UNIVERSIDADE FEDERAL DE MINAS GERAIS

Instituto De Ciências Biológicas

Programa Interunidades de Pós-graduação em Bioinformática

Carolina Silva de Carvalho

ESTRUTURA POPULACIONAL E DINÂMICA DA MISCIGENAÇÃO NA AMÉRICA LATINA: Pipelines Genéticos, Seleção Natural e Métodos de Inferência

Belo Horizonte

2024

Carolina Silva de Carvalho

ESTRUTURA POPULACIONAL E DINÂMICA DA MISCIGENAÇÃO NA AMÉRICA LATINA: Pipelines Genéticos, Seleção Natural e Métodos de Inferência

Tese apresentada ao Programa Interunidades de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito para obtenção do título de Doutora em Bioinformática.

Orientador: Prof. Dr. Eduardo Tarazona Santos

Belo Horizonte
2024

043 Carvalho, Carolina Silva de.
Estrutura populacional e dinâmica da miscigenação na América Latina: pipelines genéticos, seleção natural e métodos de inferência [manuscrito] / Carolina Silva de Carvalho. – 2024.
164 f. : il. ; 29,5 cm.

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

1. Bioinformática. 2. Genética Populacional. 3. Ancestralidade comum. 4. Miscigenação. I. Santos, Eduardo Martin Tarazona. 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

ATA

INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

DEFESA DE TESE

CAROLINA SILVA DE CARVALHO

Às oito horas do dia **28 de novembro de 2024**, reuniu-se, em formato híbrido, no Instituto de Ciências Biológicas da UFMG e através do aplicativo Zoom, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"ESTRUTURA POPULACIONAL E DINÂMICA DA MISCIGENAÇÃO NA AMÉRICA LATINA: Pipelines Genéticos, Seleção Natural e Métodos de Inferência"**, requisito para obtenção do grau de Doutora 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 à candidata, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa da candidata. Logo após, a Comissão se reuniu, sem a presença da candidata 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	Universidade Federal de Minas Gerais	**Aprovada**
Dr. Aristóteles Góes Neto	Universidade Federal de Minas Gerais	**Aprovada**
Dra. Patricia Alexandra Silva Santos	Universidade de Ferrara, Itália	**Aprovada**
Dr. Ricardo Ventura Santos	Museu Nacional/UFRJ	**Aprovada**
Dra. Sibelle Torres Vilaça	Instituto Tecnológico Vale	**Aprovada**

Pelas indicações, a candidata foi considerada: **Aprovada**

O resultado final foi comunicado publicamente à candidata 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, 28 de novembro de 2024.

Documento assinado eletronicamente por **Aristoteles Goes Neto**, **Professor do Magistério Superior**, em 28/11/2024, às 13:32, 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 **Ricardo Ventura Santos**, **Usuário Externo**, em 28/11/2024, às 13:44, 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 **Sibelle Torres Vilaca**, **Usuário Externo**, em 28/11/2024, às 19:50, 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 **Patrícia Alexandra Silva Santos**, **Usuária Externa**, em 29/11/2024, às 05:51, 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 **Eduardo Martin Tarazona Santos**, **Professor do Magistério Superior**, em 29/11/2024, às 09:44, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3758584** e o código CRC **F270D3DD**.

Referência: Processo nº 23072.268988/2024-40 SEI nº 3758584

Agradecimentos

Agradeço primeiramente a minha base e porto seguro. Minha família, que sempre me apoiou e me incentivou a seguir a diante. Meu pai, Robson Marques, pelo carinho e suporte, por me mostrar que o mundo é amplo e incentivar que eu siga em várias direções. A minha mãe, Orsilane Valadares, também pelo carinho e suporte, pelos abraços capazes de curar o mundo e por toda compreensão ao longo da vida. A minha irmã, Orsilane Carvalho, por ser minha melhor amiga, me apoiar, me distrair e me manter no foco, em equilíbrio conforme necessário. A meu marido, Guilherme Marcelino, pelos anos de parceria, por me acompanhar em tudo e por aceitar sem questionamentos meus momentos de loucura.

A toda equipe do LDGH, tanto ex-membros como os membros atuais, agradeço a receptividade e ambiente leve e colaborativo em meus 7 anos de casa. Especialmente agradeço ao Professor Eduardo por me receber em meus últimos semestres da graduação e abrir diversas portas em meu caminho até o fim de meu doutorado. À Camila agradeço imensamente por todas as oportunidades, os puxões de orelha, conselhos e (acredito que possa dizer) amizade ao longo destes anos. Aos ex-membros Julia, Ferdi, Victor, Thiago, Marla, Mateus e Isabela agradeço por todas as discussões científicas e elucidações desde o momento que entrei na equipe. A Rafael por ter sobrevivido a minha primeira experiência como orientadora, Lucas por todas as discussões científicas e cafés, a Pedro e Luca por terem contribuído tanto como meus orientandos mais recentes. Agradeço a todos com quem tive contato nessa equipe maravilhosa por todos os momentos juntos em que me senti parte do time.

Resumo

A estrutura genética latino-americana é marcada por diversos eventos históricos, incluindo as expansões populacionais pré-colombianas e os eventos de miscigenação pós-colombianos. O entendimento da estrutura populacional pode ajudar a esclarecer estes eventos, uma vez que ela é resultado da interação das populações com fatores evolutivos. O primeiro capítulo deste trabalho aborda os pipelines e metodologias de inferência da estrutura genética em dois níveis: a estrutura a nível familiar e a estrutura a nível populacional, considerando métodos baseados em genótipos e haplótipos. Além disso, apresento nossos fluxogramas do Laboratório de Diversidade Genética Humana (LDGH) usados em projetos de colaboração e próprios. Nossos pipelines visam incluir populações sub-representadas no contexto genômico mundial, como populações miscigenadas e populações Nativo-americanas, uma vez que a lacuna da diversificação afeta a aplicabilidade da genética médica globalmente. Dessa forma, iniciativas como o Projeto Genomas Peruanos se tornam cada vez mais relevantes para atender às necessidades de diversificação populacional nos estudos genéticos humanos. O segundo capítulo aborda a estruturação populacional, seleção natural e diferenciação de SNVs em nativos americanos peruanos, trazendo inferências sobre a história pré-colombiana e sua influência na distribuição de duas variantes de importância farmacogenética. No contexto pós-colombiano, ainda existem lacunas históricas sobre a dinâmica do processo de colonização ocorrido nos últimos 500 anos nas américas. O terceiro capítulo do presente trabalho apresenta a implementação de um modelo demográfico e um pipeline baseado em computação bayesiana aproximada para inferir a dinâmica da miscigenação a partir de dados de ancestralidade global e local, recuperado de um trabalho anterior do grupo e implementado com ferramentas mais atuais de inferências genético-populacionais.

Palavras-chave: Inferências Genético-populacionais, Ancestralidade, Populações Sub-representadas.

Abstract

The Latin-America genetic structure shows signs from diverse historical events, including pre-Colombian population expansions and the admixture events post-Colombian. The acknowledgment of the populational structure can elucidate these events, once it results from the interaction of the population with evolutive agents. The first chapter of this dissertation presents pipelines and methods used for genetic structure in two levels: family structure and populational structure, considering haplotype- and genotype-based methods. Furthermore, I present our flowcharts from the Genetic Human Diversity Laboratory (LDGH), used in our projects and collaborations. Our pipelines aim to include populations underrepresented in the world genomic scenario, such as admixed and Native American populations, once this gap affects the applicability of precision medicine globally. Therefore, initiatives such as the Peruvian Genomes Project became increasingly relevant to answer the calls for more population diversification in human genetics studies. The second chapter of this dissertation addresses the population structure, natural selection and differentiation of SNVs in native Americans from Peru, presenting inferences about pre-Colombian history and its influences on the distribution of two genetic variants of pharmacogenetics importance. In the post-Colombian context, there are still historical gaps in the colonization process that occurred 500 years by past in the Americas. The third chapter of this dissertation brings the implementation of a demographic model, and a pipeline based on approximated Bayesian computation to infer the admixture dynamics from global and local chromosomic ancestries, recovered from previous work from our group and implemented with more accurate tools for population genetics inferences.

Key-words: Population Genetics Inferences, Ancestry, Underrepresented Populations

LISTA DE FIGURAS

Figura 1. Estrutura familiar identificada na coorte de Bambuí utilizando ADMIXTURE, REAP e PCA. Ao utilizar indivíduos de outras coortes, ADMIXTURE (K = 7) identifica clusters de ancestralidade que correspondem a um grupo de indivíduos aparentados identificados por análise de kinship do software REAP e por uma abordagem de redes desenvolvida pelo grupo. Indivíduos do cluster preto também foram identificados pelo segundo componente do PCA (pontos vermelhos), quando aplicado somente para a coorte de Bambuí. Fonte: Kehdy et al. 2015. 18

Figura 2 Esquema dos níveis de ancestralidade. A ancestralidade populacional se refere à porcentagem que cada população parental contribuiu em uma população miscigenada. A ancestralidade individual estima a porcentagem de cada população parental que o indivíduo apresenta. E a ancestralidade cromossômica estima a população parental de origem em cada segmento cromossômico. As cores representam as populações parentais. Azul - Africana; Vermelho - Europeia; Verde - Nativo Americana. Fonte: Soares-Souza et al., 2018. 20

Figura 3 Esquema com os achados de ancestralidade local para Scliar et al. 2021. Esquema de cromossomos demonstrando Admixture-Mapping para variante associada ao aumento do IMC presente em porções genômicas de ancestralidade africana em mulheres com obesidade. Fonte: Scliar et al., 2021. 22

Figura 4 Mapa de organização do banco de dados para inferência de ancestralidade a nível individual e local cromossômica. 23

Figura 5 . Pipeline de ancestralidade e estrutura populacional aplicado no LDGH. a) Pipeline de ancestralidade individual e parentesco. b) Pipeline dos métodos baseados em haplótipos CHROMOPAINTER e FineStructure. c) Pipeline Ancestralidade Local, baseada em haplótipos. d) Pipeline Identidade por Descendência, baseada em haplótipos. 26

Figura 6 Admixture continental e análise de kinship para populações brasileiras do EPIGEN. A) Coeficientes de kinship para cada par de indivíduos e a probabilidade que eles compartilhem zero alelos idênticos por descendência (IBD) (IBD = 0). As linhas horizontais representam o coeficiente de kinship necessário que deve ser usado para serem considerados parentes. B) Regiões brasileiras, as populações estudadas e o barplot correspondente a sua ancestralidade continental individual. N representa o número de indivíduos do EPIGEN no banco de dados original (incluindo aparentados). C) Representação em PCA, incluindo populações do EPIGEN e globais, usando apenas indivíduos não aparentados. Os três gráficos derivam da mesma análise e são diferentes apenas pelo plot dos indivíduos do EPIGEN. AP, admixed population; ASW, Americans of African ancestry in USA; CEU, Utah residents with Northern and Western European ancestry; CLM, Colombians from Medellin, Colombia; EAFR, east Africa; FIN, Finnish in Finland; French B, Basque; GBR, British in England and Scotland; IBS, Iberian population in Spain; LWK, Luhya in Webuye, Kenya; ME, Middle East; MXL/MEX, Mexican ancestry from Los Angeles; N., (North) Italian; NAT, Native American; NE, northeast; NEUR, north Europe; PC, principal component; PUR, Puerto Ricans from

Puerto Rico; S, south; SE, southeast; SEUR, south Europe; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeira; WAFR, west Africa. Fonte: Kehdy et al. 2015.....28

Figura 7 Base do Shapeit5 ilustrando inferência de haplótipo. a) Todas as amostras são faseadas para variantes comuns (MAF ≥ 0.1%). b) Faseamento de uma determinada variante rara nos haplótipos para variantes comuns. Os haplótipos utilizados no condicionamento da estimativa compartilham longas porções com os targets (verde e azul) e não são monomórficos para variantes raras. Como os genótipos heterozigotos para as variantes raras não estão faseados, assume-se que os alelos menores nessas regiões estão nos dois haplótipos (homozigose forçada). c) Faseamento de singleton assumindo o novo alelo no haplótipo target com a correspondência mais curta. d) Mapeamento de eventos de heterozigotos compostos baseados no faseamento de variantes raras (a-c). Fonte: Hofmeister et al., 2023.....30

Figura 8 Barplot representando as proporções de ancestralidade individual para os indivíduos do projeto EPIGEN. As proporções de valores de ancestralidade individual foram calculadas usando de K=3 até K=10 para o banco de dados original. As populações de cada grupo étnico/geográfico são descritas abaixo da figura na mesma ordem que aparecem na imagem. Cada barra representa um indivíduo e cada cor um cluster específico de ancestralidade. *Moabitas são uma população do noroeste africano. Fonte: Kehdy et al., 2015.....32

Figura 9 Clustering baseado em haplótipo de populações parentais e inferência de admixture para populações americanas continentais miscigenadas. A) Árvore gerada com fineSTRUCTURE de indivíduos parentais. B) Contribuições de populações subcontinentais relativas à ancestralidade africana total em populações miscigenadas inferidas pela regressão MIXTURE MODEL implementada no GLOBETROTTER. C) Inferências de GLOBETROTTER de eventos de miscigenação para cada população miscigenada. A data inferida e os intervalos de 95% de confiança são representados por pontos e linhas horizontais no gráfico. O retângulo tracejado ilustra o período mais dinâmico de miscigenação. Além do gráfico da datação, estão representadas as fontes inferidas de miscigenação para eventos recentes e anteriores. O tamanho das barras representa a contribuição genética de cada fonte. Cada cor corresponde à proporção da contribuição de cada população parental. CEU, Utah Residents (CEPH) with Northern and Western Ancestry-United States, IBS, Iberian population in Spain; CLM, Colombians from Medellin; PUR, Puerto Ricans from Puerto Rico; ACB, African Caribbeans in Barbados; ASW, African Americans in Southwest United States; PLCO, African Americans from East United States. Fonte: Gouveia et al., 2020. 34

Figura 10 Plot de cariograma esquemático representando a ancestralidade local de um indivíduo brasileiro calculada utilizando RFMIX. As cores representam as ancestralidades parentais em cada região genômica para cada um dos 22 cromossomos autossômicos em um indivíduo brasileiro miscigenado. A ancestralidade local foi calculada utilizando RFMIX e o plot gerado utilizando o programa 3A do grupo. Fonte: Autoria do grupo. 35

Figura 11 Proporção de ancestralidade global e distribuição geográfica das populações Peruanas selecionadas para o Projeto Genoma Peruano. Os gráficos de rosca

mostram a proporção de miscigenação de cada população, inferida pela média do resultados de K= 3 do software ADMIXTURE. Os componentes de ancestralidade de origem Africana, Europeia e Nativo-Americana estão representados nos gráficos em azul, vermelho e verde, respectivamente. O mapa adjacente mostra a localização das diferentes populações em círculos numerados coloridos com diferentes intensidades de cinza para indicar a média de altitude da posição de cada população. Gerado para: Guio et. al. 2024, Pré-print.....37

Figura 12 Repercussão midiática do comentário publicado na revista Cell. A) Capa da reportagem publicada na DW alemã. B) Capa da reportagem publicada pela revista Science. 55

Figura 13 Resumo da variante rs874116 como potencial sinal de seleção natural. A) Plot de violino representando a distribuição dos valores de F_{ST} na comparação Amazônia versus CDX (em verde) e Andes versus CDX (em marrom). As linhas pontilhadas representam a posição do rs874116 em cada distribuição. B) Frequências alélicas da variante rs874116 em populações diversas. C) Gráfico de bifurcação do haplótipo ao redor da variante rs874116 nas populações nativas peruanas, considerando Andes, Amazônia e Costa. 58

Figura 14 Distribuição dos valores de PBSn na anotação "Processo metabólico de xenobióticos". Nessa via se encontram genes como SULT1A1, CYP26B1, CYP2C18, CYP2C19, CYP2C9, CYP2E1, TBXAS, FMO3, AIP, ACSM2B e ACSM1 com SNPs com altos valores de PBSn para a população andina (PBSn > 0.1, valor de $p < 0.05$). a) Distribuição dos valores de PBSn calculados usando as populações amazônicas como alvo. Valor de p para KS $= 2.68x\ 10^{-11}$. b) Distribuição dos valores de PBSn calculados usando as populações andinas como alvo. Valor de p para o KS $= 2.77x\ 10^{-8}$. 59

Figura 15 Padrão esquemático dos fragmentos cromossômicos de ancestralidade ao longo das gerações. 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. 62

Figura 16 Modelo demográfico de miscigenação pós-colombiana e identificação gráfica dos parâmetros m de migração. A) Características dos parâmetros $m_{i,g}$ no método aplicado nos cromossomos autossômicos. B) Definição dos pulsos de migração e ao longo do tempo e dos parâmetros $m_{i,g}$ associados a cada pulso na aplicação do método nos cromossomos autossômicos. 65

Figura 17 Ilustração da influência dos valores de mi, g nos valores de Ma, g. Na geração g = 0 a população é formada somente pela população P, tendo valores de $mP, 0$ e $MP, 0 = 1$ e os valores de $MQ, 0$ e $MR, 0$ são iguais a zero. Na geração g =1 há a chegada de um fluxo $mQ, 1 = 0{,}3$, o que faz com os valores de $MP, 1 = 0{,}7$, $MQ, 1 = 0{,}3$ e $MR, 1 = 0$. Na geração g =2 há a chegada de um fluxo $mR, 2 = 0{,}2$, o que faz com os valores de $MP, 2 = 0{,}56$, $MQ, 2 = 0{,}24$ e $MR, 2 = 0{,}20$. Na geração g =3 há a chegada de um fluxo $mQ, 3 = 0{,}5$, o que faz com os valores de $MP, 3 = 0{,}28$, $MQ, 3 = 0{,}62$ e $MR, 3 = 0{,}10$. 66

Figura 18 Análise de componentes principais obtida a partir do banco descrito na Tabela 2. PCA utilizado para seleção das amostras referência na inferência de ancestralidade local. No eixo X se encontra o PC1 e no eixo Y o PC2. 73

Figura 19 Fluxograma da metodologia ABC. 75

Figura 20 Ancestralidade populacional e individual nas populações miscigenadas Latino-Americanas. Em azul temos o componente de miscigenação Africano; em verde o componente de miscigenação Nativo-americano e em vermelho o componente de miscigenação Europeu. A) Em gráficos de pizza separados vemos a proporção do componente de cada parental para formação das populações miscigenadas nos cromossomos autossômico. B) A distribuição das ancestralidades individuais calculadas com base nos cromossomos autossômicos. 76

Figura 21 Distribuições de tracts de ancestralidade para as coortes de Pelotas, Salvador e Bambuí. A) Distribuições dos tracts de ancestralidade obtidos de todos os cromossomos autossômicos. B) Distribuições de tracts obtidas a partir dos cromossomos 1, 2, 15 e 21. 78

Figura 22 Distribuições dos parâmetros $m_{i,g}$ na população de Pelotas nos três pulsos de miscigenação. A) As linhas pontilhadas representam as distribuições a priori e as linhas contínuas representam as distribuições a posteriori. As cores azul, vermelho e verde correspondem às ancestralidades Africana, Europeia e Nativo-Americanos, respectivamente. B) Intervalos de credibilidade 90% HPD das distribuições a posteriori. 80

Figura 23 Distribuições dos parâmetros $m_{i,g}$ na população de Bambui nos três pulsos de miscigenação. A) As linhas pontilhadas representam as distribuições a priori e as linhas contínuas representam as distribuições a posteriori. As cores azul, vermelho e verde correspondem às ancestralidades Africana, Europeia e Nativo-Americanos, respectivamente. B) Intervalos de credibilidade 90% HPD das distribuições a posteriori. 81

Figura 24 Distribuições dos parâmetros $m_{i,g}$ na população de Salvador nos três pulsos de miscigenação. A) As linhas pontilhadas representam as distribuições a priori e as linhas contínuas representam as distribuições a posteriori. As cores azul, vermelho e verde correspondem às ancestralidades Africana, Europeia e Nativo-Americanos, respectivamente. B) Intervalos de credibilidade 90% HPD das distribuições a posteriori. 82

Figura 25 Ancestralidade populacional e individual nas populações miscigenadas Brasileiras inferidas com base no cromossomo X. Em azul temos o componente de miscigenação Africano; em verde o componente de miscigenação Nativo-americano e em vermelho o componente de miscigenação Europeu. A) Em gráficos de pizza separados vemos a proporção do componente de cada parental para formação das populações miscigenadas no cromossomo X. B) A distribuição das ancestralidades individuais calculadas com base no cromossomo X. 85

Figura 26 Modelo demográfico de miscigenação pós-colombiana e identificação gráfica dos parâmetros m de migração pelo cromossomo X. A) Características dos parâmetros $m_{i,g}$ no método aplicado no cromossomo X. B) Definição dos pulsos de migração e ao longo do tempo e dos parâmetros $m^{s}_{i,g}$ associados a cada pulso na aplicação do método no cromossomo X.....86

LISTA DE TABELAS

Tabela 1 SNVs presentes na região genica do gene PPP3CA contidos no array Omni Array 2.5 M. 57

Tabela 2 Descrição das populações do banco de dados de 1.5M 70

Tabela 3 Intervalos de Credibilidade 90% HPD das distribuições a posteriori e suas amplitudes para as coorte de Bambuí, Salvador e Pelotas. 83

LISTA DE FÓRMULAS

Equação 1 Índice de ficação 27

Equação 2 Coeficiente de inbreeding F para o indivíduo 27

Equação 3 Número esperado de loci em homozigose 27

Equação 4 Proporção de ancestralidade da população a na geração g $M_{a,g}$ 66

LISTA DE ABREVIAÇÕES

1KGP	*1000 Genomes Project*
ABC	*Approximate Bayesian Computation* (Computação Bayesiana Aproximada)
ABCG2	*ATP Binding Cassette eSubfamily G Member 2*
ACB	*African Caribbean in Barbados*
aDNA	DNA antigo
AFR	Africanos
ASW	*African Ancestry in Southwest US*
AWS	*Amazon Web Services*
CDX	*Chinese Dai in Xishuangbanna*
CEU	*North Europeans from Utah*
CRISPR/cas9	*Clustered Regulatory Interspaced Short Palindromic Repeats/CRISPR Associated Protein 9*
DBSNP	*Single Nucleotide Polymorphism Database*
DNA	*Deoxyribonucleic Acid*
DUOX2	*Dual Oxidase 2*
ESN	*Esan in Nigeria*
EUR	Europeus
GBR	*British from England and Scotland*
GWD	*Gambian in Western Division*
HAND2	*Heart and Neural Crest Derivatives Expressed 2*
HAND2-AS1	*HAND2 Antisense RNA 1*

HGDP	*Human Genome Diversity Project*
HPD	*High Posteriori Density* (Densidade a posteriori máxima)
HWE	*Hardy-Weinberg Equilibrium* (Equilíbrio de Hardy-Weinberg)
IBD	*Identity by descent (Identidade por descendência)*
IBS	*Identity by state (Identidade por estado)*
IBS	*Iberian Population in Spain*
IMC	Índice de Massa Corporal
KS	Kolmogorov-Smirnov
LD	*Linkage Disequilibrium* (Desequilíbrio de Ligação)
LDGH	Laboratório de Diversidade Genética Humana
LOCA	*Local Ancestry* (Ancestralidade Local)
LRH	*Long Range Haplotype*
LWK	*Luhya in Webuye, Kenya*
MASSA	*Multi-Agent System for SNP Annotation*
MCMC	*Monte Carlo Markov Chain* (Cadeia de Markov Monte Carlo)
MSL	*Mende in Sierra Leone*
NAT	Nativo-Americanos
NGS	*Next Generation Sequencing* (Sequenciamento de nova geração)
PBMC	*Peripheral Blood Mononuclear Cells*
PBSn	*Population Branch Statistic normalised*
PCA	*Principal Component Analysis* (Análise de Componente Principais)
PC	*Principal Component* (Componente Principal)

