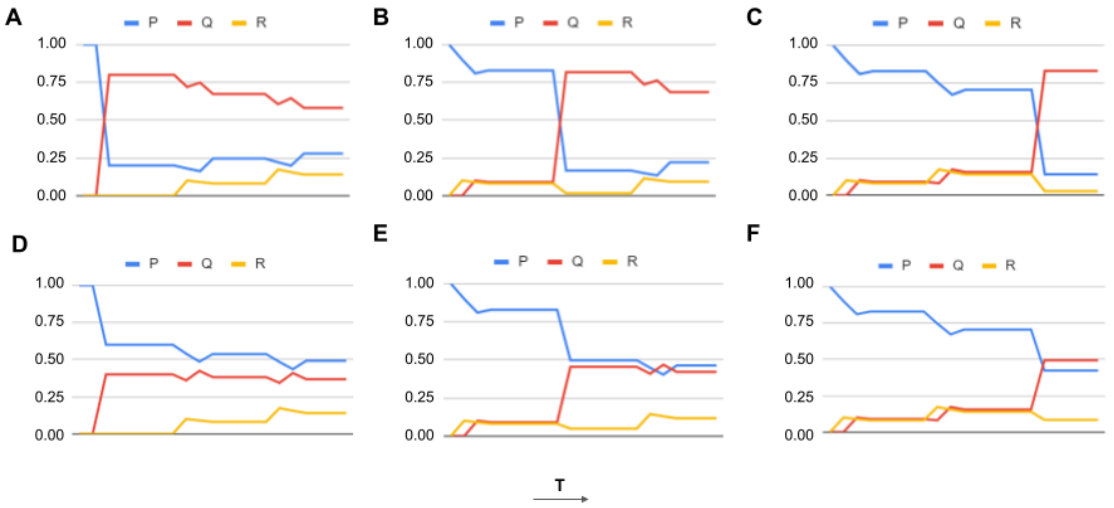

Figura 32. Proporções de ancestralidade ao longo das gerações com P e R iguais a 0 no bloco de maior pulso de Q. Durante o bloco de maior pulso de Q, as populações P e R possuem peso de pulso migracional igual a 0. **A** Cenário 1, **B** Cenário 2, **C** Cenário 3, **D** Cenário 4, **E** Cenário 5, **F** Cenário 6. O eixo horizontal representa a passagem de tempo (gerações) e no eixo vertical as proporções de ancestralidade.

Caso exista somente o pulso de Q durante o seu seu bloco de migração e os demais sejam 0 para todos os outros, observarmos que os cenários de mesmo peso de pulso, geram Ms finais iguais. Ou seja, em ausência de fluxo migracional das demais populações, o momento do pulso não importa, apenas o seu peso será capaz de variar os Ms finais (Tabela 15)(Figura 33). Mostrando que a fórmula não é capaz de gerar proposições falsas de ancestralidade caso não haja migração daquela população que não migrou.

Tabela 15. Proporção final (Ms finais) para P, Q e R através da fórmula com peso 0 para todos os pulsos.

	M Final		
	P	**Q**	**R**
Cenário 1	0.2	0.8	0
Cenário 2	0.2	0.8	0
Cenário 3	0.2	0.8	0