PGP	*Peruvian Genome Project* (Projeto Genoma Peruano)
PNAS	*Proceedings of National Academy of Sciences*
PPP3CA	*Protein Phosphatase 3 Catalytic Subunit Alpha*
PTPRC	*Protein Tyrosine Phosphatase Receptor Type C*
REAP	*Relatedness Estimates in Admixed Populations*
RNA	*Ribonucleic Acid*
RT-PCR	*Real Time Polimerase Chain Reaction*
SNV	*Single Nucleotide Variant* (Variante de Nucleotídeo Único)
SNP	*Single Nucleotide Polymorphism* (Polimorfismo de Nucleotídeo Único)
TSI	*Tuscans from Italy*
UFMG	Universidade Federal de Minas Gerais
VKORC1	*Vitamin K Epoxide Reductase Complex Subunit 1*
XP-EHH	Cross-Population Extended Homozygosity
YRI	*Yoruba in Ibadan, Nigeria*

SUMÁRIO

CAPÍTULO 1: Mini-revisão de Estrutura Populacional e ancestralidade 17

1.1. Introdução 17

1.2. Estruturação a nível familiar 27

1.2.1. Métodos baseados em genótipos 27

1.2.2. Métodos baseados em haplótipos 29

1.3. Estruturação a nível populacional 31

1.3.1. Métodos baseados em genótipo: 31

1.3.2. Métodos baseados em haplótipos: 33

CAPÍTULO 2: Genética de populações e medicina evolutiva em nativos americanos peruanos 36

2.1. Introdução 36

2.2. The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture 39

2.2.1. Contribuições 39

2.2.2. Artigo 39

2.3. The need to diversify genomic studies: Insights from Andean highlanders and Amazonians 49

2.3.1. Contribuições 49

Artigo 49

2.4. Discussão 55

2.5. Perspectivas 58

CAPÍTULO 3. Inferência da Dinâmica da Miscigenação em Latino Americanos . 62

3.1 Introdução 62

3.2 Modelo demográfico e genético populacional da miscigenação 64

3.2.1 Formalização matemática e generalização da dinâmica da miscigenação 65

3.3 Como inferir a dinâmica da miscigenação a partir de dados genéticos 67

3.3.1 ABC: como inferir a distribuição a posteriori dos parâmetros $\boldsymbol{mi, g}$ 67

3.4 Resultados 76

3.4.1 Ancestralidade Global e Distribuição dos *tracts* de ancestralidade local nas populações alvo 76

3.4.2 Parâmetros para simulação dos cenários 78

3.5 Discussão 83

3.6 Perspectivas 84
REFERÊNCIAS 87
ANEXOS 94
Anexo 1 - Artigos em colaboração 94
Anexo 1.1 - Resumo do artigo "*Rare genetic variants of NLRP12 in Admixed Latino-American Children with SARS-CoV-2-related Multisystem Inflammatory Syndrome*", publicado em colaboração com a Fiocruz PE 94
Anexo 1.2 - Resumo do artigo "*The Peruvian Genome Project: Expanding the Global Pool of Genome Diversity from South America*", em submissão em colaboração com o Projeto Genoma Peruano 95
Anexo 1.3 - Resumo do artigo "*The effect of altitude on the expression of immune-related genes in Peruvian rural indigenous*", em submissão em colaboração com o Projeto Genoma Peruano 97
Anexo 2 - Materiais suplementares de artigos em primeira autoria 99
Anexo 2.1 Material Suplementar do artigo: "The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture" 99
Anexo 2.2 Material Suplementar do artigo: "The need to diversify genomic studies: insights from Andean highlanders and Amazonians" 157

CAPÍTULO 1: Mini-revisão de Estrutura Populacional e ancestralidade

O presente capítulo, inicialmente escrito pela autora da tese, dará origem a uma mini-*review*, que será publicada em coautoria com os membros da equipe de ancestralidade do Laboratório de Diversidade Genética Humana Luca Gama e Lucas Faria. Portanto, a versão aqui apresentada já inclui contribuições de ambos os autores.

1.1. Introdução

A interação de uma população com fatores evolutivos resulta em suas características genéticas e demográficas. A informação da estrutura genética populacional pode ser extraída da genética dos indivíduos a partir de métodos baseados em genótipos de variantes independentes ou baseados em haplótipos. Os métodos baseados em haplótipo normalmente precisam de uma densidade maior de variantes e usam os padrões de desequilíbrio de ligação (LD) ao longo dos genomas dos indivíduos para inferir a fase cromossômica dos alelos e detectar a estrutura das populações (Lawson *et al.* 2012). Por usar mais informações genéticas, os métodos baseados em haplótipos podem ser mais acurados e serem capazes de detectar estrutura populacional fina, desde que o banco de dados analisado possua uma alta densidade de variantes.

Podemos considerar que a estrutura genética tem dois componentes: o nível familiar (*inbreeding*, *kinship* e parentesco), e o nível de estrutura populacional. O nível familiar é quantificado pelo cálculo dos coeficientes de *inbreeding* de um indivíduo e dos coeficientes de *kinship* e de parentesco para pares de indivíduos. Define-se *inbreeding* como o resultado do acasalamento entre indivíduos aparentados (Hedrick 2005). Ao contrário de fenômenos evolutivos como o acasalamento preferencial positivo por algum fenótipo, que afeta unicamente os *loci* associados com esse fenótipo, o *inbreeding* afeta todos os *loci* no genoma, provocando uma queda na proporção de heterozigotos de um indivíduo ou de uma população em relação ao esperado com base no equilíbrio de Hardy-Weinberg (HWE). O coeficiente de *inbreeding* mostra a probabilidade de dois alelos homólogos amostrados em um indivíduo serem idênticos por descendência (IBD, i.e. derivar de um mesmo antepassado há poucas

gerações). O coeficiente de *kinship*, de forma similar ao *inbreeding*, trata-se da probabilidade de que dois alelos homólogos coletados aleatoriamente em dois indivíduos diferentes (diploides para cromossomos autossômicos) sejam IBD.

A inferência do parentesco por ferramentas genômicas enriquece os estudos populacionais ao detectar relações entre indivíduos que não foram informadas previamente (Lee e Chen 2016). Detectar a estrutura familiar dentro de uma amostra é fundamental, uma vez que esta pode ser confundida com uma população discreta e interferir na inferência de estrutura populacional (Figura 1, extraída de Kehdy *et al.* 2015; Price *et al.* 2010).

A estrutura genética a nível populacional, assim como as análises de parentesco, é inferida com base na similaridade entre os indivíduos. Um dos componentes da estruturação populacional é a ancestralidade. Detectar geneticamente os indivíduos de uma mesma ancestralidade requer modelos que considerem isolamento, deriva genética e fluxo gênico. Para definir o conceito de ancestralidade genética, precisamos entender a diferença entre as populações miscigenadas e as populações parentais.

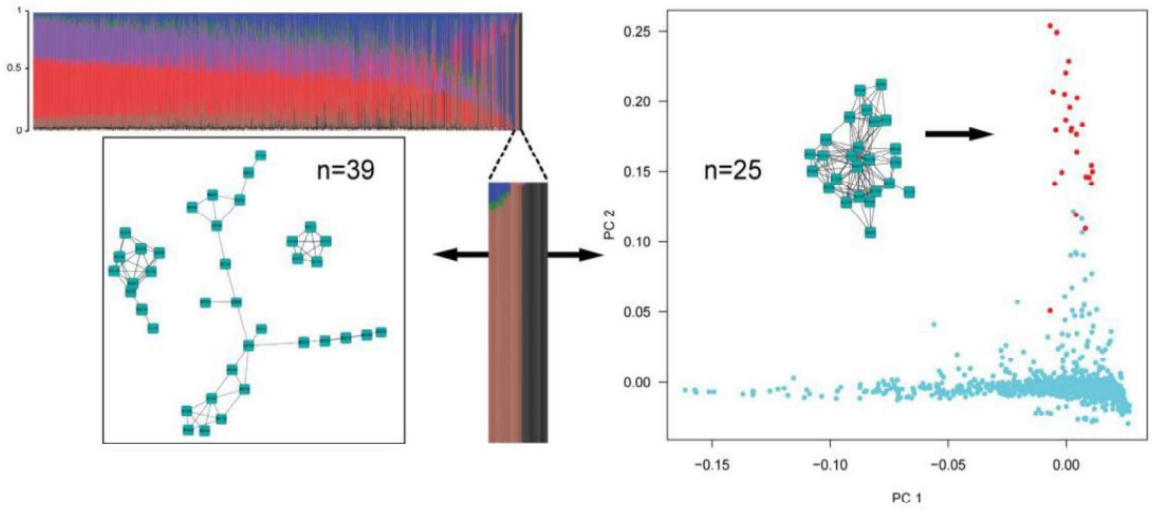


Figura 1. Estrutura familiar identificada na coorte de Bambuí utilizando ADMIXTURE, REAP e PCA. Ao utilizar indivíduos de outras coortes, ADMIXTURE (K = 7) identifica clusters de ancestralidade que correspondem a um grupo de indivíduos aparentados identificados por análise de *kinship* do *software* REAP e por uma abordagem de redes desenvolvida pelo grupo. Indivíduos do cluster preto também foram identificados pelo segundo componente do PCA (pontos vermelhos), quando aplicado somente para a coorte de Bambuí. Fonte: Kehdy et al. 2015.

Chamamos de população miscigenada aquela população formada pela mistura entre duas ou mais populações parentais, que são populações que permaneceram isoladas entre si por várias gerações. Dessa forma, os genomas dos indivíduos dessas populações são mosaicos formados pelos genomas originados das populações ancestrais, que se misturam ao

longo do tempo. A ancestralidade pode ser mensurada através da genética em três níveis: populacional, onde é calculada a proporção de contribuição de cada população parental na formação da população miscigenada (Figura 2-A), individual (Figura 2-B), em que inferimos a proporções de contribuição das populações ancestrais na formação do genoma do indivíduo, e local cromossômica (Figura 2-C), onde, utilizando dados de alta densidade de marcadores no genoma, inferimos a ancestralidade de cada fragmento de cada cromossomo de um indivíduo miscigenado.

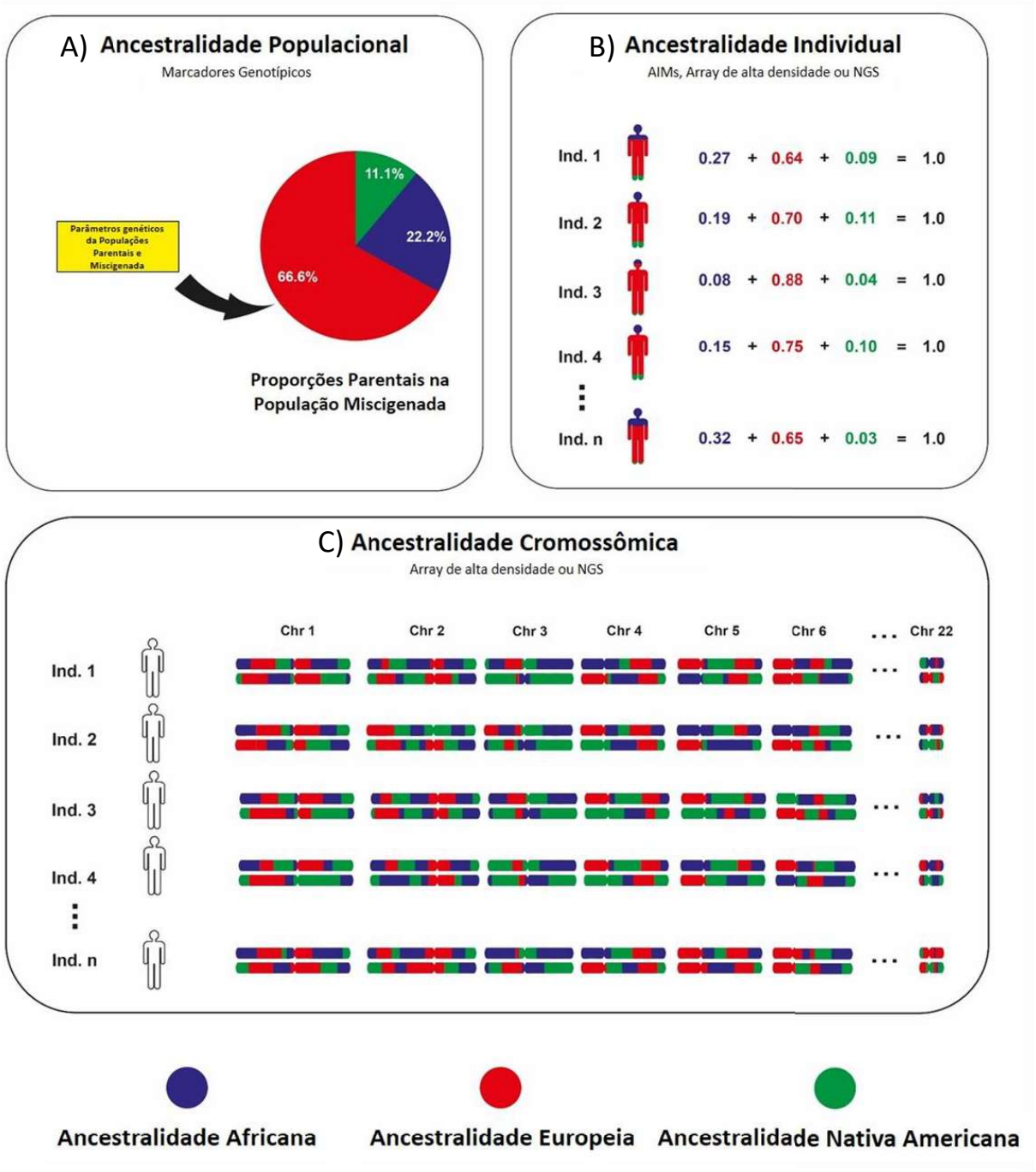


Figura 2 Esquema dos níveis de ancestralidade. A ancestralidade populacional se refere à porcentagem que cada população parental contribuiu em uma população miscigenada. A ancestralidade individual estima a porcentagem de cada população parental que o indivíduo apresenta. E a ancestralidade cromossômica estima a população parental de origem em cada segmento cromossômico. As cores representam as populações parentais. Azul - Africana; Vermelho - Europeia; Verde - Nativo Americana. Fonte: Soares-Souza et al., 2018.

As informações de ancestralidade genética são capazes de recuperar lacunas históricas em populações miscigenadas. Devido ao processo iniciado pelas grandes navegações, pouco mais de 500 anos atrás, a América Latina se tornou um dos grandes exemplos do processo de miscigenação em populações humanas, sendo sua população formada principalmente pela

miscigenação entre indivíduos de origem europeia, africana e nativo-americana. Os elementos envolvidos na história de miscigenação latino-americana são extremamente complexos e estão relacionados a uma história social violenta e estratificada que ainda se reflete nos dias atuais. Este reflexo também pode ser encontrado na genética de sua população, como demonstra Kehdy *et al.* (2015) ao evidenciar geneticamente o conhecido viés sexual, demonstrando diferenças de proporções parentais ao comparar o cromossomo X com autossômicos. Este fator pode ser relevante para estudos de doenças relacionadas ao cromossomo X, por exemplo.

A inferência da ancestralidade vai além do conhecimento da história das populações, mas também no contexto da saúde e qualidade de vida das populações. A ancestralidade pode estar associada à expressão de alguns fenótipos. Um método para detecção deste tipo de relação é o *Admixture Mapping* (Patterson *et al.* 2004), que usa as informações de ancestralidade local para encontrar associações de variantes genéticas a fenótipos complexos. Este foi o método aplicado por Scliar *et al.* 2021, onde é revelada uma variante que está associada ao aumento do IMC (índice de massa corporal) em mulheres miscigenadas que a possuem especificamente em fragmentos genômicos de origem africana (Figura 3). Outro exemplo da relação entre ancestralidade e saúde é no cálculo de *polygenic risk scores* (PRS). Alguns estudos mostram que alguns fenótipos complexos parecem estar mais associados à ancestralidade que ao próprio PRS (Curtis 2018; Fuso *et al.* 2021).

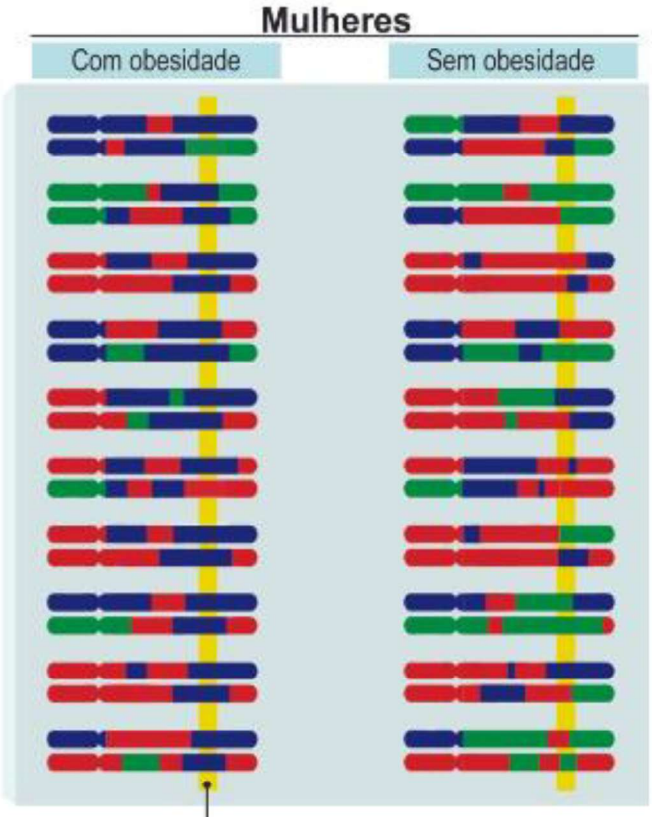


Figura 3 Esquema com os achados de ancestralidade local para Scliar et al. 2021. Esquema de cromossomos demonstrando Admixture-Mapping para variante associada ao aumento do IMC presente em porções genômicas de ancestralidade africana em mulheres com obesidade. Fonte: Scliar et al., 2021.

Para viabilizar nossas análises de ancestralidade, nosso grupo estruturou nosso banco de dados de forma que os dados dos projetos gerados e publicados pelo laboratório e bancos de dados públicos foram organizados por região geográfica (Figura 4). O objetivo é facilitar o uso de bancos de dados de referência para análises específicas de ancestralidade.

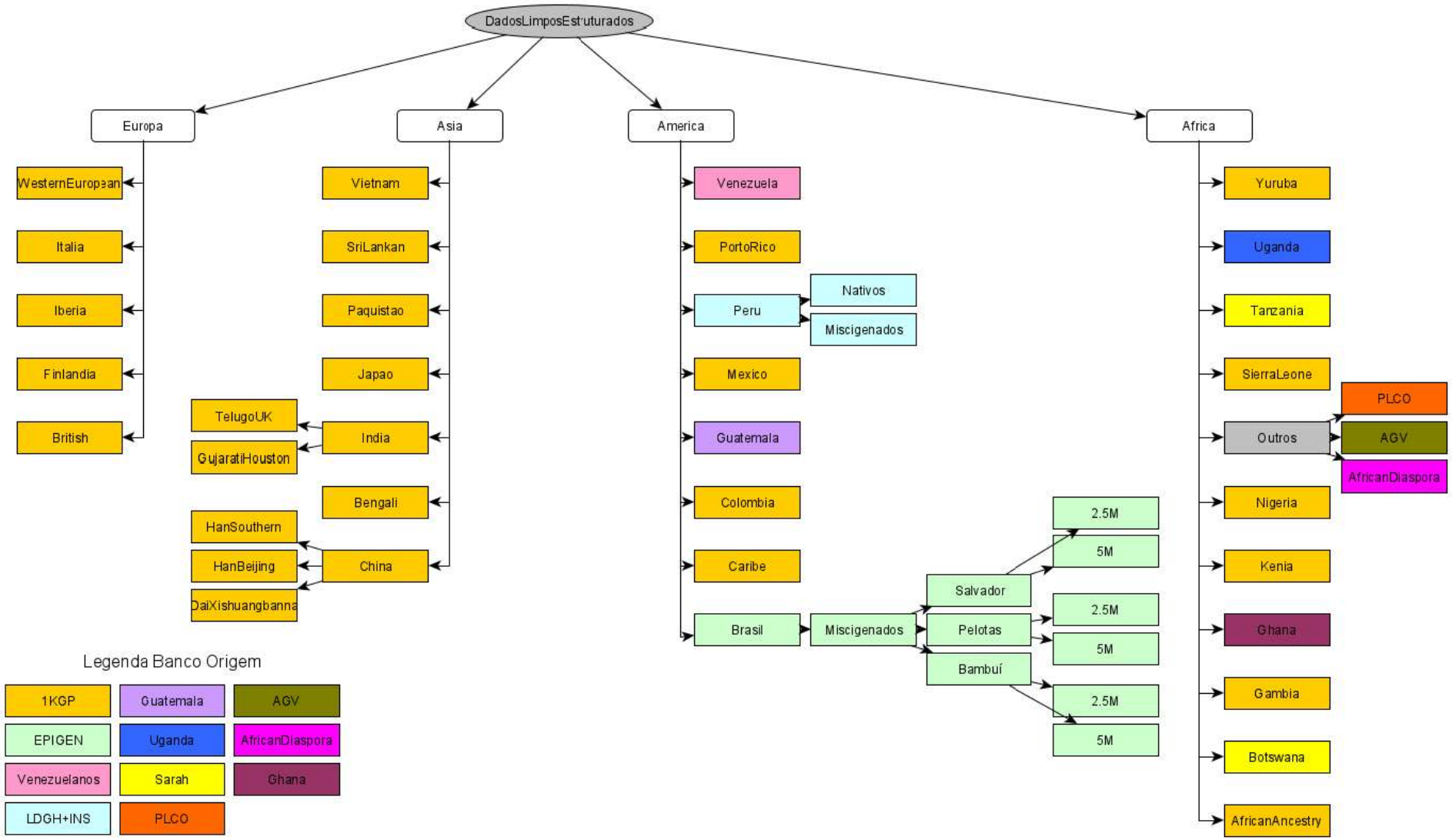


Figura 4 Mapa de organização do banco de dados para inferência de ancestralidade a nível individual e local cromossômica.

No Laboratório de Diversidade Genética Humana (LDGH), desenvolvemos um pipeline para análises de parentesco e estrutura populacional (Figura 5). Nele constam métodos baseados em genótipos e haplótipos que serão apresentados nas próximas seções deste capítulo. Atualmente contamos com parte desse pipeline automatizado, apresentado no programa 3A (https://github.com/ldgh/3A-public).

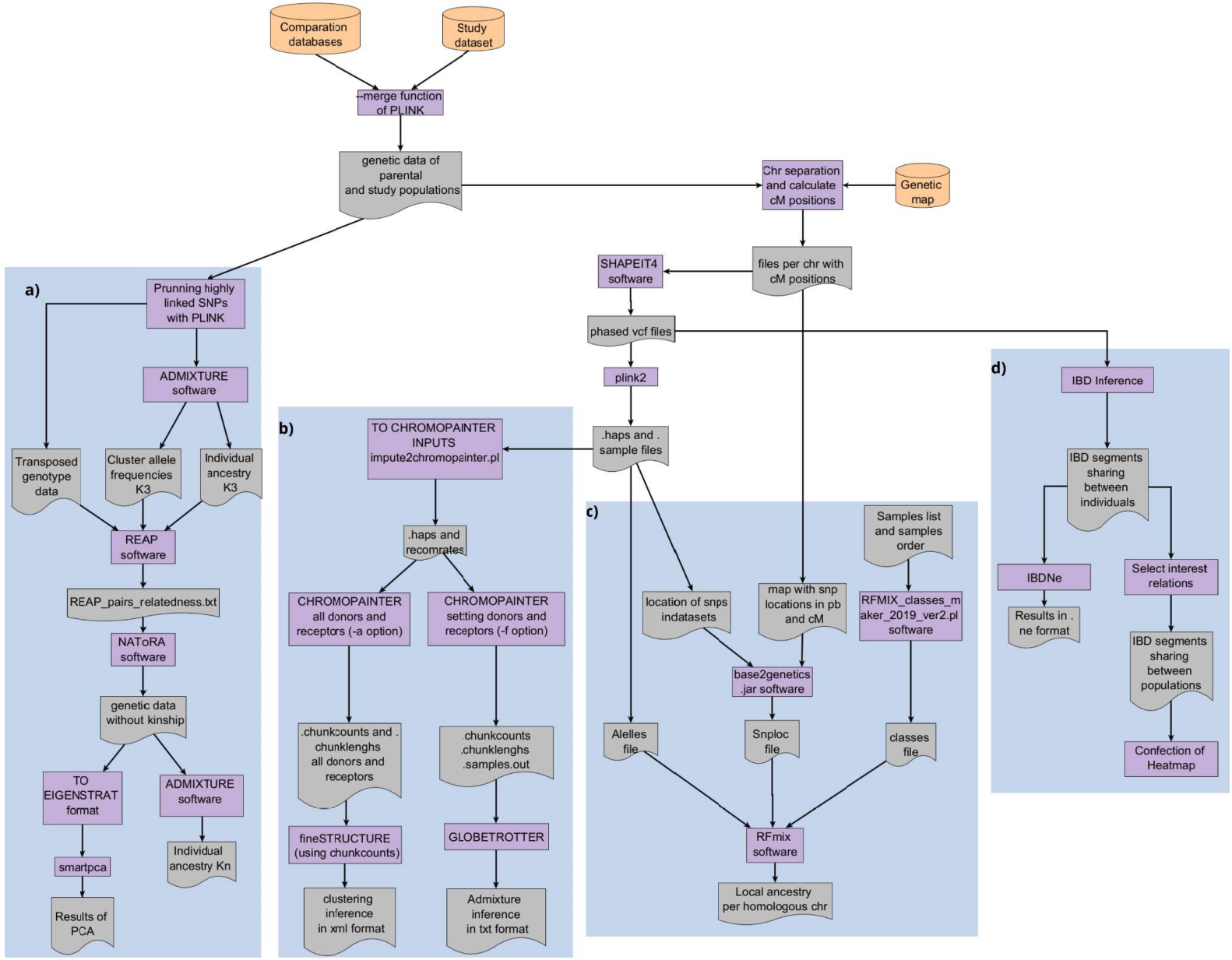

Comparation databases
Study dataset
merge function of PLINK
genetic data of parental and study populations
Chr separation and calculate cM positions
Genetic map
files per chr with cM positions
SHAPEIT4 software
phased vcf files
plink2
.haps and . sample files
a)
Prunning highly linked SNPs with PLINK
ADMIXTURE software
Transposed genotype data
Cluster allele frequencies K3
Individual ancestry K3
REAP software
REAP_pairs_relatedness.txt
NAToRA software
genetic data without kinship
TO EIGENSTRAT format
ADMIXTURE software
smartpca
Individual ancestry Kn
Results of PCA
b)
TO CHROMOPAINTER INPUTS impute2chromopainter.pl
.haps and recomrates
CHROMOPAINTER all donors and receptors (-a option)
CHROMOPAINTER setting donors and receptors (-f option)
.chunkcounts and . chunklenghs all donors and receptors
.chunkcounts .chunklenghs .samples.out
fineSTRUCTURE (using chunkcounts)
GLOBETROTTER
clustering inference in xml format
Admixture inference in txt format
c)
Samples list and samples order
location of snps indatasets
map with snp locations in pb and cM
RFMIX_classes_maker_2019_ver2.pl software
base2genetics .jar software
Alelles file
Snploc file
classes file
RFmix software
Local ancestry per homologous chr
d)
IBD Inference
IBD segments sharing between individuals
IBDNe
Results in . ne format
Select interest relations
IBD segments sharing between populations
Confection of Heatmap

Figura 5 . Pipeline de ancestralidade e estrutura populacional aplicado no LDGH. a) Pipeline de ancestralidade individual e parentesco. b) Pipeline dos métodos baseados em haplótipos CHROMOPAINTER e FineStructure. c) Pipeline Ancestralidade Local, baseada em haplótipos. d) Pipeline Identidade por Descendência, baseada em haplótipos.

1.2. Estruturação a nível familiar

1.2.1. Métodos baseados em genótipos

O cálculo do coeficiente de *inbreeding* acontece de duas formas possíveis: I) o cálculo focado na variante dentro de uma população; II) o cálculo focado no indivíduo. No primeiro caso, para cada *locus* bialélico (Hedrick 2005, Danecek *et al.* 2011), utiliza-se o número de genótipos esperados e observados para estimar o índice de fixação,

$$F_{loci} = 1 - (H_i/2p_iq_i)$$

Equação 1 Índice de fixação

no qual, para o *locus* i, p_i e q_i são as frequências alélicas, H_i é a heterozigosidade observada e $2p_iq_i$ é a heterozigosidade esperada.

Para estimar o coeficiente de *inbreeding* F para cada indivíduo, Danecek *et al*. (2011) usam a fórmula

$$F_{ind} = (O-E)/(L-E) \quad (2)$$

Equação 2 Coeficiente de inbreeding F para o indivíduo

no qual, para cada lócus bialélico, L é o número de loci usados para aquele indivíduo, O é o número de loci observados em homozigose em um indivíduo, e o número esperado de loci em homozigose baseados nas frequências alélicas p_i e q_i é dado por

$$E = 1 - 2p_iq_i$$

Equação 3 Número esperado de loci em homozigose

Considerando que é esperado que indivíduos com maior ancestralidade africana sejam mais diversos (i.e. mais heterozigotos, Campbell *et al.* 2014), em Sant'Anna *et al.* (submetido), é recomendável estratificar a estimativa de F_{ind}, considerando seis *bins* não sobrepostos de ancestralidade africana com o mesmo número de indivíduos entre eles, cada um com 32 indivíduos.

Para inferir relações de parentesco usando apenas as informações de genótipo dos indivíduos, grande parte dos métodos comparam o compartilhamento de alelos esperado com base nas frequências alélicas e observado para estimar as identidades por descendência (IBD) e por estado (*identity by state*, IBS). Como o padrão esperado de compartilhamento de alelos é mais complexo numa população miscigenada (Figura 6), alguns métodos como o REAP (*Relatedness Estimates in Admixed Populations*, Thornton *et al*. 2012) modelam a

miscigenação e estimam os coeficientes de parentesco condicionado às proporções de ancestralidade estimada por outros métodos. Uma metodologia mais recente e acurada do mesmo grupo de pesquisa, denominada PCRelate (Conomos et al. 2016), usa o Análise de Componentes Principais (PCA) e, a partir dos componentes principais e de uma matriz de parentesco inicial, estima conjuntamente seja a os coeficientes de *inbreeding* e de parentesco e as componentes principais baseadas nos genótipos que descrevem em geral a estrutura genética da população miscigenada. PCRelate explora o fato que alguns Componentes Principais, que são independentes entre si, são capazes de capturar o padrão de parentesco (IBD recente) e não a estrutura populacional.

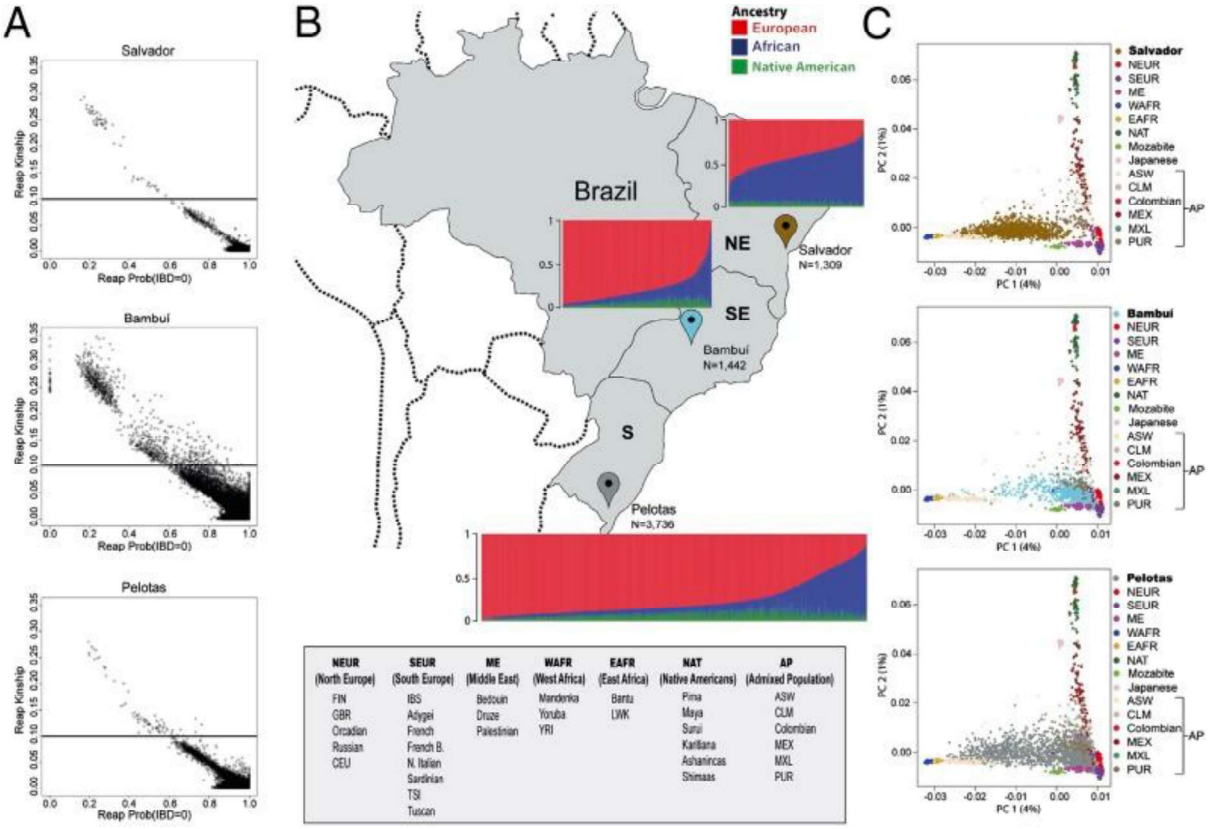


Figura 6 Admixture continental e análise de *kinship* para populações brasileiras do EPIGEN. A) Coeficientes de *kinship* para cada par de indivíduos e a probabilidade que eles compartilhem zero alelos idênticos por descendência (IBD) (IBD = 0). As linhas horizontais representam o coeficiente de *kinship* necessário que deve ser usado para serem considerados parentes. B) Regiões brasileiras, as populações estudadas e o *barplot* correspondente a sua ancestralidade continental individual. N representa o número de indivíduos do EPIGEN no banco de dados original (incluindo aparentados). C) Representação em PCA, incluindo populações do EPIGEN e globais, usando apenas indivíduos não aparentados. Os três gráficos derivam da mesma análise e são diferentes apenas pelo plot dos indivíduos do EPIGEN. AP, admixed population; ASW, Americans of African ancestry in USA; CEU, Utah residents with Northern and Western European ancestry; CLM, Colombians from Medellin, Colombia; EAFR, east Africa; FIN, Finnish in Finland; French B, Basque; GBR, British in England and Scotland; IBS, Iberian population in Spain; LWK, Luhya in Webuye, Kenya; ME, Middle East; MXL/MEX, Mexican ancestry from Los Angeles; N., (North) Italian; NAT, Native American; NE, northeast; NEUR, north Europe; PC, principal component; PUR, Puerto Ricans from Puerto Rico; S, south; SE,

southeast; SEUR, south Europe; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeira; WAFR, west Africa. Fonte: Kehdy et al. 2015.

No LDGH, por termos foco em populações da América Latina, optamos por usar o REAP, por considerar a informação de miscigenação dos indivíduos para corrigir a inferência de parentesco, evitando a superestimação de valores para indivíduos com a mesma ancestralidade.

1.2.2. Métodos baseados em haplótipos

1.2.2.1 Inferência de haplótipos

A inferência do conjunto de alelos herdados na mesma cópia de um cromossomo de um par de cromossomos homólogos é chamada de inferência de haplótipo. Os haplótipos acrescentam uma camada extra de informação que pode ser usada para refinar a inferência de estrutura populacional e ancestralidade, sendo a qualidade da inferência dos haplótipos diretamente ligada à qualidade das inferências posteriores. A observação direta dos haplótipos a partir de dados de genótipos é inviável, tornando necessária a implementação de métodos estatísticos (Browning e Browning 2015). Apesar da qualidade da inferência de haplótipos melhorar a medida que o número de marcadores aumenta, a complexidade da inferência estatística de haplótipos aumenta à medida que surgem variantes raras, sendo um dos problemas a se enfrentar para haplotipagem de dados de genômicos (Browning e Browning 2011). Além disso, à medida que a quantidade de marcadores cresce, a necessidade de capacidade computacional para processá-los também cresce.

Para lidar com esse problema, os autores da ferramenta de haplotipagem Shapeit na versão 5 trabalharam em três soluções (Hofmeister *et al*., 2023) : 1) Para variantes comuns, é usado um modelo baseado no compartilhamento de *long range haplotype* (LRH - Figura 7-A) , essa solução é a mesma usada na versão 4 do software (Delaneau *et al.* 2019); 2) Após a inferência de fase para todas as variantes comuns, as variantes raras são haplotipadas com um modelo de imputação, onde é escolhido um grupo de haplótipos informativos que contenham polimorfismo na posição do alelo raro e que sejam idênticos por descendência localmente ao indivíduo alvo (Figura 7- B); 3) Após a detecção de fase das variantes raras, os *singletons* são haplotipados usado um modelo baseado na teoria do coalescente, com o novo alelo sendo

aderido ao maior haplótipo compartilhado, assumindo um caminho geracional mais curto percorrido pelo alelo desde seu surgimento (Figura 7- C).

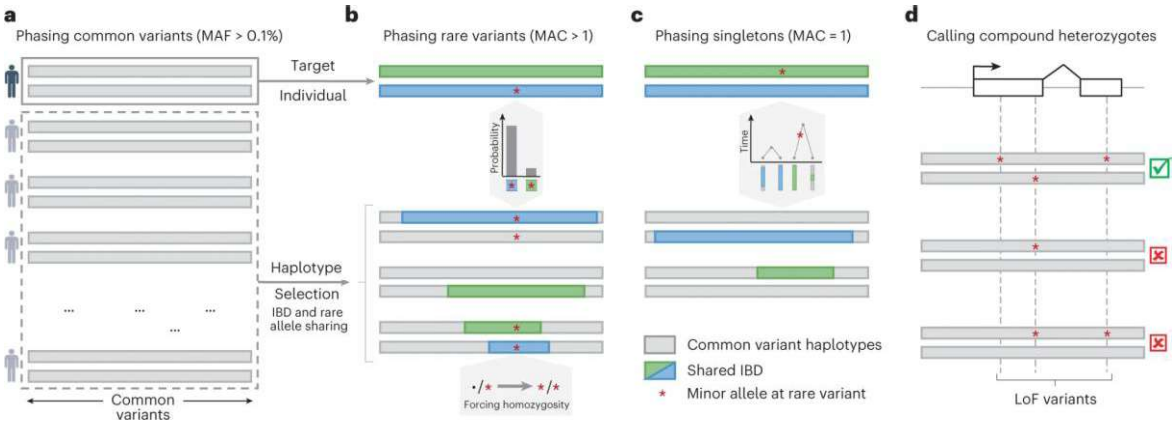


Figura 7 Base do Shapeit5 ilustrando inferência de haplótipo. a) Todas as amostras são faseadas para variantes comuns (MAF ≥ 0.1%). b) Faseamento de uma determinada variante rara nos haplótipos para variantes comuns. Os haplótipos utilizados no condicionamento da estimativa compartilham longas porções com os targets (verde e azul) e não são monomórficos para variantes raras. Como os genótipos heterozigotos para as variantes raras não estão faseados, assume-se que os alelos menores nessas regiões estão nos dois haplótipos (homozigose forçada). c) Faseamento de singleton assumindo o novo alelo no haplótipo target com a correspondência mais curta. d) Mapeamento de eventos de heterozigotos compostos baseados no faseamento de variantes raras (a-c). Fonte: Hofmeister et al., 2023.

Outros softwares usam outras estratégia, como a segunda versão do Eagle, que compacta o banco de haplótipos de referência em grupos de estruturas de árvores compactas que evitam perdas na estrutura do haplótipo, ou a versão 5.2 do Beagle que adota um algoritmo progressivo de faseamento para melhorar a performance diante de um grande número de variantes (De Marino *et al.* 2022). Porém, apesar de rápidos, a eficácia do Shapeit 5 e sua capacidade de inferência direcionada para trios o mantém no pipeline do LDGH.

1.2.2.2 Métodos de inferência de estrutura genética a nível familiar

As relações de parentesco apresentadas anteriormente se baseiam em informações normalmente de dezenas ou centenas de milhares de genótipos independentes, com os quais é mais difícil diferenciar quais alelos são realmente IBD e quais IBS. Com o aumento da densidade de variantes genéticas nos bancos de dados, em particular com as tecnologias atuais de sequenciamento, a inferência de haplótipos se torna mais acurada, e por sua vez, maior a probabilidade de encontrar fragmentos provenientes do mesmo ancestral. A inferência de

fragmentos IBD além de contribuir para a identificação das relações familiares pode revelar a estrutura populacional fina. O IBD tem sido usado, por exemplo, para inferir o local de nascimento dos indivíduos (Nait Saada *et al.*, 2020). Além disso, ao observar o tamanho dos fragmentos de IBD compartilhados entre os indivíduos é possível identificar o tempo geracional do ancestral comum mais recente entre eles, como ilustraremos no capítulo 2 desta tese no estudo de populações nativas peruanas.

A evolução dos métodos de inferência IBD passa por melhorias da qualidade da inferência e diminuição do tempo computacional, porém, assim como os demais métodos baseados em haplótipos, necessita de uma alta densidade de marcadores (Browning; Browning, 2012). Inicialmente, a inferência do IBD em nosso pipeline acontecia através do software RefinedIBD, que se apresentava adequado para inferência baseada em arrays de alta densidade, porém não é capaz de fazer o aproveitamento completo das informações do genoma completo, como das variantes raras (Browning; Browning, 2013). Dessa forma, nossas análises passam no momento pela adequação para a metodologia hapIBD (Figura 3d), do mesmo grupo de pesquisa (Zhou; Browning; Browning, 2020), que torna a inferência de IBD escalável para grandes biobancos, além de aumentar a acurácia da detecção de fragmentos entre 2-4 cM, que no RefinedIBD só era possível com várias corridas repetidas do software. Além disso, o hapIBD tem demonstrado uma taxa menor de fragmentos IBD falsos positivos quando comparado a outras metodologias atuais e escaláveis, como o phasedIBD (Freyman *et al.*, 2020).

1.3. Estruturação a nível populacional

1.3.1. Métodos baseados em genótipo:

O PCA é um tradicional método não baseado em modelo que reduz a dimensionalidade de dados complexos. No caso das informações genéticas de um grupo de indivíduos, essa redimensionalização mostra a proximidade genética entre os indivíduos que reflete a estrutura populacional e eventualmente, as relações de parentesco (Price *et al.* 2019, Patterson *et al.* 2006).

Um dos modelos mais usados para inferir as proporções de ancestralidade de um indivíduo baseando-se em genótipos independentes assume que um indivíduo é formado por até K subpopulações parentais ou clusters de ancestralidade. Ele estima as frequências alélicas

de cada população K, ajustando as frequências alélicas para que se adaptem ao equilíbrio de Hardy-Weinberg em cada subpopulação parental e maximizando o F_{ST} (a variância estandardizada das frequências alélicas) entre as K subpopulações parentais. Após estimar as frequências alélicas, o modelo é capaz de inferir a proporção de ancestralidade de cada população K em cada indivíduo (Figura 7). Este modelo foi adotado pelo método STRUCTURE (Pritchard; Stephens; Donnelly, 2000), que usava Cadeia de Markov Monte-Carlo (MCMC) para estimar o valor ótimo de *likelihood,* e posteriormente foi adotado pelo software ADMIXTURE (Alexander, Novembre e Lange 2009), que usa *Expectation-Maximization*, o que permite um número maior de variantes genéticas como entrada.

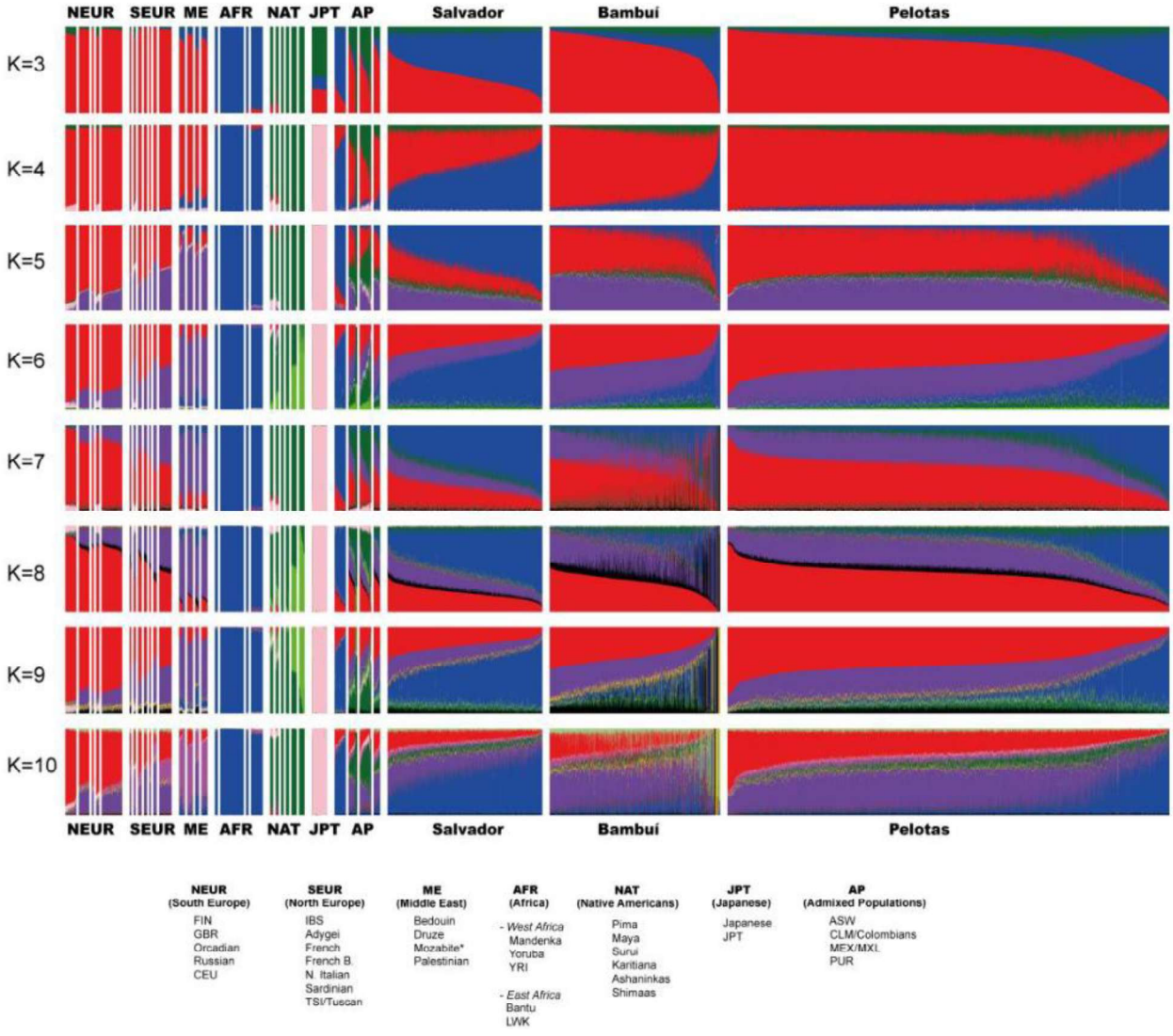


Figura 8 Barplot representando as proporções de ancestralidade individual para os indivíduos do projeto EPIGEN. As proporções de valores de ancestralidade individual foram calculadas usando de K=3 até K=10 para o banco de dados original. As populações de cada grupo étnico/geográfico são descritas abaixo da figura na mesma ordem que aparecem na imagem. Cada barra representa um indivíduo e cada cor um cluster específico de ancestralidade. *Moabitas são uma população do noroeste africano. Fonte: Kehdy et al., 2015.