Cenário 4	0.6	0.4	0
Cenário 5	0.6	0.4	0
Cenário 6	0.6	0.4	0

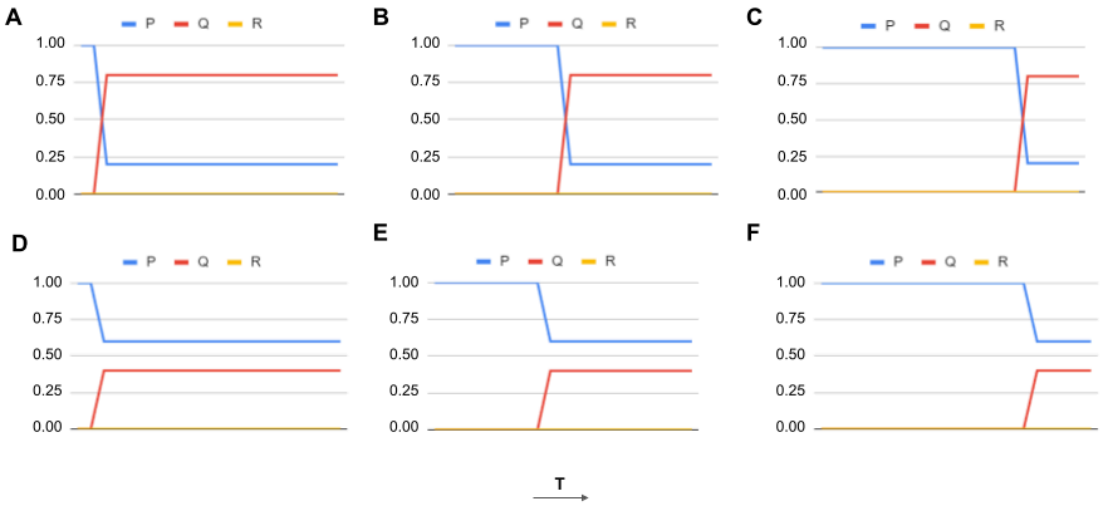

Figura 33. Proporções de ancestralidade ao longo das gerações com fluxo 0 de P e R em todos blocos de pulsos. Durante o bloco de maior pulso de Q, as populações P e R possuem peso de pulso migracional igual a 0. **A** Cenário 1, **B** Cenário 2, **C** Cenário 3, **D** Cenário 4, **E** Cenário 5, **F** Cenário 6. O eixo horizontal representa a passagem de tempo (gerações) e no eixo vertical as proporções de ancestralidade.

2.4.3 Comparação das proporções finais de ancestralidade, M, entre os cenários calculados via Fórmula e via modelo de Liang-Nielsen

Como mencionado anteriormente, o simulador de Liang e Nielsen 2014 não nos permite observar as proporções dos M's (proporções de P, Q e R) ao longo das gerações de uma maneira simples. Após a criação dos cenários e execução do algoritmo, recebemos os tamanho dos *tracts* finais por indivíduo para o cromossomo simulado. Somando todos os tamanhos de *tracts* de ancestralidade e calculando as proporções de P, Q e R, por cenário, chegamos aos valores de M's finais (Tabela 16), assim como os resultados obtidos através da fórmula.

Tabela 16. Proporção final (M's finais) para as populações P, Q e R através do simulador

	M Final (Simulador)		
	Ps	**Qs**	**Rs**
Cenário 1	0.3477454582	0.5309493032	0.1213052386

Cenário 2	0.2741974184	0.6439877376	0.08181484404
Cenário 3	0.2664020262	0.6821495133	0.05144846052
Cenário 4	0.517899811	0.3435086538	0.1385915352
Cenário 5	0.5261418241	0.3456733943	0.1281847815
Cenário 6	0.5450495286	0.3303271204	0.124623351
Cenário 7	0.6827165797	0.1619429669	0.1553404533

Comparando os M's finais via Simulador (Tabela 16) com os M's finais via fórmula (Tabela 13), é possível observar que a variação mais comum é dada pelo fato da fórmula gerar um valor de M final maior que o simulador (Tabela 17). Quanto mais próximo de 0, menor a diferença.

Tabela 17. Variação da proporção dos M's finais via Simulador e via Fórmula.

Média Simulador - Fórmula			
	ΔP	**ΔQ**	**ΔR**
Cenário 1	0.0356079162	-0.0072982168	-0.0283096994
Cenário 2	0.0033941764	0.0184789176	-0.02187309396
Cenário 3	0.0522987842	-0.0630593067	0.01076052252
Cenário 4	0.033575385	-0.0034201062	-0.0301552788
Cenário 5	0.0595320981	-0.0386530657	-0.0208790325
Cenário 6	0.1027398026	-0.1052993396	0.002559537
Cenário 7	0.0692519907	-0.0414967231	-0.0277552677

Os cenários 3 e 6 (PR1 de pesos 0.8 e 0.4, respectivamente)(Figura 34C e 34F) são os que apresentam maior diferença entre as proporções de dos M's finais. Os cenários do tipo PI1 apresentam diferenças maiores que os cenários PA1. Essa maior diferença para eventos de miscigenação recente pode estar associado com os efeitos de recombinação que possuem uma maior força que eventos antigos para alterar as proporções dos *tracts*. Além disso, é dito em Liang e Nielsen 2014 que o modelo baseado em Árvore Binária Perfeita pode acabar inferindo tamanhos de segmentos de ancestralidade menores que o esperado. Apesar disso, os resultados via formulá se aproximam do obtidos via simulador (Figura 34).