Nosso pipeline mantém a inferência de ancestralidade individual pelo software ADMIXTURE (Figura 3a). Através dos valores de *Cross-Validation Error* gerados pelo software é possível identificar o número de populações parental mais adequado para descrever as populações de um banco de dados. Outros métodos como o Graf (Jin *et al.*, 2017) tentam fazer uma abordagem mais rápida para dividir os indivíduos entre grupos pré-determinados, usando um painel SNVs informativos de ancestralidade para classificar os indivíduos. Entretanto, na nossa experiência, essa abordagem perde poder e comumente não consegue classificar corretamente indivíduos que são miscigenados homogeneamente entre mais de três populações ancestrais, o que se torna um limitante para populações diversas como a da América Latina.

1.3.2. Métodos baseados em haplótipos:

Nesta seção, exploraremos os métodos que usam padrões de similaridade dos haplótipos para detecção de estrutura populacional. Um dos métodos de inferência é o CHROMOPAINTER (Lawson *et al.*, 2012), no qual é feita a contagem e verificação do comprimento de segmentos de haplótipos (*chunks*) compartilhados entre os indivíduos. Assim como os fragmentos IBD podem ser usados para detectar grupos populacionais, os *chunks* cromossômicos podem servir ao mesmo propósito, porém sem o poder de detecção de estrutura familiar. A partir da contagem de *chunks* compartilhados entre os indivíduos é construída uma matriz de coancestralidade que serve como *input* para o algoritmo fineSTRUCTURE, que detecta estruturação populacional fina, eventualmente permitindo a reconstrução da relação filogenética entre as populações. A matriz de coancestralidade também permite identificar a ancestralidade populacional, porém não permite a inferência da ancestralidade a nível local cromossômico. Em nosso grupo, usamos fineStructure e o CHROMOPAINTER (Figura 3b) para identificar mais finos de estruturação africana nas Américas (Gouveia *et al.* 2020), e em particular, porque a metodologia GLOBETROTTER permite uma datação dos eventos de miscigenação (Figura 9).

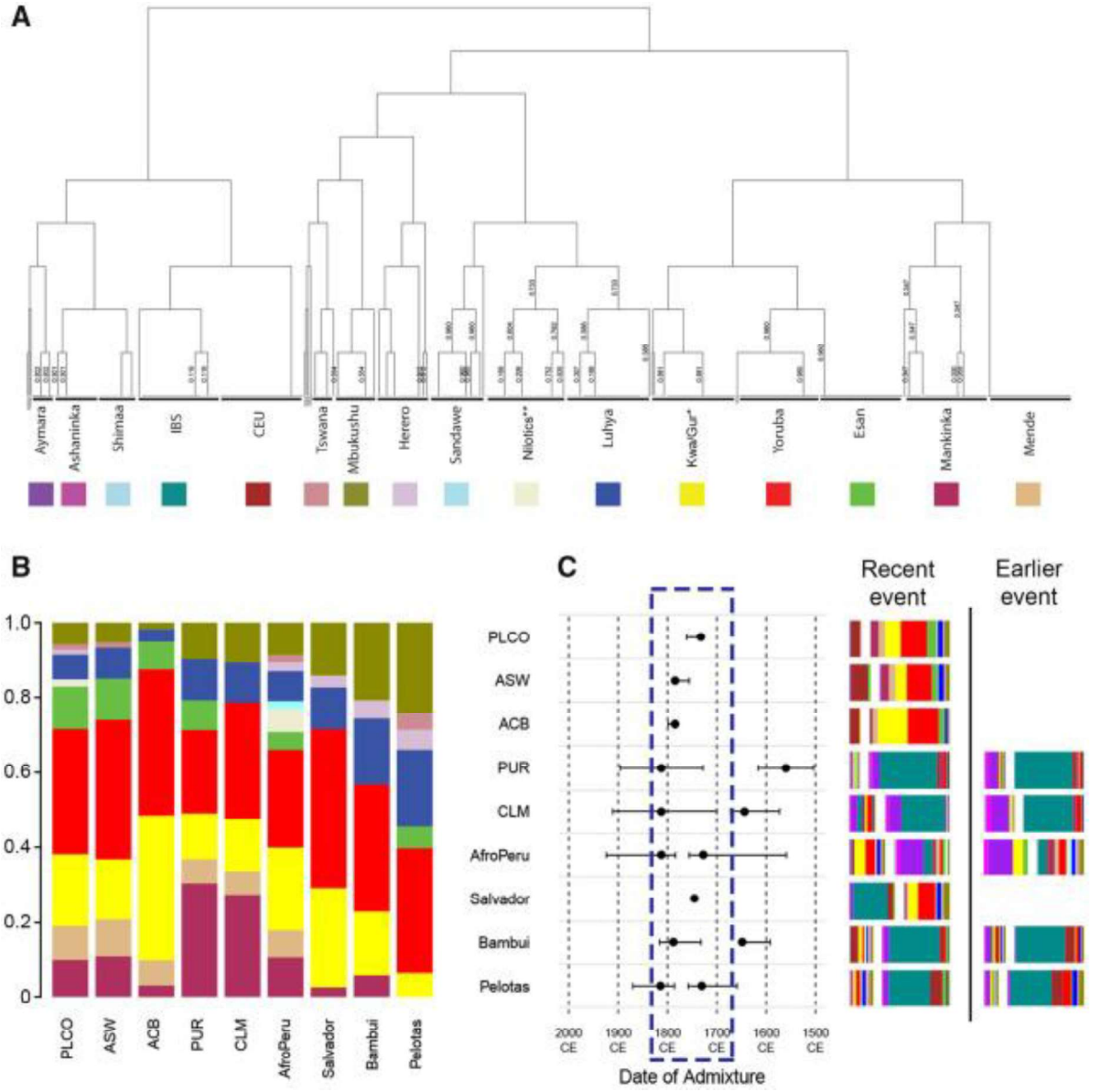


Figura 9 Clustering baseado em haplótipo de populações parentais e inferência de admixture para populações americanas continentais miscigenadas. A) Árvore gerada com fineSTRUCTURE de indivíduos parentais. B) Contribuições de populações subcontinentais relativas à ancestralidade africana total em populações miscigenadas inferidas pela regressão MIXTURE MODEL implementada no GLOBETROTTER. C) Inferências de GLOBETROTTER de eventos de miscigenação para cada população miscigenada. A data inferida e os intervalos de 95% de confiança são representados por pontos e linhas horizontais no gráfico. O retângulo tracejado ilustra o período mais dinâmico de miscigenação. Além do gráfico da datação, estão representadas as fontes inferidas de miscigenação para eventos recentes e anteriores. O tamanho das barras representa a contribuição genética de cada fonte. Cada cor corresponde à proporção da contribuição de cada população parental. CEU, Utah Residents (CEPH) with Northern and Western Ancestry-United States, IBS, Iberian population in Spain; CLM, Colombians from Medellin; PUR, Puerto Ricans from Puerto Rico; ACB, African Caribbeans in Barbados; ASW, African Americans in Southwest United States; PLCO, African Americans from East United States. Fonte: Gouveia et al., 2020.

A estimativa da ancestralidade local cromossômica requer inferência de haplótipos. Apesar da literatura não indicar um número mínimo de variantes para inferir a ancestralidade local, a acurácia da inferência está diretamente associada ao número de variantes e à sua distribuição ao longo do cromossomo, uma vez que regiões com baixa densidade de variantes

podem gerar segmentos contínuos de uma determinada ancestralidade com tamanhos maiores que o real. Como parâmetro mínimo para inferência de ancestralidade no LDGH, nós adotamos um mínimo de 600 mil marcadores genéticos compartilhados entre o banco de dados alvo e a referência.

Para inferência de ancestralidade local (Figura 10), usamos o software RFMix (Maples *et al.* 2013; Figura 3c). Esse método trabalha dividindo o cromossomo em janelas, que terão sua ancestralidade inferida a partir de um campo aleatório condicional, parametrizado por *random forests* treinadas de acordo com o banco de dados de referência. A segunda versão do software apresenta um componente de *Machine Learning* para adaptar o método a bancos de dados maiores, porém não chegou a ser publicada. Outra alternativa para inferência de ancestralidade local cromossômica é o software Gnomix (Hilmarsson *et al.*, 2021), que usa ferramentas de inteligência artificial para solucionar o problema da escalabilidade para grandes bancos de dados como o UKBiobank (Sudlow *et al.*, 2015).

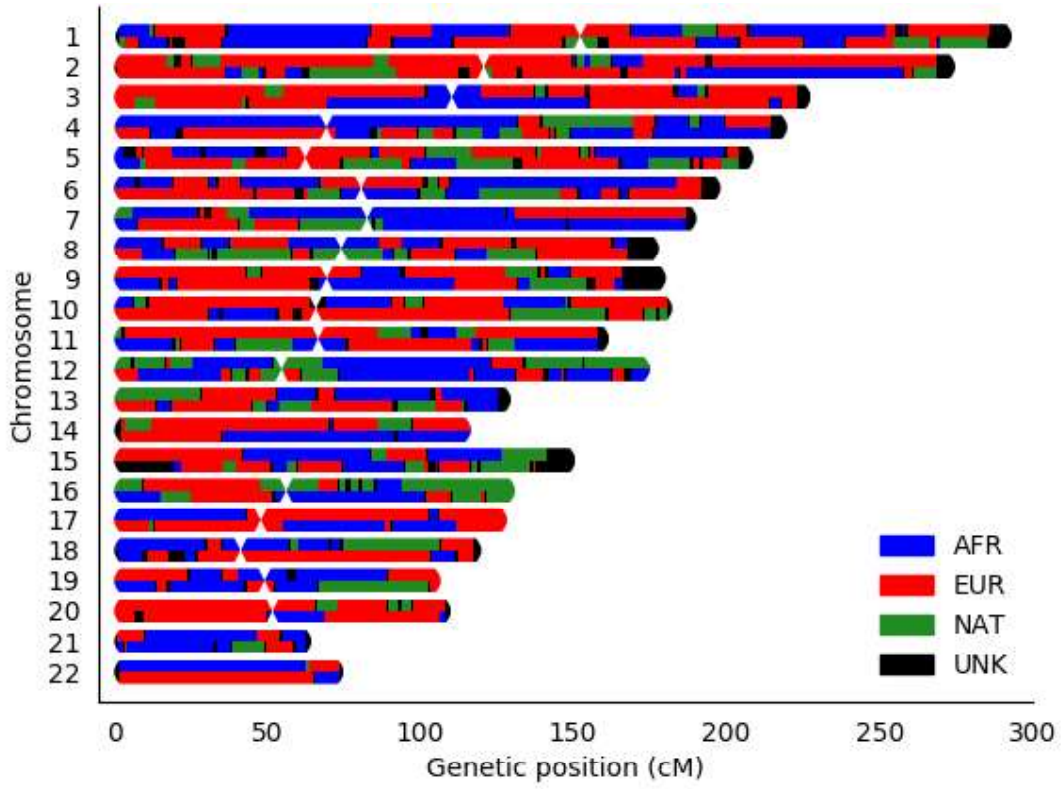


Figura 10 Plot de cariograma esquemático representando a ancestralidade local de um indivíduo brasileiro calculada utilizando RFMIX. As cores representam as ancestralidades parentais em cada região genômica para cada um dos 22 cromossomos autossômicos em um indivíduo brasileiro miscigenado. A ancestralidade local foi calculada utilizando RFMIX e o plot gerado utilizando o programa 3A do grupo. Fonte: Autoria do grupo.

CAPÍTULO 2: Genética de populações e medicina evolutiva em nativos americanos peruanos

2.1. Introdução

O cenário mundial de estudos genômicos permanece desigual, sendo a maioria dos artigos já publicados sobre populações europeias. O apelo da comunidade científica por estudos que sejam mais globalmente representativos cresce a cada dia mais (Sirugo *et al.*, 2019, Alvim *et al.*, 2024), uma vez que os resultados de estudos que englobam participantes de etnias sub representadas revelam informações genéticas que ainda não tinham sido detectadas anteriormente, como novas associações que não poderiam ser reveladas em um estudo que só representasse a população europeia (Scliar *et al.* 2021; Naslavsky *et al.* 2022b; Leal *et al.* 2023). Dentre as etnias sub representadas nos estudos genômicos, as populações Nativas Americanas, que consistem nas populações que já ocupavam as Américas antes da chegada dos europeus (Mills e Rahal 2020; Corpas *et al.* 2024)

Diante da necessidade de diminuir a disparidade populacional dentro dos dados genômicos, surge o *Peruvian Genome Project* (PGP), iniciativa do Ministério da Saúde Perunano, que através do Laboratório de Diversidade Genética Humana envolve uma colaboração entre Universidade Federal de Minas Gerais (UFMG) e Instituto Nacional de Saúde Peruano (INS). Desde 2011, o desenvolvimento do PGP busca explorar a diversidade genética peruana de indivíduos nativo-americanos e miscigenados (Guio *et al.* 2024 - Pré-print; Anexo 2). Nele, estão presentes amostras das ecorregiões Costeira, Andina e Amazônica peruanas, daqui em diante definida como Oeste Sul-Americano (Figura 11).

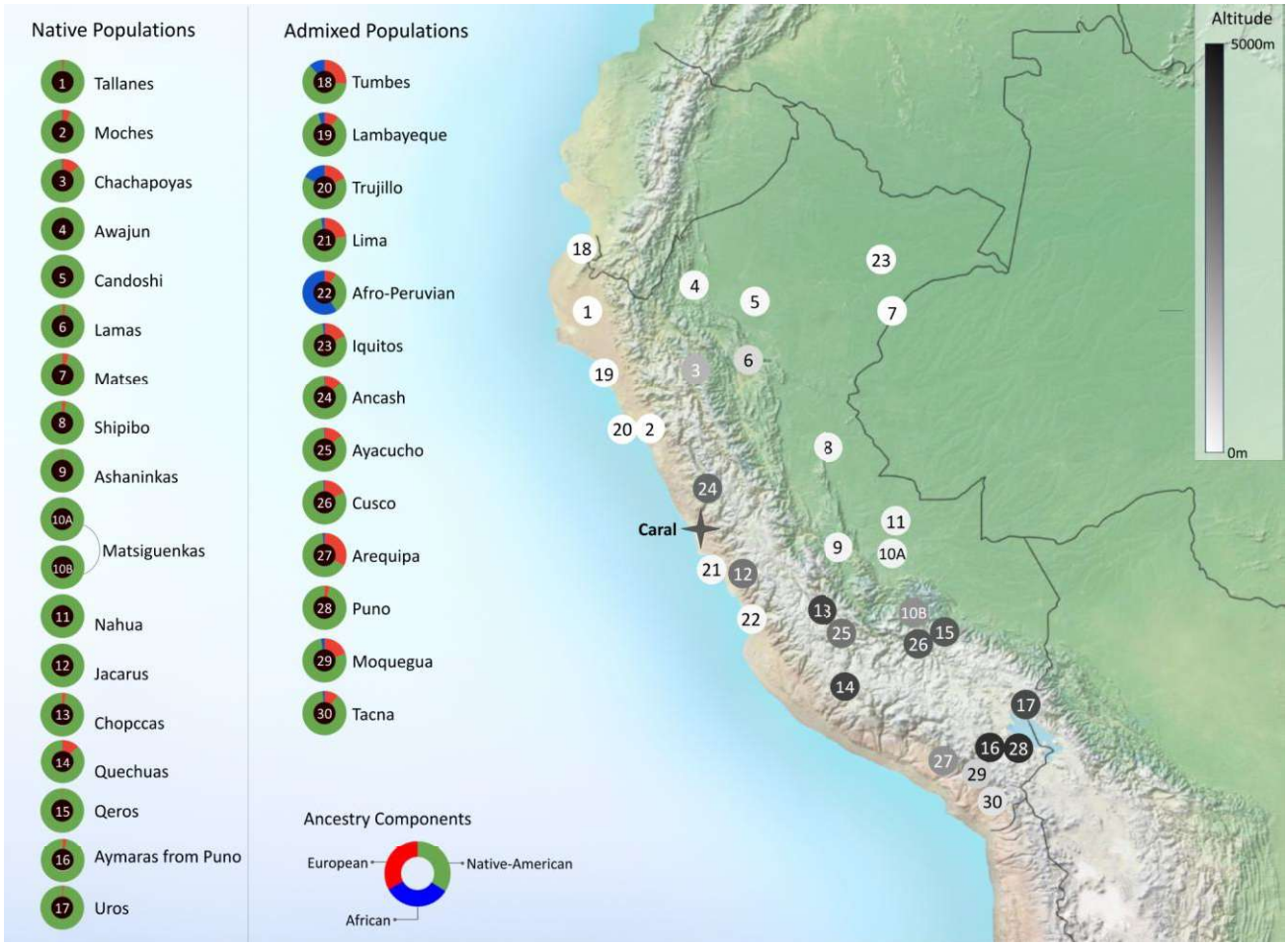


Figura 11 Proporção de ancestralidade global e distribuição geográfica das populações Peruanas selecionadas para o Projeto Genoma Peruano. Os gráficos de rosca mostram a proporção de miscigenação de cada população, inferida pela média do resultados de K= 3 do software ADMIXTURE. Os componentes de ancestralidade de origem Africana, Europeia e Nativo-Americana estão representados nos gráficos em azul, vermelho e verde, respectivamente. O mapa adjacente mostra a localização das diferentes populações em círculos numerados coloridos com diferentes intensidades de cinza para indicar a média de altitude da posição de cada população. Gerado para: Guio et. al. 2024, Pré-print.

Entre as populações nativo-americanas, as do Oeste Sul-Americano, um dos berços da civilização humana (Solis, Haas e Craemer 2001; Roustain *et al*. 2024), se destacam pelo povoamento de dois biomas contrapostos: a cordilheira andina e a floresta amazônica. Os Andes apresentam um ambiente árido, com alta incidência de radiação UV, hipóxia e frio, características que dificilmente seriam contornadas por adaptações culturais. A Amazônia proporciona um ambiente quente, úmido, com baixa incidência solar e alta biodiversidade, refletida nos patógenos humanos.

Porém, é importante ressaltar que tanto o ambiente andino quanto o amazônico possuem regiões de transição, onde características de ambos estão presentes simultaneamente. Os Andes Centrais podem ser divididos latitudinalmente em Central Árido, que coincide com as maiores altitudes habitadas no Peru, e o Norte Fértil, onde a cadeia de montanhas perde

altitude (Lumbreras 2015). No Leste Andino, por outro lado, existe a região da Baixa Amazônia, onde estão as regiões alagáveis, e a região de Yunga, também conhecida como Alta Amazônia, onde acontece uma transição longitudinal entre Andes e Amazônia.

Estudos realizados com populações dos Andes Centrais mostraram a homogeneidade genética interpopulacional nesta região, homogeneidade que também se reflete na pouca diversidade linguística. Nos Andes Áridos são faladas apenas duas famílias linguísticas, enquanto na Amazônia pode ser observada uma maior diversidade genética e linguística (Campbell 2000; Tarazona-Santos *et al.* 2001; Harris *et al.* 2018). Dessa forma, vemos claramente uma dicotomia Andes/Amazônia descrita na literatura. Porém, estudos antropológicos mostram a existência de interações socioculturais entre estas populações, principalmente na região norte peruana (Solis, Haas e Creamer 2001; Lumbreras 2015).

Frente a esta dicotomia leste/oeste de ambientes e diversidade genética, além da evidência de interação sociocultural, uma série de questões históricas e adaptativas a respeito destas populações pode ser levantada. Entre elas estão: I) A homogeneização genética entre as populações andinas e a dicotomia genética encontrada entre as populações dos Andes Centrais e as populações da Amazônia se estende para a região norte peruana? II) As interações culturais entre Andes e Amazônia foram acompanhadas de fluxo gênico? III) Quando a homogeneização das populações dos Andes centrais começou e qual sua relação com a história andina? IV) Como a adaptação destas populações aos ambientes andino e amazônico se reflete no genoma? Essas quatro perguntas foram exploradas no manuscrito anexado no tópico 2.2 deste capítulo, no qual compartilho a primeira autoria com os doutores Victor Borda, Marla Mendes e Isabela Alvim. Ele foi publicado em 2020 na revista *Proceedings of the National Academy of Sciences* (PNAS), mostramos que este contraste ambiental fez com que as populações nativas dessa região fossem moldadas geneticamente pela interação entre ambiente, geografia e cultura.

A elucidação da estrutura populacional nativo-peruana trazida em Borda *et al.* 2020 reforça a existência de diversidade entre as populações nativas americanas, que já foram erroneamente tratadas como homogêneas dentro das três Américas (de Ulloa, 1772; Hrdlička, 1928). Além disso, os resultados dos testes de seleção natural e de diferenciação entre as populações abrem as portas para um problema mais profundo: a variabilidade de frequências alélicas entre as populações nativas e seu impacto nas variantes de importância médica e farmacêutica. Para enriquecer o apelo por estudos em populações de origem não-europeia, em mais uma primeira autoria compartilhada com a doutora Isabela Alvim, publicamos um

comentário na revista *Cell* (tópico 2.3 deste capítulo) onde evidenciamos a diferença entre as frequências alélicas em populações andinas e amazônicas de duas variantes clinicamente acionáveis de importância farmacogenética: o rs2231142, localizado no gene *ABCG2*, e o rs9923231, localizado no gene *VKORC1*.

2.2. The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture

2.2.1. Contribuições

Como uma das primeiras autoras no artigo "*The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture*", participei das discussões e reuniões para desenvolvimento do trabalho, da escrita e do processo de submissão e revisão. Tive participação na execução de alguns dos métodos aplicados no estudo, entre eles ajudei ao autor Victor Borda na aplicação de controles de qualidade aos bancos de dados após a incorporação de dados de novas populações, aplicação de testes de Análise de Componente Principais (PCA) (Patterson, Price, e Reich 2006) e ADMIXTURE (Alexander, Novembre, e Lange 2009), além da aplicação da estatística D (Green *et al*. 2010; Durand *et al.* 2011). Juntamente ao autor Giordano Bruno Soares Souza apliquei testes de *cross-population Extended Haplotype Homozygosity* (XP-EHH), para detecção de seleção natural monogênica. Além disso, com ajuda da autora Isabela Alvim, fiz a análise de desequilíbrio de ligação de dados de sequência dos genes detectados sob seleção natural para a versão submetida do artigo. Além disso, atuei na anotação de variantes detectadas pelos testes de seleção natural monogênica, utilizando a ferramenta de anotação Sistema Multiagente para Anotação de SNPs (MASSA - Soares-Souza, 2014).

2.2.2. Artigo

Link para o artigo online: The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture | PNAS

Os materiais suplementares deste artigo se encontram no Anexo 2.1 da presente tese.

The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture

Víctor Borda[a,b,c,1], Isabela Alvim[a,1], Marla Mendes[a,1], Carolina Silva-Carvalho[a,1], Giordano B. Soares-Souza[a], Thiago P. Leal[a], Vinicius Furlan[a], Marilia O. Scliar[a,d], Roxana Zamudio[a], Camila Zolini[a,e,f], Gilderlanio S. Araújo[g], Marcelo R. Luizon[a], Carlos Padilla[c], Omar Cáceres[c,h], Kelly Levano[c], César Sánchez[c], Omar Trujillo[i], Pedro O. Flores-Villanueva[c], Michael Dean[j], Silvia Fuselli[k], Moara Machado[a,j], Pedro E. Romero[l], Francesca Tassi[k], Meredith Yeager[j], Timothy D. O'Connor[m,n,o], Robert H. Gilman[l,p], Eduardo Tarazona-Santos[a,l,q,2,3], and Heinner Guio[c,h,r,2,3]

[a]Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil; [b]Laboratório de Bioinformática (LABINFO), Laboratório Nacional de Computação Científica, Petrópolis, RJ, 25651-076, Brazil; [c]Laboratorio de Biotecnología y Biología Molecular, Instituto Nacional de Salud, Lima 9, Peru; [d]Human Genome and Stem Cell Research Center, Biosciences Institute, University of São Paulo, São Paulo, SP, 05508-090, Brazil; [e]Beagle, Belo Horizonte, MG, 31270-901, Brazil; [f]Mosaico Translational Genomics Initiative, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil; [g]Laboratório de Genética Humana e Médica, Programa de Pós Graduação em Genética e Biologia Molecular, Universidade Federal do Pará, Belém, PA, 66075-110, Brazil; [h]Carrera de Medicina Humana, Facultad de Ciencias de la Salud, Universidad Científica del Sur, Lima, 150142, Peru; [i]Centro Nacional de Salud Intercultural, Instituto Nacional de Salud, Lima 11, Peru; [j]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD 20892; [k]Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara, 44121, Italy; [l]Universidad Peruana Cayetano Heredia, Lima 31, Peru; [m]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201; [n]Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD 21201; [o]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201; [p]Department of International Health, Johns Hopkins School of Public Health, Baltimore, MD 21205; [q]Instituto de Estudos Avançados Transdisciplinares, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil; and [r]Facultad de Ciencias de la Salud, Universidad de Huánuco, Huánuco, 10001, Peru

Edited by Anne C. Stone, Arizona State University, Tempe, AZ, and approved October 26, 2020 (received for review July 9, 2020)

Western South America was one of the worldwide cradles of civilization. The well-known Inca Empire was the tip of the iceberg of an evolutionary process that started 11,000 to 14,000 years ago. Genetic data from 18 Peruvian populations reveal the following: 1) The between-population homogenization of the central southern Andes and its differentiation with respect to Amazonian populations of similar latitudes do not extend northward. Instead, longitudinal gene flow between the northern coast of Peru, Andes, and Amazonia accompanied cultural and socioeconomic interactions revealed by archeology. This pattern recapitulates the environmental and cultural differentiation between the fertile north, where altitudes are lower, and the arid south, where the Andes are higher, acting as a genetic barrier between the sharply different environments of the Andes and Amazonia. 2) The genetic homogenization between the populations of the arid Andes is not only due to migrations during the Inca Empire or the subsequent colonial period. It started at least during the earlier expansion of the Wari Empire (600 to 1,000 years before present). 3) This demographic history allowed for cases of positive natural selection in the high and arid Andes vs. the low Amazon tropical forest: in the Andes, a putative enhancer in *HAND2-AS1* (heart and neural crest derivatives expressed 2 antisense RNA1, a noncoding gene related to cardiovascular function) and rs269868-C/Ser1067 in *DUOX2* (dual oxidase 2, related to thyroid function and innate immunity) genes and, in the Amazon, the gene encoding for the CD45 protein, essential for antigen recognition by T and B lymphocytes in viral–host interaction.

Native Americans | human population genetics | natural selection | gene flow

Significance

Native Americans are neglected in human genetics studies, despite recent interest in the study of ancient DNA of their ancestors. Our findings on Andean and Amazonian populations exemplify how the current pattern of genetic diversity in human populations is influenced by the interaction of history and environment. In the present case, this pattern is influenced by 1) altitudinal and climatic differences among the northern, lower, and fertile Andes versus the southern, higher, and arid Andes and 2) the sharp differences between the Andean highlands and the Amazon lowlands, where natural selection and other evolutionary forces acted for millennia, shaping differences in the frequencies of genetic variants related to immune response, drug response, and cardiovascular and hematological functions.

Living Native Americans, the object of this study, are among the most neglected populations in human genetics studies, despite the increasing interest in the study of ancient DNA (aDNA) of their ancestors (1, 2). Western South America was one of the cradles of civilization in the Americas and the world (3). When the Spanish conqueror Francisco Pizarro arrived in 1532, the pan-Andean Inca Empire ruled in the Andean region and had achieved levels of socioeconomic development and

Author contributions: V.B., C.P., O.C., C.S., E.T.-S., and H.G. designed research; V.B., I.A., M.Me., C.S.-C., G.B.S.-S., T.P.L., R.Z., C.Z., G.S.A., M.R.L., C.P., O.C., K.L., C.S., O.T., P.O.F.-V., S.F., M.Ma., P.E.R., and F.T. performed research; G.B.S.-S., T.P.L., V.F., M.D., S.F., M.Ma., M.Y., and R.H.G. contributed new reagents/analytic tools; V.B., I.A., M.Me., C.S.-C., M.O.S., C.P., O.C., C.S., P.E.R., F.T., T.D.O., and H.G. analyzed data; V.B., I.A., M.Me., and E.T.-S. wrote the paper; G.B.S.-S., T.P.L., V.F., G.S.A., M.R.L., S.F., and M.Ma. performed analysis or provided bioinformatics resources for the analyses; R.Z., K.L., O.T., P.O.F.-V., and R.H.G. collected samples, processed them, or generated the genetic data; C.Z. and E.T.-S. coordinated the research teams in Brazil; C.P., O.C, and C.S. collected samples, processed them, and generated genetic data; M.D. and M.Y. provided unpublished comparative datasets; H.G. was the coordinator of the Peruvian Genome Project; and all authors read different versions of the manuscript, providing suggestions and discussing it.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

[1]V.B., I.A., M.Me., and C.S.-C. contributed equally to this work.

[2]E.T.-S. and H.G. contributed equally to this work.

[3]To whom correspondence may be addressed. Email: edutars@gmail.com or heinnerguio@gmail.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2013773117/-/DCSupplemental.

First published December 4, 2020.

Downloaded from https://www.pnas.org by UNIVERSIDADE FEDERAL DE MINAS GERAIS on June 12, 2024 from IP address 150.164.28.250.

population density unmatched in other parts of South America. The Inca Empire, which lasted for around 200 years before the conquest, with its emblematic architecture such as Machu Picchu and the city of Cuzco, was just the "tip of the iceberg" of a millenary cultural and biological evolutionary process (4, 5). This process started 11,000 to 14,000 years ago (6–8) with the peopling of this region, hereafter called western South America, that involves the entire Andean region and its adjacent and narrow Pacific coast.

Tarazona-Santos et al. (9) proposed in 2001 that cultural exchanges and gene flow along time have led to a current relative genetic, cultural, and linguistic homogeneity between the populations of western South America compared with those of eastern South America (a term that hereafter refers to the region adjacent to the eastern slope of the Andes and eastward, including Amazonia), where populations remained more isolated from each other. For instance, only two languages (Quechua and Aymara) of the Quechumaran linguistic stock predominate in the entire Andean region, whereas in eastern South America natives speak a different and broader spectrum of languages classified into at least four linguistic families (5, 9, 10). This spatial pattern of genetic diversity and its correlation with geography and environmental, linguistic, and cultural diversity was confirmed, enriched, and rediscussed by us and others (2, 4, 5, 9–15).

There are, however, pending issues. The first is whether the current dichotomic organization of genetic variation characterized by the between-population homogeneous southern Andes vs. between-population heterogeneous central Amazon extends northward. This is important because scholars from different

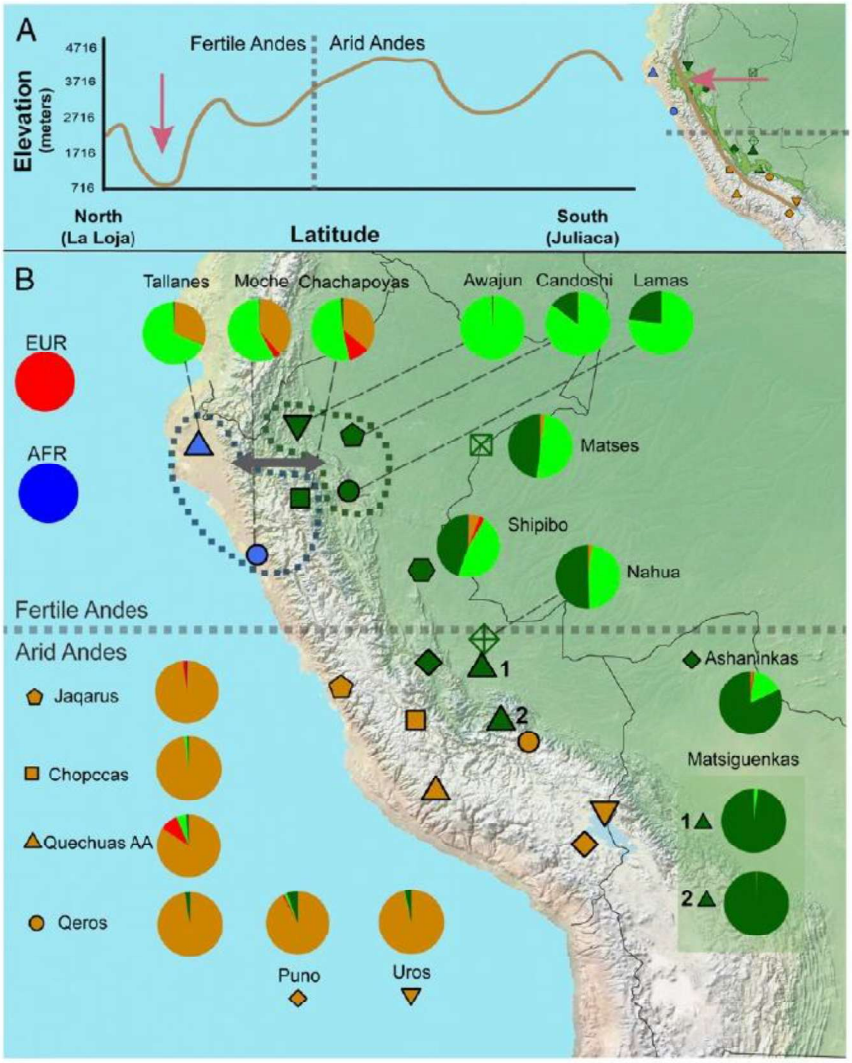


Fig. 1. Genetic and geographic landscape for western South American natives. (*A*) Elevation vs. latitude plot from a cross-section line in the Andean region from La Loja (Ecuador) to Juliaca (Peru). A vertical line indicates the division between Fertile and Arid Andes (16). The map shows the path used to create the plot elevation and latitude plot and a green area delimiting the area of the Amazon Yunga region. Altitude data were obtained from Google Maps (https://www.google.com.br/maps). (*B*) Geographic distribution and genetic structure for 18 native populations from the coast, Andes, and Amazon inferred by ADMIXTURE result (K = 5, corresponding to the lowest cross-validation error). Pie charts show the average percentages over individuals for the five ADMIXTURE clusters in each population. Three clusters were related to Native American groups: one Andean (brown) and two Amazonian (green clusters). Two clusters were associated with non-Native continental ancestries (European [red] and African [dark blue]). Blue and green dashed lines delimit the groups that showed a highly significant *D* statistic value, indicating gene flow (gray arrow, |Z score| > 4, *SI Appendix*, Figs. S14–S17). Matsiguenkas 1= Matsiguenkas-Sepahua, Matsiguenkas 2= Matsiguenkas-Shimaa. The gray horizontal dashed line in the center of the map shows the approximate division between the fertile Andes and arid Andes (16).

Downloaded from https://www.pnas.org by UNIVERSIDADE FEDERAL DE MINAS GERAIS on June 12, 2024 from IP address 150.164.28.250.

disciplines emphasize that western South America is not latitudinally homogeneous, differentiating a northern and, in general, lower and wetter fertile Andes and a southern, higher, and more arid Andes (16) (Fig. 1*A*). These environmental and latitudinal differences are correlated with demography and culture, including different histories and spectra of domesticated plants and animals. Indeed, the development of agriculture, in the first urban centers such as Caral (3) and its associated demographic growth, occurred earlier in the northern fertile Andes (around 5,000 years ago) than in the southern arid Andes (and their associated coast), with products such as cotton, beans, and corn domesticated in the fertile north and the potato, quinoa, and South American camelids domesticated in the arid south (16). In human population genetics studies, the region where the between-population homogeneity was ascertained by Tarazona-Santos et al. (9) was the arid Andes. Consequently, here we test whether the between-population homogenization of western South America and the dichotomy of arid Andes/Amazonia extend to the northward fertile Andes.

A second open issue is the evolutionary relationship between Andean and Amazonian populations, particularly with the culturally, linguistically, and environmentally different neighboring populations of the Amazon Yunga (the rain forest transitional region between the Andes and Lower Amazonia). Harris et al. (5) inferred that Andean and Amazonian populations diverged around 12,000 years ago. Archaeological findings of recent decades have rejected the traditional view of the Amazonian environment as incompatible with complex pre-Columbian societies and have revealed that the Amazonian basin has produced the earliest ceramics of South America, that endogenous agricultural complex societies developed there, and that population sizes were larger than previously thought (17). Population genetics studies (18) have reported episodes of gene flow in Amazonia which suggest that Amazonian populations were not necessarily isolated groups. Moreover, the ancestors of people living on the Peruvian coast, in the Andes, and in the Amazon Yunga had cultural and commercial interactions during the last millennia, sharing practices such as sweet potato and manioc cultivation, ceramic iconography and styles (e.g., Tutishcanyo, Kotosh, Valdivia, and Corrugate), and traditional coca chewing (19). Therefore, here we address whether gene flow accompanied the cultural and socioeconomic interactions between the ancestors of current Andean and Amazon Yunga populations.

Despite some controversy about definitions and chronology, archeologists identify a unique cultural process in western South America which includes three temporal horizons, Early, Middle, and Late, that correspond to periods of cultural dispersion involving a wide geographic area (20) (Fig. 2). In particular, the Middle and Late Horizons are associated with the expansions of the Wari (~1,000 to 1,400 years before present [YBP]) and Inca (~524 to 466 YBP) states, respectively (21–23). The between-population homogeneity currently observed in the arid Andes results from high levels of gene flow in this region, which is commonly associated with the Inca Empire (20). However, Isbell (22) has suggested that the former Wari expansion led to the spread of the Quechua language in the central Andes and that the Wari were pioneers in developing a road system in the Andes called *Wari ñam*, which was later used by the Incas to develop their network of roads (the *Qapaq ñam*) (16). A third relevant question is, therefore, when the current between-population genetic homogenization started in the context of the arid Andean chronology (Fig. 2). Particularly, is this a phenomenon restricted to the period of the Inca Empire (Late Horizon), or did it extend backward to the Middle/Wari Horizon?

Finally, Native Americans had to adapt to different and contrasting environments and stresses. The high and arid Andes are characterized by high ultraviolet radiation, cold, dryness, and hypoxia (a stress that does not allow for cultural adaptations and requires biological changes) (24, 25). The Amazon has a low incidence of light, a warm and humid climate typical of the rain forest, and high biodiversity, including pathogens (26). Here we infer episodes of genetic adaptation to the arid Andes and the Amazonian tropical forest.

Results and Discussion

We used data from Harris et al. (5) for 74 indigenous individuals and additional data from 289 unpublished individuals from 18 Peruvian Native populations, genotyped for ~2.5 million single nucleotide polymorphisms (SNPs) (Fig. 1*B* and Dataset S1). For population genetics analyses, we created three datasets with different SNP densities and populations (27–30) (*SI Appendix*, Fig. S1 and section 1.3, and Datasets S2 and S3). The institutional review boards of participants' institutions approved this research. The study was led by Peruvian institutions and investigators who have a long record of community engagement activities as an intrinsic component of their research protocols. Bioinformatics pipelines are described in (31).

The Between-Population Homogenization of Western South America and the Dichotomy of Arid Andes/Amazonia do not Extend to the Northward Fertile Andes. By applying ADMIXTURE (32) and principal component analyses (Fig. 1*B* and *SI Appendix*, Figs. S2–S7), as well as haplotype-based methods (33, 34) (*SI Appendix*, Figs. S8–S13 and sections 2.1.1 and 2.1.2), we confirmed that populations in the arid Andes are genetically homogeneous, appearing as an almost panmictic unit, with an ancestry pattern differentiated with respect to Amazonian populations (Fig. 1*B*). Conversely, populations of the northern coast (Moches and Tallanes) and in the northern Amazon Yunga (i.e., Chachapoyas) share the same ancestry profile between them (Fig. 1*B* and *SI Appendix*, Figs. S8–S13), which is different from the populations from the arid Andes. Thus, the between-population homogenization of the arid Andes and its differentiation with respect to Amazonian populations of similar latitudes do not extend northward and are not characteristic of all western South America. Instead, the genetic structure of western South Amerindian populations recapitulates the environmental and cultural differentiation between the northern fertile Andes and the southern arid Andes. Nakatsuka et al. (2) (their figure 2), studying aDNA from 86 pre-Columbian individuals, showed that some level of north–south population structure predates the arrival of Spaniards to Peru in 1532. They claim that there was a strong pre-Columbian north–south population structure in the western Andes in pre-Columbian times. However, their claim partly depends on removing from the results of their figure 2 sixteen out of the 86 studied pre-Columbian individuals whom they call "outliers" (18% of their aDNA dataset). The inclusion of these so-called outliers [see SI Appendix, figure S4 of Nakatsuka et al. (2)] shows that the north–south pre-Columbian population structure was not as strong as they claimed.

Longitudinal Gene Flow between the North Coast, Andes, and Amazonia Accompanied the Well-Documented Cultural and Socioeconomic Interactions. Haplotype-based inferences (ChromoPainter/Globetrotter methods) (33, 34) (Fig. 1*B* and *SI Appendix*, Figs. S11–S13 and section 2.1.3), statistical tests of treeness (35) (Fig. 1*B* and *SI Appendix*, Figs. S14 and S15 and section 3.2.1), and admixture graphs (35) (*SI Appendix*, Figs. S16–S19 and section 3.2.2) reveal genetic signatures of gene flow between coastal/Andean and Amazon Yunga populations in latitudes of the northern fertile Andes but not in the southern arid Andes. Thus, longitudinal gene flow between the north coast, Andes, and Amazonia accompanied cultural and socioeconomic interactions documented by archeology, which include ceramic styles and crops, as well as the critical role that Chachapoyas may have played (see Introduction and *SI Appendix*, section 3.1). This pattern of gene flow recapitulates the

Downloaded from https://www.pnas.org by UNIVERSIDADE FEDERAL DE MINAS GERAIS on June 12, 2024 from IP address 150.164.28.250.

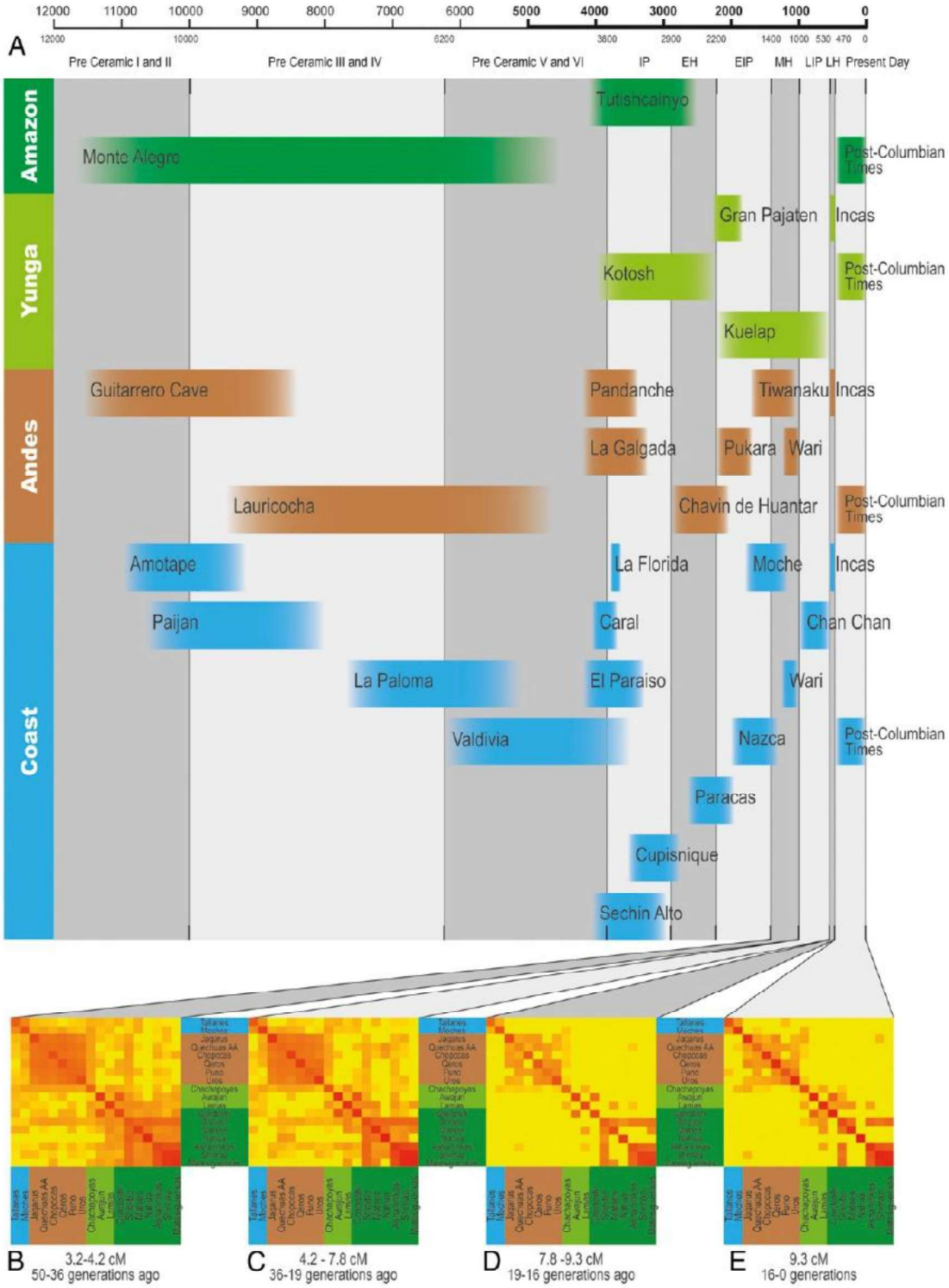


Fig. 2. Changes in IBD sharing over time between the Pacific coast, central Andes, Amazon Yunga, and Amazon and its relationship with the archaeological chronology of the Andes. (*A*) Key historical events (cultures and archeological sites) of Peruvian history in four Peruvian longitudinal regions, coast, Andes, Amazon Yunga, and Amazonia. This is a simplified chronology of Peruvian archaeological history based on different dating records. To account for temporal uncertainties, we depicted the events in the chronology plot without clearly defined chronological borders. The timeline on the top and bottom is represented in years before present. IP: initial period, EH: Early Horizon, EIP: early intermediate period, MH: Middle Horizon, LIP: late intermediate period, LH: Late Horizon. Adapted from ref. 4, which is licensed under CC BY 4.0. (*B–E*) Heat maps of the average pairwise relatedness (85) among Native Americans of the Natives 1.9M dataset. Each heat map represents an interval of IBD segment lengths, which correspond to interval times (36).