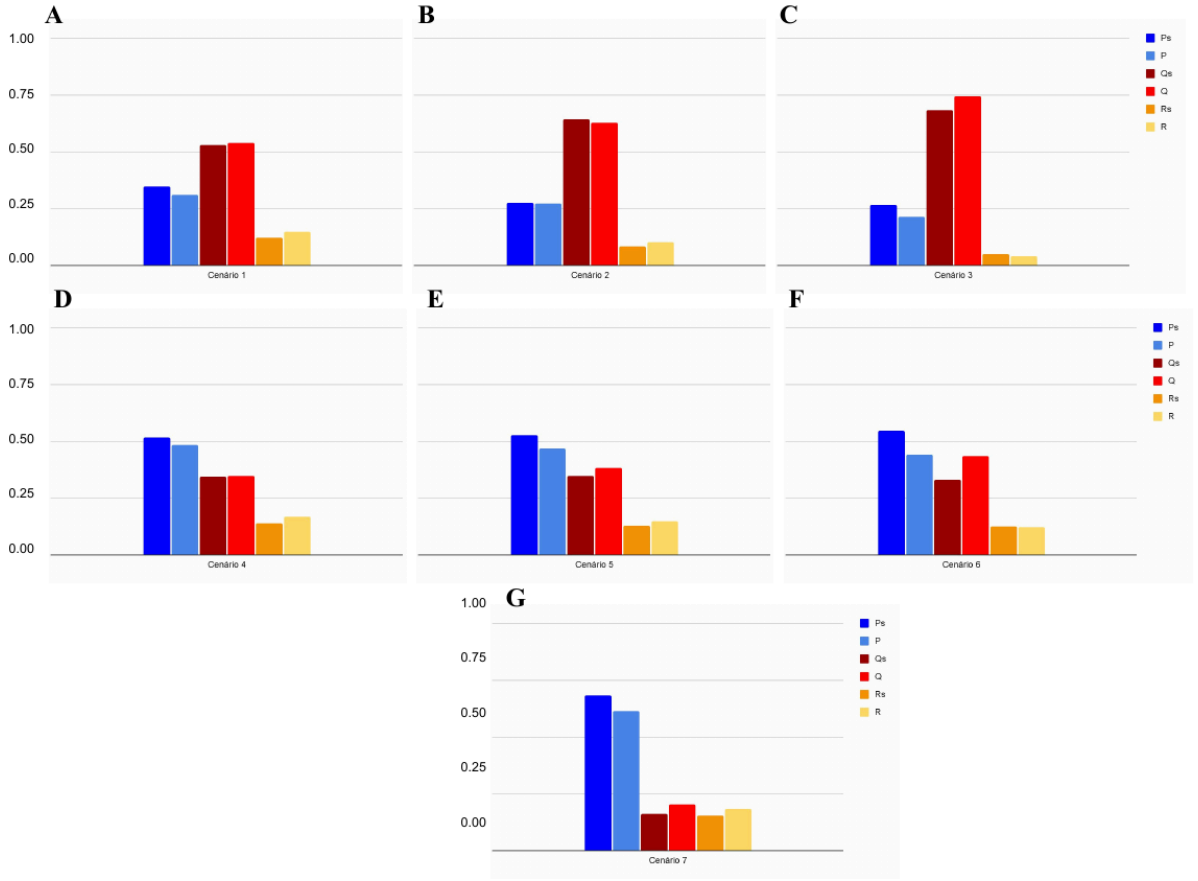

Figura 34. Comparação entre os Ms finais via Fórmula e Liang-Nielsen. Para os Ms finais de cada cenário obtido via simulador (Ps, Qs e Rs) e via Fórmula (P, Q e R) realizamos a comparação entre os cenários iguais. **A** Cenário 1 via Simulador contra Cenário 1 via Fórmula. **B** Cenário 2 via Simulador contra Cenário 2 via Fórmula. **C** Cenário 3 via Simulador contra Cenário 3 via Fórmula. **D** Cenário 4 via Simulador contra Cenário 4 via Fórmula. **E** Cenário 5 via Simulador contra Cenário 5 via Fórmula. **F** Cenário 6 via Simulador contra Cenário 6 via Fórmula. **G** Cenário 7 via Simulador contra Cenário 7 via Fórmula.

Ao observarmos a distribuição das médias de ancestralidade (M) da população Qs (ancestralidade Q via simulador) nos dados simulados para os cenários não homogêneos (isto é, excluindo o Cenário 7), os Cenários 1, 2 e 4 possuem a média via fórmula pertencente ao intervalo dos percentis de 5% e 95%. Os demais, o valor de M através do modelo teórico proposto está fora do intervalo. Isso ocorre pelo fato de que o simulador leva em conta o processo de recombinação e que o modelo de árvore binária utilizado em Liang-Nielsen tende a gerar fragmentos menores que o esperado, como alertado pelos autores, tornando a média dos Qs finais menores que o da fórmula.