Downloaded from https://www.pnas.org by UNIVERSIDADE FEDERAL DE MINAS GERAIS on June 12, 2024 from IP address 150.164.28.250.

differentiation between the fertile north, where altitudes are lower, and the arid south, where the Andes altitudes are higher (Fig. 1*A*) and may have acted as a barrier to gene flow, imposing a sharper environmental differentiation between the Andes and the Amazon Yunga. Formal comparison of admixture graphs (35) (*SI Appendix*, Figs. S16–S19) representing different scenarios shows that gene flow was more intense from the north coast to the Amazon than in the opposite direction and that in latitudes of the fertile north, gene flow included important ethnic groups such as the current Chachapoyas of the Amazon Yunga, as well as eastward Lower Amazonian populations such as those of the Jivaro linguistic family (Awajun and Candoshi) and Lamas (Fig. 1*B* and *SI Appendix*, Figs. S16–S19). These results are consistent with those of Nakatsuka et al. (2) based on current and pre-Hispanic individuals.

The Homogenization of the Central Arid Andes Started at least during the Wari Expansion (1,400 to 1,000 YBP). We analyzed the distribution of identity-by-descent (IBD) segment lengths between individuals of different arid Andean populations, which is informative about the dynamics of past gene flow (5, 36). We observed a signature of gene flow in the interval between 1,400 and 1,000 YBP, within the Wari expansion in the Middle Horizon (Fig. 2). Thus, the homogenization of the central arid Andes is not only due to migrations during the Inca Empire or later during the Spanish Viceroyalty of Peru, when migrations (often forced) occurred (37). The Wari expansion (1,400 to 1,000 YBP) was also accompanied by intensive gene flow whose signature is still present in the between-population genetic homogeneity of the arid central Andes region. We also observed that during the Wari/Middle Horizon the effective population size (N_e) was rising in the arid Andes (*SI Appendix*, Fig. S22), a trend that stopped with the European contact, when N_e started to decline, consistent with demographic records (38) and with genetic studies by Lindo et al. (39). Because IBD analysis on current individuals does not allow for inferences of gene flow that occurred more than 75 generations ago (36), ancient DNA analysis at the population level will be necessary to infer whether the between-population homogenization of the Andes started even earlier.

Episodes of Genetic Adaptation Occurred in the Arid Andes and the Amazonian Tropical Forest. Populations from the high and arid Andes and those from the Amazon (Fig. 1*B*) settled in these contrasting environments more than 5,000 years ago (40) and show little evidence of gene flow between them (i.e., that would homogenize allele frequencies, potentially concealing the effect of diversifying natural selection). We performed genome-wide scans in these two groups of populations using two tests of positive natural selection: 1) population branch statistics (PBSn) comparing arid Andeans (Chopccas, Quechuas_AA, Qeros, Puno, Jaqarus, and Uros; $n = 102$) vs. Amazonian populations (Ashaninkas, Matsiguenkas, Matses, and Nahua; $n = 75$) with a Chinese population (Dai in Xishuangbanna, China; $n = 100$) from 1000 Genomes as an out-group (41) (*SI Appendix*, section 5.2.1) and 2) long-range haplotypes (xpEHH) (42) estimated for the two groups of populations (Fig. 3 and *SI Appendix*, Figs. S24–S27 and section 5.2.2). The complete lists of SNPs with high PBSn and xpEHH statistics for Andean and Amazonian populations are in Datasets S4–S7.

The gene with the consensually strongest signal of adaptation (both from PBSn and xpEHH statistics: PBSn = 0.205, *P* value = 0.003; xpEHH = 4.481, *P* value < 0.00001) to the Andean environment (Fig. 3 and Dataset S4) is a long noncoding RNA gene called *HAND2-AS1* (heart and neural crest derivatives expressed 2 RNA antisense 1, chromosome 4), that modulates cardiogenesis by regulating the expression of the nearby *HAND2* gene (43, 44). This result is consistent with 1) the natural selection genome-wide scan by Crawford et al. (41), who identified three genes related to the cardiovascular system in Andeans, including *TBX5*, which works together with *HAND2* in reprogramming fibroblasts to cardiac-like myocytes (45, 46), and 2) a pattern of adaptation of Andean populations preferentially mediated by the cardiovascular system. The derived allele rs2877766-A (frequencies: Amazonians, 0.453; Andeans, 0.880) is the core of the extended haplotype. *HAND2-AS1* is located in the antisense 5′ region of *HAND2*, and the positively selected six SNPs core haplotype is ∼18-kilobase and encompasses a putative human enhancer (GeneHancer identifier GH04J173536, *SI Appendix*, Fig. S29). Considering the limitation of our data that come from genotyping arrays, we further recovered from the sequencing data by Harris et al. (5) all nearby SNPs in linkage disequilibrium in Andean populations ($r^2 > 0.80$) with the core SNP rs2877766. We found that the positively selected haplotype includes the SNP rs3775587, mapped within the putative enhancer GH04J173536. Altogether, these results suggest (but do not demonstrate) that the *HAND2-AS1* signature of natural selection is related to regulation of gene expression by an enhancer and reflects cardiovascular adaptations. Andeans have cardiovascular adaptations to high altitude that differ from those of lowlanders exposed to hypoxia and from those of other highlanders, showing higher pulmonary vasoconstrictor response to hypoxia, lower resting middle cerebral flow velocity than Tibetans, and higher uterine artery blood flow than Europeans and lowlanders raised in high altitude (47).

DUOX2 (dual oxidase 2, chromosome 15) is the gene with the highest signal of adaptation to the Andean environment by PBSn analysis (PBSn = 0.22, *P* value = 0.002) (Fig. 3 and *SI Appendix*, Fig. S24). It has already been reported as a natural selection target in the Andes (48, 49). *DUOX2* encodes a transmembrane component of an NADPH oxidase, which produces hydrogen peroxide (H_2O_2), and is essential for the synthesis of the thyroid hormone and for the production of the microbicidal hypothiocyanite anion ($OSCN^-$) during mucosal innate immunity response against bacterial and viral infections in the airways and intestines (50, 51). Mutations in *DUOX2* produce inherited hypothyroidism (52). Here we report the following: 1) The PBSn signal for *DUOX2* comprises several SNPs, including two missense mutations (rs269868: C > T: Ser1067Leu, C allele frequencies: Amazon, 0.01, Andean, 0.53; rs57659670: T > C: His678Arg, C allele frequencies: Amazon, 0.01, Andean, 0.53); 2) bioinformatics analysis reveals that rs269868 is located in an A-loop, 1064-1078 amino acids, which is a region of interaction of *DUOX2* with its coactivator *DUOXA2*. Mutations in this region of the protein can affect the stability and maturation of the dimer and, consequently, the conversion of the intermediate product O_2 to the final product H_2O_2 and their released proportions (53). If the natural selection signal is related to this effect, then the standing ancestral allele has been positively selected in the Andes. It is not clear whether the *DUOX2* natural selection signal is related to thyroid function or innate immunity. Before the introduction of the public health program of supplementing manufactured salt with iodine, one of the environmental stresses of the Andes for human populations was iodine deficiency, which impairs thyroid hormone synthesis, increasing the risk of developing hypothyroidism, goiter, obstetric complications, and cognitive impairment (54, 55).

Natural selection studies in Amazon populations are scarce. Studies targeting rain forest populations in Africa and Asia have found natural selection signals in genes related to height and immune response (56). In the Amazon region, the strongest natural selection PBSn signal (PBSn = 0.302, *P* value = 0.002) is in a long noncoding RNA gene on chromosome 18 with unknown function (Dataset S5 and *SI Appendix*, Fig. S25). The second-highest signal (which also shows a significant long-range haplotype signal: PBSn = 0.265, *P* value = 0.004; xpEHH = −4.222, *P* value = 0.0003) corresponds to the gene *PTPRC* (Fig. 3), which encodes the protein CD45, essential in antigen recognition by T and B lymphocytes in pathogen–host interaction, in particular for

Downloaded from https://www.pnas.org by UNIVERSIDADE FEDERAL DE MINAS GERAIS on June 12, 2024 from IP address 150.164.28.250.

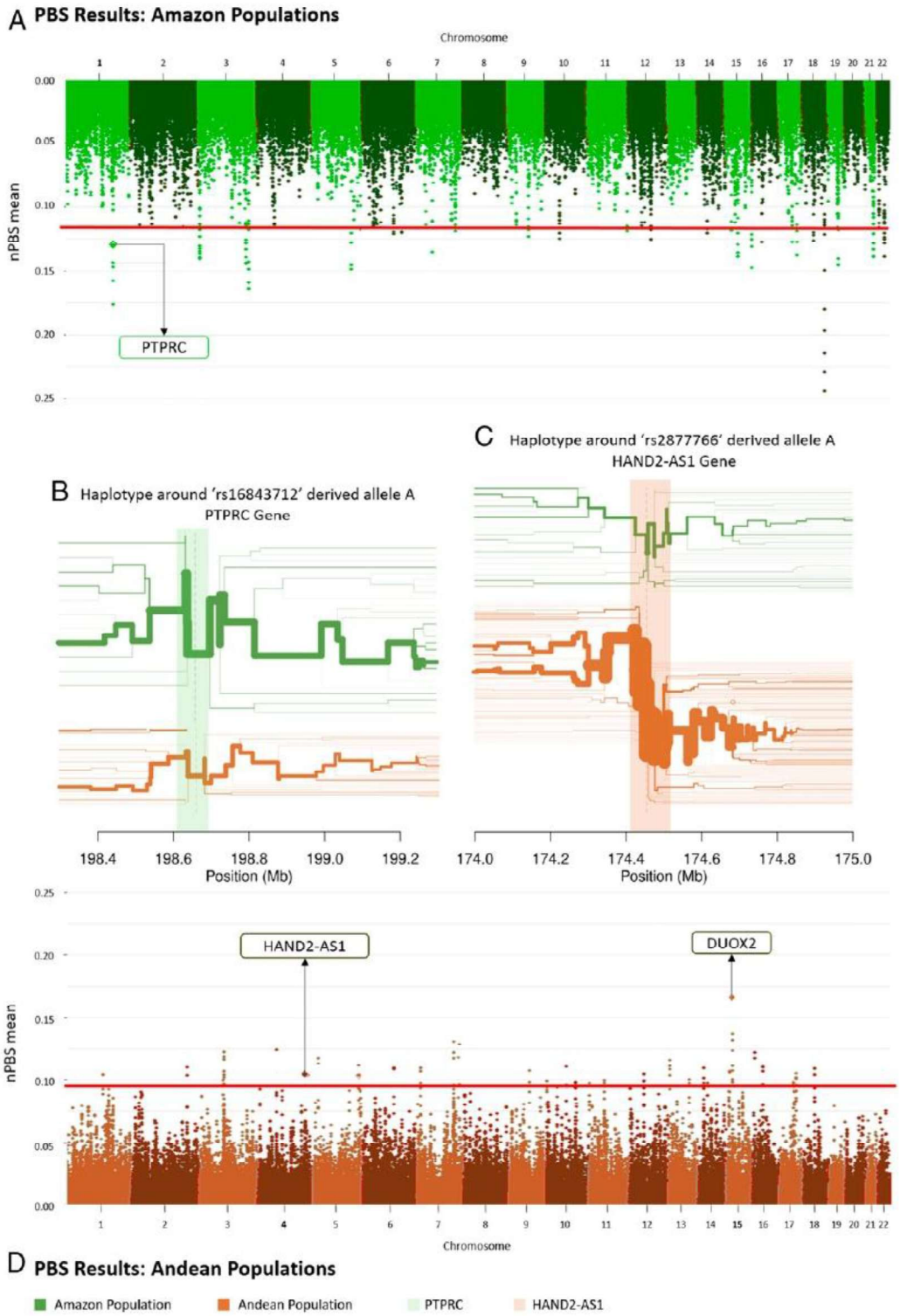


Fig. 3. Sandwich-like representation of natural selection signatures in the arid Andes and the Amazon tropical forest. Manhattan plots (the bread of the sandwich) correspond to the PBSn estimated from the sliding window in Amazon (*A*) and Andean (*D*) populations. The horizontal red line shows the 99.95th percentile of PBSn values. The filling of the sandwich is the long-range haplotype representations. (*B*) Long-range haplotypes flanking the rs16843712 derived allele A (frequencies: Amazon, 0.811; Andean, 0.324) in the PTPRC gene (light green vertical shading). (*C*) Haplotype flanking the rs2877766 derived allele A (frequencies: Amazon, 0.453; Andean, 0.880) in the HAND2-AS1 gene (light brown vertical shading). Green plots refer to the Amazon populations; brown plots refer to the Andean populations.

viruses such as human adenovirus type 19 (57), HIV-1-induced cell apoptosis (58, 59), hepatitis C (60, 61), and herpes simplex virus 1 (62), even if we cannot exclude a role for unknown viruses endemic in the Amazon region. The core haplotype flanks the rs16843712 derived allele A (frequencies: Amazonia, 0.811; Andes, 0.324), within the putative human enhancer GH01J198660

Downloaded from https://www.pnas.org by UNIVERSIDADE FEDERAL DE MINAS GERAIS on June 12, 2024 from IP address 150.164.28.250.

(sensu GeneHancer; *SI Appendix*, Fig. S30), and includes the A (Thr193) allele of the nonsynonymous SNP rs4915154 (A > G: Thr193Ala) in exon 6 that affects alternative splicing and alters a potential O- and N-linked glycosylation site. The positively selected allele A (Thr193) has been associated (63) with a lower proportion of CD45R0+ T memory cells and an increased amount of naive phenotype T cells expressing A (exon 4), B (exon 5), and C (exon 6) isoforms. This result is consistent with the hypothesis of CD45 evolution driven by a host–virus arms race model (64).

In addition to the natural selection PBSn and xpEHH signals, we used the bioinformatics platform MASSA (Multi-Agent System for SNP Annotations) (65) to annotate the 1,985 (0.1%) most differentiated SNPs (F_{CT} > 0.318) between the same Andean and Amazonian groups that we tested for natural selection. Notably, we found three *TMPRSS6* (transmembrane serine protease 6) variants, rs855791-T (2246T > C Val727Ala: Andean = 0.60, Amazon = 0.92), rs4820268-G (Andean = 0.59 Amazon = 0,98), and rs2413450-T (Andean = 0.60, Amazon = 0.98; Dataset S8), more common in the Amazon region and associated with a broad spectrum of hematological phenotypes such as lower hemoglobin, iron, ferritin, and glycated hemoglobin and higher hepcidin/ferritin ratio (a hormone that decreases iron absorption and distribution) levels in blood, as well as mean corpuscular volume (sensu Genome-Wide Association Study [GWAS] Catalog, that includes GWASs with Latin American admixed individuals) (66–68).

We use DANCE [Disease Ancestry Network (69)] to present the allele frequencies of our total Native American samples for 30,270 GWAS hits and its associated complex phenotypes (*sensu* GWAS Catalog, https://www.ebi.ac.uk/gwas/), in comparison with African, European, and Asian allele frequencies from the 1000 Genome Project. While this information is relevant, we recall that the allelic architecture of the complex diseases presented in the GWAS Catalog is biased by the underrepresentation of individuals with non-European ancestry in genetic studies.

In conclusion, in western South America, there is an environmental and cultural differentiation between the fertile north of the Andes, where altitudes are lower, and the arid south of the Andes, where these mountains are higher, defining sharp environmental differences between the Andes and Amazonia. This has influenced the genetic structure of western South Amerindian populations. Indeed, the between-population homogenization of the central southern Andes and its differentiation with respect to Amazonian populations of similar latitudes do not extend northward. Gene flow between the northern coast of Peru, the Andes, and Amazonia accompanied cultural and socioeconomic interactions revealed by archeology, but in the central southern Andes, these mountains have acted as a genetic barrier to gene flow (70). We provide insights on the dynamics of the genetic homogenization between the populations of the arid Andes which is not only due to migrations during the Inca Empire or the subsequent colonial period but started at least during the earlier expansion of the pre-Inca Wari Empire (600 to 1,000 YBP). Nakatsuka et al. (2), comparing ancient with modern individuals from western South America, make the general claim that the genetic structure of current populations "strongly echoed" and "are most closely related to the ancient individuals from their region" (i.e., 500 to 2,000 years ago). However, this general statement is not supported by their own results (see their SI Appendix, figure S7). From nine ancient (500 to 2,000 years ago) vs. current comparisons of populations from the same region, this statement is true only for the five cases of the Southern Highlands of Peru and for Chile (their SI Appendix, figure S7 J and K) and not for the four comparisons from the Peruvian coast and north of Peru (their SI Appendix, figure S7 F–I). Thus, Nakatsuka et al.'s (2) results emphasize and add a temporal perspective to the dichotomy observed by us between the current northern fertile Andes (more associated with trans-Andean gene flow) and the southern arid Andes (more homogeneous between populations and differentiated from the Amazonia). The evolutionary journey of western South Amerindians was accompanied by episodes of adaptive natural selection to the high and arid Andes vs. the low Amazon tropical forest: the noncoding gene *HAND2-AS1* (related to cardiovascular function and with the positively selected haplotype encompassing a putative human enhancer) and *DUOX2* (related to thyroid function and innate immunity) in the Andes. In the Amazon forest, the gene encoding for the protein CD45, essential for antigen recognition by T and B lymphocytes and viral–host interactions, shows a signature of positive natural selection, consistent with the host–virus arms race hypothesis. Our results and other studies (70) continue to show how Andean highlanders and Amazonian dwellers provide examples of how the interplay between geography and culture influences the genetic structure and adaptation of human populations.

Materials and Methods

The protocol for the Peruvian Genome Diversity Project was approved by the Research and Ethics Committee (OI003-11 and OI-087-13) of the Peruvian National Institute of Health, and all participants who had samples collected in this project provided informed consent. We genotyped 289 present-day Native Americans from Peru using the Human Omni array of Illumina for 2.5 million SNPs as part of the Peruvian Genome Diversity Project. Quality control was performed using PLINK (71) and Laboratório de Diversidade Genética Humana bioinformatics protocols and scripts (31). We merged our individuals with public datasets (1, 28–30) and Kaqchikel individuals from M.D. lab from National Cancer Institute. For *D* statistics and admixture graph analyses, we generate masked data, after phasing our datasets with SHAPEIT2 (72) and inferring the non-Native DNA segments with RFMix (73). To infer population structure, we used two approaches: 1) principal component analysis in Eigenstrat (74) and genetic clustering on ADMIXTURE software (32) using a linkage disequilibrium pruned dataset and 2) fineSTRUCTURE (33), MIXTURE MODEL (34, 75), and SOURCEFIND (76) for haplotype-based analyses, after phase inference. Historical relationships were inferred using *D* statistics (77) and Admixture Graphs (35). IBD was inferred using refinedIBD (78) and IBDNe (79). For the genetic differentiation analyses, the pairwise genetic distances (F statistics) between Native South American groups (F_{ST}) and between populations within groups (F_{SC}) were calculated for multilocus and individual loci using 4P software (80) and the hierfstat R package (81), respectively. The linkage disequilibrium was inferred by the software Haploview (82). Natural selection scans were performed using population branch statistics (41, 83) and xpEHH from the package Selscan (42, 84).

Data Availability. Data have been deposited in the European Genome-phenome Archive (EGA), https://www.ebi.ac.uk/ega/home (accession nos. EGAD00010001958, EGAD00010001990, EGAD00010001991, EGAD00010001992).

ACKNOWLEDGMENTS. We thank the Peruvian populations for their participation. We thank the members of the Laboratório de Diversidad Genética Humana, Mateus Gouveia, Kelly Nunes, Garrett Hellenthal, Mark Lipson, Marcia Beltrame, Fabrício Santos, Claudio Struchiner, Ricardo Santos, Luis Guillermo Lumbreras, Sandra Romero-Hidalgo, Víctor Acuña-Alonzo, Miguel Ortega, and Juliana Lacerda, for discussions or technical assistance; Harrison Montejo, Silvia Capristano, Juana Choque, and Marco Galarza from Laboratorio de Biotecnologia y Biologia Molecular of Instituto Nacional de Salud (Peru) for collaborating with the Peruvian Genome Project and conducting the genotyping; and Rafael Tou, Lucas Faria, Livia Metzker, and Alex Teixeira for their final reading of *SI Appendix*. This work was supported by the Peruvian National Institute of Health (INS), the Brazilian Conselho Nacional de Desenvolvimento Científico e Tecnológico, Pró-Reitoria de pesquisa at the Universidade Federal de Minas Gerais (UFMG), Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG, grant number RED00314-16), and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) programs: the Programa de Excelência Acadêmica (PROEX) and the Programa Institucional de Internacionalização (PRINT). V.B. was a CAPES/Programa de Estudantes-Convênio de Pós-Graduação (PEC-PG) fellow (grant number 88882.195664/2018-01). P.E.R. was funded by the Fondo Nacional de Desarrollo Científico, Tecnológico y de Innovación Tecnológica (Fondecyt - Perú) (grant number 34-2019, "Proyecto de Mejoramiento y Ampliación de los Servicios del Sistema Nacional de Ciencia, Tecnología e Innovación Tecnológica"). Datasets were

Downloaded from https://www.pnas.org by UNIVERSIDADE FEDERAL DE MINAS GERAIS on June 12, 2024 from IP address 150.164.28.250.

processed in the Sagarana HPC cluster at the Centro de Laboratórios Multiusuários at Instituto de Ciências Biológicas-UFMG. This work is a product of the collaboration between investigators from the Peruvian Genome Project at the INS and the Genomics and Bioinformatics group of the Project Proproject Epidemiologia Genômica de Coortes Brasileiras de base populagional (EPIGEN-Brazil, https://epigen.grude.ufmg.br/), funded by the Departamento de Ciência e Tecnologia/Ministério de Saúde (DECIT-MS, Brazil).

1. C. Posth *et al.*, Reconstructing the deep population history of central and South America. *Cell* **175**, 1185–1197.e22 (2018).
2. N. Nakatsuka *et al.*, A paleogenomic reconstruction of the deep population history of the Andes. *Cell* **181**, 1131–1145.e21 (2020).
3. R. S. Solis, J. Haas, W. Creamer, Dating Caral, a preceramic site in the Supe Valley on the central coast of Peru. *Science* **292**, 723–726 (2001).
4. M. O. Scliar *et al.*, Bayesian inferences suggest that Amazon Yunga natives diverged from Andeans less than 5000 ybp: Implications for South American prehistory. *BMC Evol. Biol.* **14**, 174 (2014).
5. D. N. Harris *et al.*, Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E6526–E6535 (2018).
6. C. Lahaye *et al.*, New insights into a late-Pleistocene human occupation in America: The Vale da Pedra Furada complete chronological study. *Quat. Geochronol.* **30**, 445–451 (2015).
7. T. D. Dillehay *et al.*, Monte Verde: Seaweed, food, medicine, and the peopling of South America. *Science* **320**, 784–786 (2008).
8. T. D. Dillehay *et al.*, New archaeological evidence for an early human presence at Monte Verde, Chile. *PLoS One* **10**, e0141923 (2015).
9. E. Tarazona-Santos *et al.*, Genetic differentiation in South Amerindians is related to environmental and cultural diversity: Evidence from the Y chromosome. *Am. J. Hum. Genet.* **68**, 1485–1496 (2001).
10. L. Campbell, *American Indian Languages: The Historical Linguistics of Native America* (Oxford University Press, 2000).
11. S. Fuselli *et al.*, Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. *Mol. Biol. Evol.* **20**, 1682–1691 (2003).
12. C. M. Lewis Jr., J. C. Long, Native South American genetic structure and prehistory inferred from hierarchical modeling of mtDNA. *Mol. Biol. Evol.* **25**, 478–486 (2008).
13. S. Wang *et al.*, Genetic variation and population structure in native Americans. *PLoS Genet.* **3**, e185 (2007).
14. J. R. Sandoval *et al.*; Genographic Project Consortium, The genetic history of indigenous populations of the Peruvian and Bolivian Altiplano: The legacy of the Uros. *PLoS One* **8**, e73006 (2013).
15. G. A. Gnecchi-Ruscone *et al.*, Dissecting the pre-Columbian genomic ancestry of Native Americans along the Andes-Amazonia divide. *Mol. Biol. Evol.* **36**, 1254–1269 (2019).
16. L. G. Lumbreras, *Los orígenes de la civilización en el Perú*; (Instituto Andino de Estudios Arqueológico-Sociales, 2015).
17. A. Roosevelt, "The maritime, highland, forest dynamic and the origins of complex culture" in *The Cambridge History of the Native Peoples of the Americas*, F. Salomon, S. B. Schwartz, Eds. (Cambridge University Press, 1999), pp. 264–349.
18. C. Barbieri *et al.*, The current genomic landscape of western South America: Andes, Amazonia and Pacific coast. *Mol. Biol. Evol.* **36**, 2698–2713 (2019).
19. H. Silverman, W. Isbell, Eds., *The Handbook of South American Archaeology* (Springer, New York, 2008).
20. J. Haas, S. Pozorski, T. Pozorski, *The Origins and Development of the Andean State* (Cambridge University Press, 1987).
21. E. P. Lanning, *Peru before the Incas* (Prentice-Hall, 1967).
22. W. H. Isbell, "Wari and Tiwanaku: International identities in the central Andean Middle Horizon" in *The Handbook of South American Archaeology*, H. Silverman, W. H. Isbell, Eds. (Springer, New York, 2008), pp. 731–759.
23. G. Valverde *et al.*, Ancient DNA analysis suggests negligible impact of the Wari Empire expansion in Peru's central coast during the Middle Horizon. *PLoS One* **11**, e0155508 (2016).
24. E. Tarazona-Santos, M. Lavine, S. Pastor, G. Fiori, D. Pettener, Hematological and pulmonary responses to high altitude in Quechuas: A multivariate approach. *Am. J. Phys. Anthropol.* **111**, 165–176 (2000).
25. L. G. Moore, Measuring high-altitude adaptation. *J. Appl. Physiol. (1985)* **123**, 1371–1385 (2017).
26. C. E. G. Amorim, J. T. Daub, F. M. Salzano, M. Foll, L. Excoffier, Detection of convergent genome-wide signals of adaptation to tropical forests in humans. *PLoS One* **10**, e0121557 (2015).
27. G. R. Abecasis *et al.*; 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
28. D. Reich *et al.*, Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
29. M. Raghavan *et al.*, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
30. S. Mallick *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
31. W. C. S. Magalhães *et al.*; Brazilian EPIGEN Consortium, EPIGEN-Brazil initiative resources: A Latin American imputation panel and the scientific workflow. *Genome Res.* **28**, 1090–1095 (2018).
32. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
33. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
34. G. Hellenthal *et al.*, A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
35. N. Patterson *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
36. P. F. Palamara, T. Lencz, A. Darvasi, I. Pe'er, Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.* **91**, 809–822 (2012).
37. N. D. Cook, "Migration in colonial Peru: An overview" in *Migration in Colonial Spanish America*, D. J. Robinson, Ed. (Cambridge University Press, 1990), pp. 41–61.
38. N. Sanchez-Albornoz, *The Population of Latin America: A History* (University of California Press, Berkeley, 1974).
39. J. Lindo *et al.*, The genetic prehistory of the Andean highlands 7000 years BP though European contact. *Sci. Adv.* **4**, eaau4921 (2018).
40. L. Eriksen, *Nature and Culture in Prehistoric Amazonia: Using G.I.S. to Reconstruct Ancient Ethnogenetic Processes from Archeology, Linguistics, Geography, and Ethnohistory* (Department of Human Geography, Human Ecology Division, Lund University, 2011).
41. J. E. Crawford *et al.*, Natural selection on genes related to cardiovascular health in high-altitude adapted Andeans. *Am. J. Hum. Genet.* **101**, 752–767 (2017).
42. P. C. Sabeti *et al.*; International HapMap Consortium, Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
43. K. M. Anderson *et al.*, Transcription of the non-coding RNA upperhand controls Hand2 expression and heart development. *Nature* **539**, 433–436 (2016).
44. X. Cheng, H. Jiang, Long non-coding RNA HAND2-AS1 downregulation predicts poor survival of patients with end-stage dilated cardiomyopathy. *J. Int. Med. Res.* **47**, 3690–3698 (2019).
45. H. Hashimoto *et al.*, Cardiac reprogramming factors synergistically activate genome-wide cardiogenic stage-specific enhancers. *Cell Stem Cell* **25**, 69–86.e5 (2019).
46. A. Fernandez-Perez *et al.*, Hand2 selectively reorganizes chromatin accessibility to induce pacemaker-like transcriptional reprogramming. *Cell Rep.* **27**, 2354–2369.e7 (2019).
47. C. G. Julian, L. G. Moore, Human genetic adaptation to high altitude: Evidence from the Andes. *Genes (Basel)* **10**, 150 (2019).
48. D. Zhou *et al.*, Whole-genome sequencing uncovers the genetic basis of chronic mountain sickness in Andean highlanders. *Am. J. Hum. Genet.* **93**, 452–462 (2013).
49. V. C. Jacovas *et al.*, Selection scan reveals three new loci related to high altitude adaptation in Native Andeans. *Sci. Rep.* **8**, 12733 (2018).
50. A. van der Vliet, K. Danyal, D. E. Heppner, Dual oxidase: A novel therapeutic target in allergic disease. *Br. J. Pharmacol.* **175**, 1401–1418 (2018).
51. X. De Deken, B. Corvilain, J. E. Dumont, F. Miot, Roles of DUOX-mediated hydrogen peroxide in metabolism, host defense, and signaling. *Antioxid. Redox Signal.* **20**, 2776–2793 (2014).
52. Y. Maruo *et al.*, Natural course of congenital hypothyroidism by dual oxidase 2 mutations from the neonatal period through puberty. *Eur. J. Endocrinol.* **174**, 453–463 (2016).
53. T. Ueyama *et al.*, The extracellular A-loop of dual oxidases affects the specificity of reactive oxygen species release. *J. Biol. Chem.* **290**, 6495–6506 (2015).
54. E. A. Pretell *et al.*, Elimination of iodine deficiency disorders from the Americas: A public health triumph. *Lancet Diabetes Endocrinol.* **5**, 412–414 (2017).
55. L. Pan, Z. Fu, P. Yin, D. Chen, Pre-existing medical disorders as risk factors for pre-eclampsia: An exploratory case-control study. *Hypertens. Pregnancy* **38**, 245–251 (2019).
56. S. Fan, M. E. B. Hansen, Y. Lo, S. A. Tishkoff, Going global by adapting local: A review of recent human adaptation. *Science* **354**, 54–59 (2016).
57. M. Windheim *et al.*, A unique secreted adenovirus E3 protein binds to the leukocyte common antigen CD45 and modulates leukocyte functions. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E4884–E4893 (2013).
58. A. R. Anand, R. K. Ganju, HIV-1 gp120-mediated apoptosis of T cells is regulated by the membrane tyrosine phosphatase CD45. *J. Biol. Chem.* **281**, 12289–12299 (2006).
59. S. Meer, Y. Perner, E. D. McAlpine, P. Willem, Extraoral plasmablastic lymphomas in a high human immunodeficiency virus endemic area. *Histopathology* **76**, 212–221 (2020).
60. R. Dawes *et al.*, Altered CD45 expression in C77G carriers influences immune function and outcome of hepatitis C infection. *J. Med. Genet.* **43**, 678–684 (2006).
61. J.-L. Hsiao, W.-S. Ko, C.-J. Shih, Y.-L. Chiou, The changed proportion of $CD45RA^+$/$CD45RO^+$ T cells in chronic hepatitis C patients during pegylated Interferon-α with ribavirin therapy. *J. Interferon Cytokine Res.* **37**, 303–309 (2017).
62. G. Caignard *et al.*, Genome-wide mouse mutagenesis reveals CD45-mediated T cell function as critical in protective immunity to HSV-1. *PLoS Pathog.* **9**, e1003637 (2013).
63. T. Stanton *et al.*, A high-frequency polymorphism in exon 6 of the CD45 tyrosine phosphatase gene (PTPRC) resulting in altered isoform expression. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5997–6002 (2003).
64. N. Thiel, J. Zischke, E. Elbasani, P. Kay-Fedorov, M. Messerle, Viral interference with functions of the cellular receptor tyrosine phosphatase CD45. *Viruses* **7**, 1540–1557 (2015).
65. G. Soares-Souza, *Novas Abordagens para Integração de Bancos de Dados e Desenvolvimento de Ferramentas Bioinformáticas para Estudos de Genética de Populações* (PhD Thesis, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 2014).

Downloaded from https://www.pnas.org by UNIVERSIDADE FEDERAL DE MINAS GERAIS on June 12, 2024 from IP address 150.164.28.250.

66. C. J. Hodonsky *et al.*, Genome-wide association study of red blood cell traits in Hispanics/Latinos: The Hispanic community health study/study of Latinos. *PLoS Genet.* **13**, e1006760 (2017).
67. M. H. Kowalski *et al.*; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; TOPMed Hematology & Hemostasis Working Group, Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* **15**, e1008500 (2019).
68. L. M. Raffield *et al.*, Genome-wide association study of iron traits and relation to diabetes in the Hispanic community health study/study of Latinos (HCHS/SOL): Potential genomic intersection of iron and glucose regulation? *Hum. Mol. Genet.* **26**, 1966–1978 (2017).
69. G. S. Araújo *et al.*, Integrating, summarizing and visualizing GWAS-hits and human diversity with DANCE (Disease-ANCEstry networks). *Bioinformatics* **32**, 1247–1249 (2016).
70. M. Mendes, I. Alvim, V. Borda, E. Tarazona-Santos, The history behind the mosaic of the Americas. *Curr. Opin. Genet. Dev.* **62**, 72–77 (2020).
71. S. Purcell *et al.*, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
72. O. Delaneau, J. Marchini, J.-F. Zagury, A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
73. B. K. Maples, S. Gravel, E. E. Kenny, C. D. Bustamante, RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
74. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
75. S. Leslie *et al.*; Wellcome Trust Case Control Consortium 2; International Multiple Sclerosis Genetics Consortium, The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
76. J.-C. Chacón-Duque *et al.*, Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat. Commun.* **9**, 5388 (2018).
77. R. E. Green *et al.*, A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
78. B. L. Browning, S. R. Browning, Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
79. S. R. Browning, B. L. Browning, Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J. Hum. Genet.* **97**, 404–418 (2015).
80. A. Benazzo, A. Panziera, G. Bertorelle, 4P: Fast computing of population genetics statistics from large DNA polymorphism panels. *Ecol. Evol.* **5**, 172–175 (2015).
81. J. Goudet, Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Resour.* **5**, 184–186 (2005).
82. J. C. Barrett, B. Fry, J. Maller, M. J. Daly, Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
83. X. Yi *et al.*, Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
84. Z. A. Szpiech, R. D. Hernandez, selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* **31**, 2824–2827 (2014).
85. S. Baharian *et al.*, The Great Migration and African-American Genomic Diversity. *PLoS Genetics* **12**, e1006059 (2016).

Downloaded from https://www.pnas.org by UNIVERSIDADE FEDERAL DE MINAS GERAIS on June 12, 2024 from IP address 150.164.28.250.

2.3. The need to diversify genomic studies: Insights from Andean highlanders and Amazonians

2.3.1. Contribuições

No comentário "*Genomics studies in underrepresented people: clinically actionable genotypes in Andeans and Amazonians*", publicado na revista Cell, pude trabalhar nas análises e interpretação dos resultados. Além disso, como uma das primeiras autoras, participei da discussão da formatação da figura e apresentação dos resultados, processo de submissão e resposta aos revisores.

Artigo

Link para o artigo online: The need to diversify genomic studies: Insights from Andean highlanders and Amazonians

Os materiais suplementares para este artigo se encontram no Anexo 2.2 da presente tese.

Please cite this article in press as: Alvim et al., The need to diversify genomic studies: Insights from Andean highlanders and Amazonians, Cell (2024), https://doi.org/10.1016/j.cell.2024.07.009

Cell
Leading Edge

CellPress

Commentary

The need to diversify genomic studies: Insights from Andean highlanders and Amazonians

Isabela Alvim,[1,2,3,4,18] Carolina Silva-Carvalho,[1,18] Marla Mendes de Aquino,[1,5] Victor Borda,[1,6,7] Cesar Sanchez,[8] Carlos Padilla,[8] Omar Cáceres,[8] Isabela Rezende-Diniz,[1] Julia Saraiva-Duarte,[1] Lucas Faria-Costa,[1] Meddly L. Santolalla,[1,9] Fernanda Rodrigues-Soares,[10] Camila Zolini,[1] Adrián Llerena,[11] Timothy D. O'Connor,[6,12,13,14] Robert H. Gilman,[15] Heinner Guio,[8,16,17] and Eduardo Tarazona-Santos[1,9,*]

[1]Department of Genetics, Ecology and Evolution, Biological Sciences Institute, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[2]School of Natural Sciences, Massey University, Palmerston North, New Zealand
[3]School of BioSciences, University of Melbourne, Melbourne, VIC, Australia
[4]St. Vincent's Institute of Medical Research, Melbourne, VIC, Australia
[5]The Centre for Applied Genomics and Genetics and Genome Biology Program, The Hospital for Sick Children, Toronto, ON, Canada
[6]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA
[7]The University of Maryland-Institute for Health Computing, University of Maryland School of Medicine, North Bethesda, MD 20852, USA
[8]Instituto Nacional de Salud, Lima, Peru
[9]Emerge, Emerging Diseases and Climate Change Research Unit, School of Public Health and Administration, Universidad Peruana Cayetano Heredia, Lima, Peru
[10]Department of Pathology, Genetic and Evolution, Biological and Natural Sciences Institute, Universidade Federal do Triângulo Mineiro, Uberaba, Brazil
[11]INUBE Extremadura Biosanitary Research Institute, Faculty of Medicine, University of Extremadura, Badajoz, Spain
[12]Program in Health Equity and Population Health, University of Maryland School of Medicine, Baltimore, MD, USA
[13]Program in Personalized Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA
[14]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA
[15]Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
[16]INBIOMEDIC Research and Technological Center, Lima, Peru
[17]Universidad de Huanuco, Huanuco, Peru
[18]These authors contributed equally
*Correspondence: edutars@gmail.com
https://doi.org/10.1016/j.cell.2024.07.009

More globally diverse perspectives are needed in genomic studies and precision medicine practices on non-Europeans. Here, we illustrate this by discussing the distribution of clinically actionable genetic variants involved in drug response in Andean highlanders and Amazonians, considering their environment, history, genetic structure, and historical biases in the perception of biological diversity of Native Americans.

The knowledge gap in medical genomics

Global precision medicine is permeated by an injustice: we know far more about the genetic basis of diseases in European-ancestry individuals and populations than in other people of the world. For this reason, genomic scientists are compellingly calling for additional genomic studies on non-European/non-US-white populations,[1] particularly on vulnerable groups such as Native Americans (i.e., individuals and groups on the American continent that predominantly descend from the autochthonous groups that were present on the continent before the arrival of Europeans and Africans since 1492). Genomic studies of underrepresented populations could facilitate the application of precision medicine in these populations, reducing the knowledge gap in medical genomics and its effect on health disparities. While these calls and discussions for additional studies on these populations are pertinent and necessary, they are biased because most of them express the perspectives of investigators from industrialized countries that are not part of the societies in which these genomic studies are necessary. Indeed, more global perspectives are needed for a richer real-world discussion, and few investigators from the Global South or US-minority groups have expressed their views.[2] Here, as a global and ethnically diverse network of young and senior investigators (several of whom were affiliated with institutions in both the North and the South during their careers), based on our experience with years of fair North-South collaboration, we complement the calls for more human genomics studies on underrepresented populations by presenting and discussing examples regarding indigenous populations from South America, showing (1) the complexity of applying precision medicine to vulnerable populations, (2) the caveats of conceiving underrepresented ethnic groups as homogeneous, and (3) the relationships between diversity and inclusion in research groups and fair community engagement practices.

Western South America was one of the cradles of civilization in the world.[3,4] Hereafter, we consider the entire Andean region and its adjacent Pacific Coast and Amazon Yunga (the rainforest transitional region between the Andean mountains and the lower Amazonia) as Western South America. The Andean highlands and its east-adjacent Amazon tropical forest are important South American biomes that pose contrasting environmental stressors to

All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Please cite this article in press as: Alvim et al., The need to diversify genomic studies: Insights from Andean highlanders and Amazonians, Cell (2024), https://doi.org/10.1016/j.cell.2024.07.009

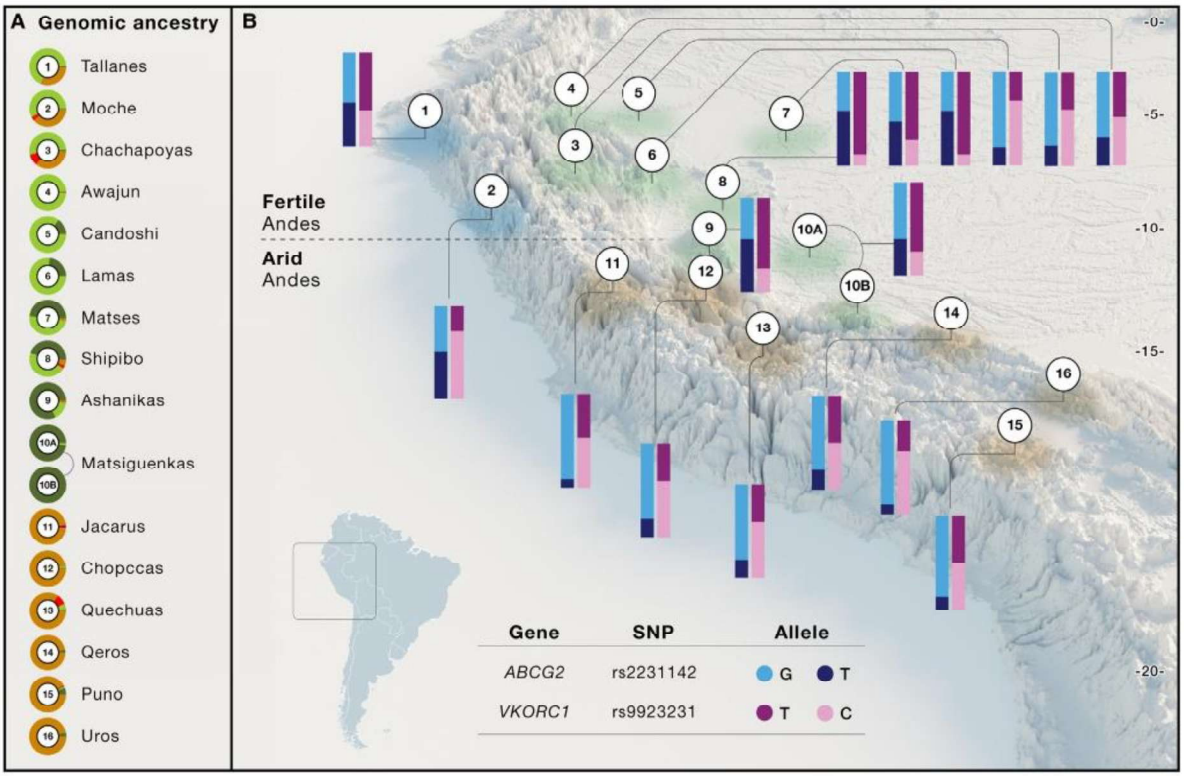


Figure 1. The genetic and geographic landscape of Western South American natives and clinically actionable pharmacoalleles sharply differentiated between the Andean and Amazonian populations

(A) Genetic structure of sixteen native populations from the focus region considered in this study. ADMIXTURE inferences for populations from the coast (1 and 2), Amazon (3–10), and Andes (11–16; K = 5, corresponding to the lowest cross-validation error) are from Borda et al.[4] Colored donut graphics show the average percentages over individuals for five ancestry clusters in each population. Three clusters were associated with distinct Native American groups: one Andean (brown) and two Amazonian (green) clusters. Two clusters were associated with non-native continental ancestries (European [red] and African [dark blue]), caused by admixture with these populations in the last 500 years. There are two populations from the ethnic group Matsiguenka (Sepahua [10A] and Shimaa [10B]).

(B) Information for clinically actionable pharmacoalleles sharply differentiated between the Andean and Amazonian populations. The colored bars indicate the frequency of the pharmacoalleles. The coastal, Andean, and Amazonian populations are located in blue, brown, and green areas, respectively.

their populations. The region of focus of this study (Figure 1) hosts more than thirty million inhabitants with a high prevalence of Native American ancestry. Comparatively, this area is larger than the Iberian Peninsula, France, Switzerland, and Austria together. Studying the genomic diversity (~1.9 million single nucleotide variants [SNVs]) of 294 individuals from 17 indigenous Peruvian populations, Borda et al.[4] (Dataset S1 therein) revealed how the genetic structure and genetic adaptation of Andean highlanders and Amazonians have been influenced by the interplay between geography, history, and culture. Latitudinally, there is an environmental dichotomy between the northern, lower, and so-called *fertile Andes* region (including its adjacent coast) versus the southern, higher, and so-called *arid Andes* region (including its adjacent coast; Figure 1). This north/south dichotomy is recapitulated by the genetic structure of the populations: the patterns of genetic diversity between Andeans and their east-adjacent Amazonian populations differ in the northern fertile Andes with respect to the southern arid Andes. In the northern fertile Andes, west/east gene flow has shaped a shared pattern of ancestry (see populations 1–6 in Figure 1A; Table S1). In the southern arid Andes, the higher altitudes and sharper environmental differences have driven a stronger genetic differentiation between Andean highlanders and Amazonian lowlanders (see Andean populations 11–16 vs. Amazonian populations 7–10 in Figure 1A; Table S1).[4] In addition, this environmental and genetic differentiation between the arid Andean and Amazonian populations has allowed natural selection and other evolutionary forces to act over millennia, shaping differences in the frequencies of genetic variants, including genes related to immune response (*CD45* and *DUOX2*); genes with thyroid (*DUOX2*), cardiovascular (*HAND2-AS1*), and hematological (*TMPRSS6*) functions;[4] and genes related to drug response. The latter are the focus of this Commentary. To exemplify the importance of more human genomics studies on underrepresented populations, we further analyzed the Borda et al.[4] dataset and

Please cite this article in press as: Alvim et al., The need to diversify genomic studies: Insights from Andean highlanders and Amazonians, Cell (2024), https://doi.org/10.1016/j.cell.2024.07.009

identify which of the 19 SNVs that are clinically relevant in pharmacogenetics showed large differences between the Andean and Amazonian populations. We considered as clinically relevant in pharmacogenetics the variants with the top levels (1 and 2) of clinical annotations in the PharmGKB knowledge base, which correspond to higher and moderate evidence of clinical relevance, respectively, following the standardized criteria of PharmGKB (Table S2). Our analysis revealed two clinically actionable variants with surprisingly sharp differentiation between the arid Andean and Amazonian populations (Figure 1B): rs2231142 in *ABCG2* and rs9923231 in *VKORC1* (Table S3).