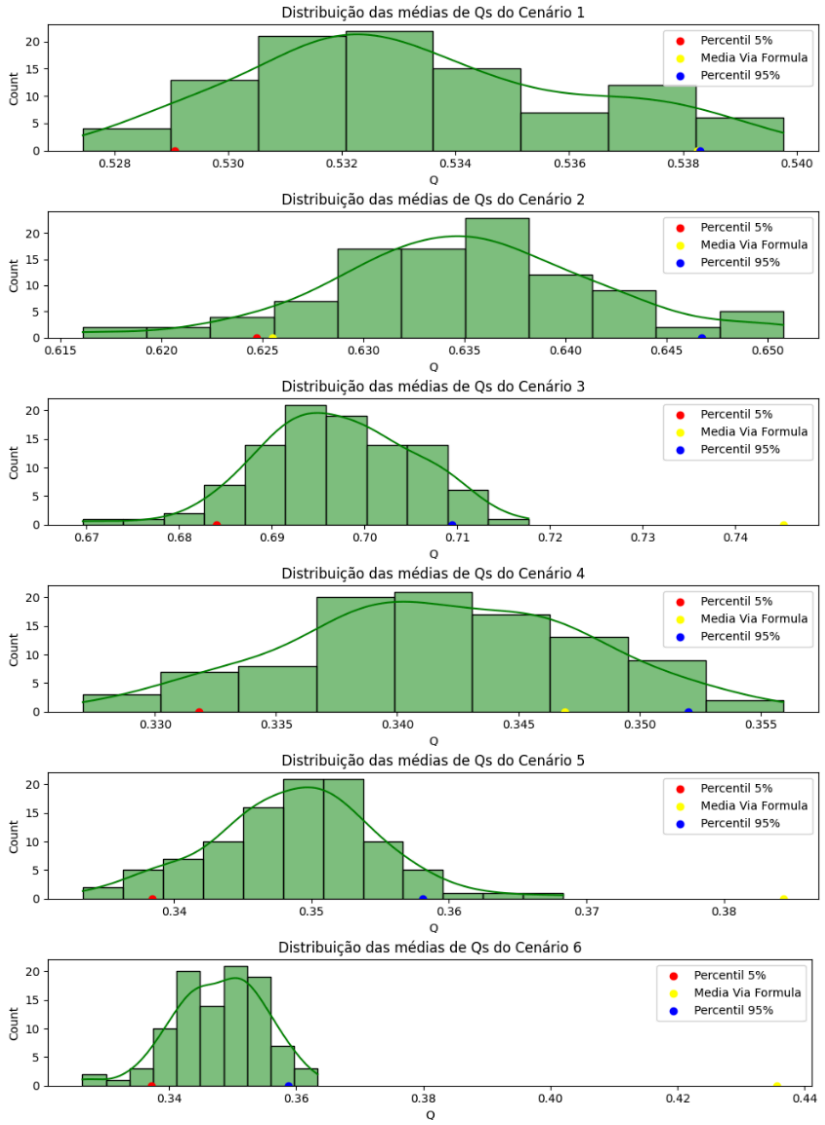

Figura 35. Intervalo de percentile 5% 95% das médias Qs via simulador vs Q via Fórmula. Tomamos 100 simulações de cada cenário para observar a dispersão das médias observadas (barras verticais em verde) e seus percentis 5% (Vermelho) e percentil 95% (Azul).

2.4.3 Discussão

Os resultados obtidos via simulador nos mostram que para uma observação da real diferença entre os cenários é preciso a simulação de grandes quantidades de dados pseudo-observados (PODs) e que se torna mais evidente quando observamos os maiores *tracts* de ancestralidade. Além disso, Liang-Nielsen trabalha com modelos estocásticos e aleatórios, tornando mais difícil a previsão e controle dos resultados. Por outro lado, a fórmula teórica descreve um modelo determinista no qual ao definirmos um parâmetro de entrada, teremos o mesmo resultado para os M's. Isso nos possibilita um controle maior na previsão dos resultados e a manipulação dos valores ao longo das gerações. Por se tratar de um modelo descritivo de

fluxo migracional, não consideramos alguns eventos como *crossing-over*. Isso pode levar a diferença nos resultados de M's quando comparado ao Liang-Nielsen, pois a fórmula descreve um modelo de proporções enquanto o simulador leva em consideração outros efeitos. Portanto, uma vez que estamos apenas interessados em avaliar as proporções de ancestralidade final populacional a partir de eventos migratórios e desenhar um cenário que se aproxima da verdadeira dinâmica de miscigenação, podemos considerar que a fórmula é mais prática.

2.4.4 Conclusão

A capacidade do simulador de Liang-Nielsen gerar resultados diferentes para cenários de fluxos migracionais diferentes nos indica um caminho a ser seguido para a descrição dos processos de dinâmica de miscigenação de uma população ao aplicarmos a metodologia proposta por Kehdy *et al.* 2015. Enquanto isso, a fórmula nos permite acompanhar como um cenário ideal de fluxo migracional define as proporções de ancestralidade de uma população caso nenhum evento além da migração ocorra.

REFERÊNCIAS

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655-1664. doi:10.1101/gr.094052.109

Borda V, Alvim I, Mendes M, Silva-Carvalho C, Soares-Souza GB, Leal TP, Furlan V, Scliar MO, Zamudio R, Zolini C, Araújo GS, Luizon MR, Padilla C, Cáceres O, Levano K, Sánchez C, Trujillo O, Flores-Villanueva PO, Dean M, Fuselli S, Machado M, Romero PE, Tassi F, Yeager M, O'Connor TD, Gilman RH, Tarazona-Santos E, Guio H. The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture. Proc Natl Acad Sci U S A. 2020 Dec 22;117(51):32557-32565. doi: 10.1073/pnas.2013773117

Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 2013 Jun;194(2):459-71. doi: 10.1534/genetics.113.150029. Epub 2013 Mar 27. PMID: 23535385; PMCID: PMC3664855.

Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free Estimation of Recent Genetic Relatedness. Am J Hum Genet. 2016 Jan 7;98(1):127-48. doi: 10.1016/j.ajhg.2015.11.022. PMID: 26748516; PMCID: PMC4716688.

Curtis D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. Psychiatr Genet. 2018;28(5):85-89. doi:10.1097/YPG.0000000000000206

Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011 Dec 4;9(2):179-81. doi: 10.1038/nmeth.1785

Fuzo CA, da Veiga Ued F, Moco S, et al. Contribution of genetic ancestry and polygenic risk score in meeting vitamin B12 needs in healthy Brazilian children and adolescents. Sci Rep. 2021;11(1):11992. Published 2021 Jun 7. doi:10.1038/s41598-021-91530-7

Gouveia MH, Borda V, Leal TP, Moreira RG, Bergen AW, Kehdy FSG, Alvim I, Aquino MM, Araujo GS, Araujo NM, Furlan V, Liboredo R, Machado M, Magalhaes WCS, Michelin LA, Rodrigues MR, Rodrigues-Soares F, Sant Anna HP, Santolalla ML, Scliar MO, Soares-Souza G, Zamudio R, Zolini C, Bortolini MC, Dean M, Gilman RH, Guio H, Rocha J, Pereira AC, Barreto ML, Horta BL, Lima-Costa MF, Mbulaiteye SM, Chanock SJ, Tishkoff SA, Yeager M, Tarazona-Santos E. Origins, Admixture Dynamics, and Homogenization of the African Gene Pool in the Americas. Mol Biol Evol. 2020 Jun 1;37(6):1647-1656. doi: 10.1093/molbev/msaa033. PMID: 32128591; PMCID: PMC7253211.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. A genetic atlas of human admixture history. Science. 2014 Feb 14;343(6172):747-751. doi: 10.1126/science.1243518. PMID: 24531965; PMCID: PMC4209567.

Hellenthal G. Methodological challenges and opportunities for inferring human demography. J Anthropol Sci. 2021 Oct 2;99:175-177. doi: 10.4436/JASS.99010. Epub ahead of print. PMID: 34601461.

Kehdy FS, Gouveia MH, Machado M, et al. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proc Natl Acad Sci U S A. 2015;112(28):8696-8701. doi:10.1073/pnas.1504447112

Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012 Jan;8(1):e1002453. doi: 10.1371/journal.pgen.1002453. Epub 2012 Jan 26. PMID: 22291602; PMCID: PMC3266881.

Leal TP, Furlan VC, Gouveia MH, et al. NAToRA, a relatedness-pruning method to minimize the loss of dataset size in genetic and omics analyses. Comput Struct Biotechnol J. 2022;20:1821-1828. Published 2022 Apr 9. doi:10.1016/j.csbj.2022.04.009

Liang M, Nielsen R. The lengths of admixture tracts. Genetics. 2014 Jul;197(3):953-67. doi: 10.1534/genetics.114.162362. Epub 2014 Apr 26. PMID: 24770332; PMCID: PMC4096373.

Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013;93(2):278-288. doi:10.1016/j.ajhg.2013.06.020

Moorjani P, Hellenthal G. Methods for Assessing Population Relationships and History Using Genomic Data. Annu Rev Genomics Hum Genet. 2023 Aug 25;24:305-332. doi: 10.1146/annurev-genom-111422-025117. Epub 2023 May 23. PMID: 37220313.

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190. doi:10.1371/journal.pgen.0020190

Popejoy, A., Fullerton, S. Genomics is failing on diversity. Nature 538, 161–164 (2016). https://doi.org/10.1038/538161a

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug;38(8):904-9. doi: 10.1038/ng1847. Epub 2006 Jul 23. PMID: 16862161.

Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019 Mar 21;177(1):26-31. doi: 10.1016/j.cell.2019.02.048. Erratum in: Cell. 2019 May 2;177(4):1080. PMID: 30901543; PMCID: PMC7380073.

Soares-Souza G., Borda, V., Kehdy, F. et al. Admixture, Genetics and Complex Diseases in Latin Americans and US Hispanics. Curr Genet Med Rep. 2018; 498(6):208–223. https://doi.org/10.1007/s40142-018-0151-z

Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. Am J Hum Genet. 2012;91(1):122-138. doi:10.1016/j.ajhg.2012.05.024

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007). PLINK: a toolset for wholegenome association and population-based linkage analysis. Am. J. Hum. Genet. 81, 559-575.

Rosenberg NA. Admixture Models and the Breeding Systems of H. S. Jennings: A GENETICS Connection. Genetics. 2016 Jan;202(1):9-13. doi: 10.1534/genetics.115.181057. PMID: 26733663; PMCID: PMC4701105.

Wangkumhang P, Hellenthal G. Statistical methods for detecting admixture. Curr Opin Genet Dev. 2018 Dec;53:121-127. doi: 10.1016/j.gde.2018.08.002. Epub 2018 Sep 20. PMID: 30245220.