The pharmacogenetics of rosuvastatin and warfarin in Andean highlanders and Amazonians

Rosuvastatin is a popular statin prescribed to reduce cholesterol levels and prevent cardiovascular events. The rosuvastatin efflux transporter *ABCG2* (ATP Binding Cassette Subfamily G Member 2, Junior Blood Group) has a level 1A PharmGKB clinical annotation and was recently the object of an updated guideline from the Clinical Pharmacogenetics Implementation Consortium (CPIC) involving statins in general and variants in the genes *SLCO1B1* and *CYP2C9* that participate in the metabolism of other statins.[5] rs2231142 (G/T) is a well-known *ABCG2* missense polymorphism also called Q141K or Gln141Lys.[5] Because the allele rs2231142-T (141K) has reduced efflux transporter activity,[5] it causes increased plasma rosuvastatin exposure and is therefore associated with a better response to rosuvastatin therapy but also a greater chance of side effects. In contrast, individuals with the rs2231142-GG genotype (homozygous for Q141) may have a reduced response to therapy to reduce LDL-C. Notably, differences in the frequency of rs2231142-T between the arid Andeans (populations 11–16, average allele frequency = 0.14; Figure 1B) and Amazonians (populations 3–10, average allele frequency = 0.38, F_{ST} = 0.25 between the two groups; Table S2) are greater than those observed between Europeans (0.12) and East Asians (0.31, reported as the highest frequency worldwide). In terms of individuals (observed rs2231142-TT genotypes), following the CPIC, only 2% of Andeans would require a lower initial dose of rosuvastatin or prescription of a different statin compared to 16% of individuals from the Amazonian traditional populations. This result exemplifies the problems of considering Native Americans as a homogeneous group and shows how ancestry and likely environmental differences may contribute to the apportionment of genetic diversity within and between Native American populations, as we also reported for other pharmacoalleles of CYP drug-metabolizer enzymes.[6]

Another more complex example is warfarin, an inexpensive and effective anticoagulant prescribed worldwide to treat thrombosis. The rs9923231 genotypes of the *VKORC1* gene (Vitamin K Epoxide Reductase Complex Subunit 1) have the highest level (1A) of PharmGKB clinical annotation, have been the objects of CPIC guidelines,[7] and should be included in algorithms to calculate individual doses of warfarin. The *VKORC1* protein is a warfarin target in the organism, and individualization of the warfarin dose is important for preventing bleeding if the dose is excessive or preventing further ischemic events if the dose is insufficient. Newer alternative drugs to warfarin such as dabigatran, rivaroxaban, or apixaban, which are commonly used in industrialized countries, are >10 times more expensive than warfarin, preventing their use in low-resource contexts where a relatively simple pharmacogenetic test to assist in warfarin dosage may be more cost-effective and more realistic. *VKORC1* rs9923231 genotypes explain 25%–30% of warfarin dose variation, and patients carrying more rs9923231-T alleles require lower doses of warfarin.[7,8] In the Amazonian region adjacent to the northern fertile Andes, where events of west/east gene flow between these two biomes have been inferred by Borda et al.,[4] the frequencies of rs9923231-T are not particularly different between coastal populations of the fertile Andean region (populations 1 and 2) and Amazonian populations (populations 3–10; Figure 1B). However, while the frequency of rs9923231-T in the southern arid Andean populations fell within the interval of 0.31–0.52 (populations 11–16, average allele frequency = 0.45), the interval in the east-adjacent Amazonian populations is within 0.75–0.88 (populations 7–10, average allele frequency = 0.78, F_{ST} = 0.19 between the two groups; Table S2). The average frequency of rs9923231-T in Amazonian populations 7–10 of Figure 1B, adjacent to the arid Andes, was much greater than in Europeans (~0.40) and West Africans (0.04) and second worldwide only to East Asian populations (0.91) from the 1000 Genomes Project. Thus, differences in rs9923231-T allele frequencies between arid Andean populations and their east-adjacent Amazonian populations are of similar magnitude to those between Europeans and East Asians. This result implies that the health systems and cardiologists from countries such as Peru and Bolivia should expect indigenous Amazonian individuals to more frequently require lower doses of warfarin than indigenous individuals from the arid Andes. Considering the genotype distributions of rs9923231 and the fact that existing warfarin dosage algorithms reduce the initial dosage for CT and TT individuals with respect to CC genotypes, 69% of individuals in the arid Andean populations (populations 11–16) vs. 93% of individuals in the adjacent Amazonian populations (populations 7–10) would require a dosage reduction. However, there are caveats that show how knowing genotype distributions does not suffice. Warfarin dosage algorithms, whether or not genetic information is included, should be adjusted for or be specific for populations/ethnic groups and may be affected by environmental factors. However, in Latin America, algorithms using genetic information have been developed only for admixed Puerto Ricans, Colombians, and Brazilians with low Native American ancestry.[9] No algorithm for warfarin dosage (including or not including genetic information) exists for any predominantly indigenous Latin American population. Moreover, the hypoxic stress in the Andean highlands (either acute or chronic) adds another layer of complexity because it may interact with the cardiovascular system and the mechanism of action of warfarin on clotting pathways. This is critical for Native populations settled in the Andes, which, unlike Asian and African highlanders, have a pattern of genetic adaptation to hypoxia characterized by changes in the cardiovascular system.[10] Importantly, the discussed examples of sharp genetic differentiation

Please cite this article in press as: Alvim et al., The need to diversify genomic studies: Insights from Andean highlanders and Amazonians, Cell (2024), https://doi.org/10.1016/j.cell.2024.07.009

in pharmacoallele frequencies between the arid Andean and Amazonian populations for rosuvastatin/ABCG2-rs2231142 and warfarin/VKORC1-rs9923231 refer to traditional indigenous communities.[3,4] However, emerging urban centers have developed in recent decades, fostered by economic activities such as mining, forestry, and nontraditional agriculture, located in the Amazon Yunga. These centers are predominantly indigenous and congregate individuals from the Andes and the Amazonia, as well as individuals who are the product of recent admixture between these ancestries. Therefore, the cases of rosuvastatin/*ABCG2*-rs2231142 and warfarin/*VKORC1*-rs9923231 drug-gene interactions in Western South America exemplify (1) how population genomic studies on underrepresented populations reveal clinically relevant information, (2) the inadequacy of ignoring the internal genetic structure of ethnic groups such as Native Americans and therefore the limits of concepts such as "ethnic/racial pharmacogenetics," and (3) how environmental stressors on the cardiovascular system, such as hypoxia in the Andes, pose new challenges for understanding the effect of genetic variation on drug response.

In the 18th century, the Spanish scientific traveler and Jesuit Antonio de Ulloa stated that "if you have seen one Indian, you can say that you have seen all of them in terms of skin color and body contexture."[11] Similarly, the influential American-Czech anthropologist Aleš Hrdlička (1869–1943) was a modern representative of this vision that overemphasizes the biological diversity between Native Americans and other groups and neglects the diversity between Native American groups.[12] This oversimplification and intellectual *Ulloa-Hrdlička bias*, even if attenuated, may persist until today. It may derive from a lack of familiarity of investigators from industrialized countries with the local historical, socioeconomic, and environmental context and its nuances.

The social context of genomics research and community engagement

In our experience with fair North-South collaboration to study Peruvian Native American populations,[3,4] the ethnic and social diversity of the members of research groups makes effective community engagement easier. In recent years, engagement of indigenous communities has been critical for the genomics scientific community of developed countries such as the United States and Canada in the Americas, as well as Australia and New Zealand. The Summer Internshle in Genomics initiative is an example of these efforts (the SING initiatives). However, the sociocultural and national contexts of community engagement practices differ by country. The former countries (as well as their scientific community) are predominantly white, and their indigenous populations have almost been exterminated during the last centuries and are currently a minority group. In contrast, in Latin America, most individuals are admixed of European, African, and Native American ancestries, and in particular, in countries such as Mexico, Guatemala, Peru, and Bolivia, indigenous people are not a minority, even if they are socioeconomically disadvantaged. In Peru, for instance, the general population has >60% Native American ancestry, even in urban areas.[3] In this context of Latin American countries where most individuals are admixed, productive scientists and socially committed research institutions frequently have decades of experience with fair interaction with indigenous groups, and importantly, investigators (seniors, students, or technicians) frequently share indigenous genetic ancestry or cultural traits or are related to indigenous people. On the other hand, Latin America is among the most socially unequal regions worldwide, and this inequality frequently permeates academic institutions. For instance, it may happen that in some top-ranked Latin American academic institutions, research groups may be predominantly composed of socioeconomically advantaged people, which may hinder fair community engagement practices because investigators may be ethnically, socially, and culturally far from disadvantageous indigenous populations. Thus, in our experience, it is necessary (but not sufficient) that research groups should be ethnically, socially, and geographically diverse to make fair community engagement more natural. Research and academic institutions, particularly those considered as *elite* institutions in Latin American countries, should adopt these inclusive practices. Moreover, the global genomics community should not uncritically attempt to translate the current experiences and models of community engagement of the Global North to admixed and predominantly indigenous countries such as Mexico, Guatemala, Peru, and Bolivia. Instead, formal efforts toward fair North-South collaboration should allow Global North countries to consider fair experiences of community engagement implemented in the Global South.

In conclusion, additional genomic studies are necessary to reduce our knowledge gap about the genetic architecture of disease in European vs. non-European individuals and populations. Beyond the presented case for clinically actionable genotypes in Andeans and Amazonians, another illustrative example regards an admixed Brazilian population-based cohort, mostly a product of the admixture of Europeans, Africans, and Native Americans. In this admixed cohort, potential loss-of-function variants mapped in genomic fragments of non-European origins were nearly three times less likely to have a ClinVar annotation than potential loss-of-function variants mapped in fragments of European origin.[13] Moreover, developing genomic studies and applying precision medicine to vulnerable populations requires (1) considering their biological, environmental, and social specificities; (2) fair local and international scientific collaboration between investigators and institutions; and (3) appropriate community engagement practices. Fair collaboration should leverage institutions and research groups of excellence in the Global South by sharing leadership in research and authorship,[3,4,6] as well as appropriate funding from global sources for training of young investigators from the Global South to perform research in their countries and strengthen local research capacities. Funding agencies should objectively assess these practices and incorporate them into their policies and funding decisions. Examples of efforts following these practices include (1) the CRONICAS Center for Excellence in Noncommunicable Diseases in Peru (https://www.cronicas-upch.pe/en/); (2) the research network led by Robert H. Gilman from the Johns Hopkins School of

Please cite this article in press as: Alvim et al., The need to diversify genomic studies: Insights from Andean highlanders and Amazonians, Cell (2024), https://doi.org/10.1016/j.cell.2024.07.009

Public Health that has maintained for more than 30 years a productive research group and infrastructure in Peru focused on infectious diseases (https://magazine.jhsph.edu/2012/la-familia-gilman); (3) the *Rede Iberoamericana de Farmacogenética* (RIBEF),[6] which in addition to performing collaborative research recently launched The Mérida/T'Hó Declaration about the application of pharmacogenetics in indigenous Latin Americans[14] in four autochthonous languages (Miskito, Nahualt, Maya, and Quichua), warning about the need to consider traditional medicine practices (that frequently coexist with allopathic medicine) in pharmacogenetic and clinical studies; (4) the Peruvian Genome Project[4] led by the Peruvian government Instituto Nacional de Salud, which was conceived and executed with a transdisciplinary and intercultural approach; and (5) the Mexican Biobank Project.[15]

As we show here, a correct assessment of the apportionment of human genetic diversity for different genetic variants in Native American groups, as well as new models of fair global scientific collaboration, will favor the application of precision medicine to vulnerable populations in their complex environmental and social contexts, allowing the maxim "Act locally, think globally" to be applied to human genomics studies.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.07.009.

ACKNOWLEDGMENTS

We thank Hanaísa Sant'Anna, Irene Gallego-Romero, and Prof. Sérgio Pena for editorial suggestions and Norbert Tavares from the Chan Zuckerberg Initiative because an old exchange of emails between him and E.T.-S. inspired the paragraph about community engagement. This study was supported by the Instituto Nacional de Salud del Peru (INS), Brazilian Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq 440238/2022-6, 402533/2022-4, 314344/2020-9, 407046/2023-2, 406913/2022-6, and 312807/2022-8), CNPq-Amazon Web Services (440066/2020-4), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Minas Gerais State Agency for Research (FAPEMIG APQ-04418-22, RED-00089-23, and APQ-04228-24), Programa Nacional de Genômica e Saúde de Precisão – Genomas SUS from the Brazilian Ministry of Health, and the Peruvian National Council for Science, Technology and Innovation (CONCYTEC N°136–2018). M.L.S. is supported by D43 TW007393 awarded by the Fogarty International Center of the U.S. National Institutes of Health. M.M.A. was supported by the CGEn HostSeq/CIHR fellowship (CGE 185054) and the SickKids Restracomp Fellowship.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019 Mar 21). The Missing Diversity in Human Genetic Studies. Cell *177*, 26–31. https://doi.org/10.1016/j.cell.2019.02.048.
2. Fatumo, S., Yakubu, A., Oyedele, O., Popoola, J., Attipoe, D.A., Eze-Echesi, G., Modibbo, F.Z., Ado-Wanka, N., et al.; 54gene Team, NCD-GHS Consortium (2022). Promoting the genomic revolution in Africa through the Nigerian 100K Genome Project. Nat. Genet. *54*, 531–536.
3. Harris, D.N., Song, W., Shetty, A.C., Levano, K.S., Cáceres, O., Padilla, C., Borda, V., Tarazona, D., Trujillo, O., Sanchez, C., et al. (2018). Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. Proc. Natl. Acad. Sci. USA *115*, E6526–E6535. https://doi.org/10.1073/pnas.1720798115.
4. Borda, V., Alvim, I., Mendes, M., Silva-Carvalho, C., Soares-Souza, G.B., Leal, T.P., Furlan, V., Scliar, M.O., Zamudio, R., Zolini, C., et al. (2020). The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture. Proc. Natl. Acad. Sci. USA *117*, 32557–32565.
5. Cooper-DeHoff, R.M., Niemi, M., Ramsey, L.B., Luzum, J.A., Tarkiainen, E.K., Straka, R.J., Gong, L., Tuteja, S., Wilke, R.A., Wadelius, M., et al. (2022). The Clinical Pharmacogenetics Implementation Consortium Guideline for SLCO1B1, ABCG2, and CYP2C9 genotypes and Statin-Associated Musculoskeletal Symptoms. Clin. Pharmacol. Ther. *111*, 1007–1021.
6. Rodrigues-Soares, F., Peñas-Lledó, E.M., Tarazona-Santos, E., Sosa-Macías, M., Terán, E., López-López, M., Rodeiro, I., Moya, G.E., Calzadilla, L.R., Ramírez-Roa, R., et al. (2020). Genomic Ancestry, CYP2D6, CYP2C9, and CYP2C19 Among Latin Americans. Clin. Pharmacol. Ther. *107*, 257–268. https://doi.org/10.1002/cpt.1598.
7. Johnson, J.A., Caudle, K.E., Gong, L., Whirl-Carrillo, M., Stein, C.M., Scott, S.A., Lee, M.T., Gage, B.F., Kimmel, S.E., Perera, M.A., et al. (2017). Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Pharmacogenetics-Guided Warfarin Dosing: 2017 Update. Clin. Pharmacol. Ther. *102*, 397–404. https://doi.org/10.1002/cpt.668.
8. Oscanoa, T.J., Guevara-Fujita, M.L., Fujita, R.M., Muñoz-Paredes, M.Y., Acosta, O., and Romero-Ortuño, R. (2023). Association between polymorphisms of the VKORC1 and CYP2C9 genes and warfarin maintenance dose in Peruvian patients. Br. J. Clin. Pharmacol. *90*, 769–775. Published online November 8. https://doi.org/10.1111/bcp.15958.
9. Roche-Lima, A., Roman-Santiago, A., Feliu-Maldonado, R., Rodriguez-Maldonado, J., Nieves-Rodriguez, B.G., Carrasquillo-Carrion, K., Ramos, C.M., da Luz Sant'Ana, I., Massey, S.E., and Duconge, J. (2020). Machine Learning Algorithm for Predicting Warfarin Dose in Caribbean Hispanics Using Pharmacogenetic Data. Front. Pharmacol. *10*, 1550. https://doi.org/10.3389/fphar.2019.01550.
10. Crawford, J.E., Amaru, R., Song, J., Julian, C.G., Racimo, F., Cheng, J.Y., Guo, X., Yao, J., Ambale-Venkatesh, B., Lima, J.A., et al. (2017). Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans. Am. J. Hum. Genet. *101*. https://doi.org/10.1016/j.ajhg.2017.09.023.
11. de Ulloa, A. (1772). Noticias Americanas: Entretenimientos phisicos-historicos, sobre la América Meridional, y la Septentrional Oriental: Comparación general de los territorios, climas y produciones.
12. Hrdlička, A. (1928). The Origin and Antiquity of Men in America. Bull. N. Y. Acad. Med. *4*, 802.
13. Naslavsky, M.S., Scliar, M.O., Nunes, K., et al. (2021). Biased pathogenic assertions of loss of function variants challenge molecular diagnosis of admixed individuals. Am. J. Med. Genet. C Semin. Med. Genet. *187*, 357–363. https://doi.org/10.1002/ajmg.c.31931.
14. Sosa-Macías, M., Fricke-Galindo, I., Fariñas, H., et al. (2023). Pharmacogenetics: ethnicity, treatment and health in Latin American populations. Pharmacogenomics *24*, 489–492. https://doi.org/10.2217/pgs-2023-0098.
15. Sohail, M., Palma-Martínez, M.J., Chong, A.Y., Quinto-Cortés, C.D., Barberena-Jonas, C., Medina-Muñoz, S.G., Ragsdale, A., Delgado-Sánchez, G., Cruz-Hervert, L.P., Ferreyra-Reyes, L., et al. (2023). Mexican Biobank advances population and medical genomics of diverse ancestries. Nature *622*, 775–783. https://doi.org/10.1038/s41586-023-06560-0.

Nossa publicação na revista *Cell*, precisamente por enfatizar a necessidade de diversificar estudos genômicos teve um importante impacto na mídia internacional, com reportagens na Science News (https://www.science.org/content/article/not-all-drugs-work-same-indigenous-groups-very-large-genetic-differences, Perez-Ortega 2024) e na radio-televisão pública alemã Deustsche Welle (https://www.dw.com/en/closing-the-genomics-gap-how-to-improve-indigenous-health/a-69890592, Schwaller 2024) (Figura 12).

Figura 12 Repercussão midiática do comentário publicado na revista Cell. A) Capa da reportagem publicada na DW alemã. B) Capa da reportagem publicada pela revista Science.

2.4. Discussão

O problema da falta de diversidade nos bancos de dados genéticos não é algo novo no mundo científico (Need e Goldstein 2009; Bustamante, De La Vega e Burchard 2011; Popejoy e Fullerton 2016; Sirugo *et al.* 2019). A geração de dados do PGP ainda continua, com o objetivo de chegar a 1000 genomas peruanos completos (Guio *et al.* 2024). Atualmente, já foram produzidos dados de sequenciamento de 150 indivíduos peruanos nativos e miscigenados, além da genotipagem de 850 indivíduos, incluindo aqueles no banco de dados apresentado em Borda et al. 2020. Nosso banco de dados já tem contribuído para o objetivo de aumentar a representatividade nativo americana nos estudos genômicos e garantir a replicabilidade da pesquisa, sendo depositado no *European Genome-phenome Archive*

(https://ega-archive.org/datasets/) para a publicação de 2020. Ele tem sido usado como painel de referência para nativos em outros estudos, como em Guimarães Alves *et al.* 2021, Naslavsky *et al.* 2022a e Naslavsky *et al.* 2022b.

Além de enriquecer a diversidade genética dos bancos de dados disponíveis, nosso artigo trouxe elucidações a respeito da adaptação aos ambientes andino e amazônico. A replicação de sinais de seleção natural com diferentes bancos de dados e métodos de genotipagem pode ser tratada como um indicativo que o sinal em si não seja um falso positivo. Dois dos genes nos quais encontramos sinais de seleção natural nos Andes foram detectados como candidatos à seleção natural no mesmo ambiente após a publicação de 2020 (Caro-Consuegra *et al.* 2022). O gene *HAND2-AS1* codifica um RNA anti-senso que regula o gene *HAND2*, associado à função cardíaca. O gene *DUOX2*, que já se tratava de uma replicação de resultado (Zhou *et al.* 2013; Jacovas *et al.* 2018), apresenta função tanto no estresse oxidativo quanto na atividade ante patogênica na via respiratória. Além disso, ele codifica uma proteína fundamental no processo de produção do hormônio tireoidiano.

Na região amazônica ainda não pudemos observar uma replicação de nosso sinal no gene *PTPRC* (Caro-Consuegra *et al*. 2022; Lea *et al.* 2022; Couto-Silva *et al.* 2023). Couto-Silva *et al.* 2023 publicaram um sinal de seleção natural positiva no SNV rs2659540, no gene *PPP3CA*, apontado como potencial proteção para doença de Chagas. Observando a importância deste resultado para populações latino-americanas, revisitamos nossos resultados de seleção natural. Considerando o fato de que o desenho experimental executado por Couto-Silva *et al.* (2023) compara populações amazônicas com populações externas à América, verificamos os valores de F_{ST} obtidos da comparação Amazônia e CDX e na comparação Andes e CDX. Foi possível observar a presença de valores altos de F_{ST}, localizados acima do 99º percentil tanto para a população amazônica quanto para a população andina (Tabela 1; Figura 13A). Especificamente, a variante rs874116, localizada numa região promotora do gene *PPP3CA*, apresenta o maior valor de F_{ST} dentro do gene nas duas populações. Para verificar a possibilidade de extensão de haplótipos, plotamos um gráfico de bifurcação de haplótipos (Figura 13 C), onde é possível visualizar um haplótipo de aproximadamente 1 Mb contendo o alelo T da variante rs874116 com alta frequência nas populações nativas. Este resultado pode ser um indicativo de que o sinal encontrado por Couto-Silva pode ser extensível à população andina.

Tabela 1 SNVs presentes na região genica do gene PPP3CA contidos no array Omni Array 2.5 M.

SNV	F_{ST} Andes vs. CDX	F_{ST} Amazônia vs. CDX	Percentil AND	Percentil AMZ	Região	REF	AMZ	AND	Costa	YRI	IBS	CDX
rs874116	0.5111	0.6176	0.9889	0.9933	Promotor	C	0.2632	0.2921	0.4538	1	0.9860	0.8626
rs79174673	0.2959	0.5468	0.9142	0.9856	Intron	T	0.2225	0.2892	0.2923	0.9907	0.9626	0.7033
rs71614685	0.3206	0.5405	0.9307	0.9848	Intron	G	0.2147	0.2598	0.2769	0.9167	0.7850	0.6978
rs34505183	0.3078	0.5340	0.9224	0.9839	Intron	C	0.2147	0.2647	0.2769	0.8611	0.7850	0.6923
rs17030795	0.3141	0.5340	0.9265	0.9839	Intron	A	0.2147	0.2598	0.2769	0.9167	0.7757	0.6923
rs73833252	0.3307	0.5331	0.9365	0.9837	Intron	C	0.2723	0.3137	0.4077	0.8380	0.9486	0.7527
rs2659504	0.3105	0.5128	0.9244	0.9806	Intron	T	0.2696	0.3137	0.4	0.8148	0.9439	0.7363
rs1506804	0.3105	0.5128	0.9244	0.9806	Intron	C	0.2696	0.3137	0.4077	0.8194	0.9439	0.7363
rs12639641	0.2575	0.4989	0.8823	0.9782	Intron	T	0.2408	0.2892	0.3154	0.8426	0.9533	0.6758
rs17030767	0.2697	0.4963	0.8938	0.9776	Intron	T	0.2421	0.2794	0.3077	0.9028	0.9533	0.6758
rs6532920	0.2634	0.4899	0.8879	0.9763	Intron	G	0.2147	0.2647	0.2692	0.5370	0.6449	0.6593
rs1506803	0.4057	0.4625	0.9673	0.9696	Intron	C	0.3063	0.2990	0.3923	0.4954	0.9206	0.7967
rs1395475	0.4057	0.4625	0.9673	0.9696	Intron	C	0.3079	0.2990	0.3984	0.3009	0.9206	0.7967
rs2850373	0.3536	0.4198	0.9479	0.9565	Intron	G	0.2696	0.2647	0.3923	0.8287	0.9486	0.7363
rs72681029	0.2728	0.4066	0.8965	0.9514	Intron	G	0.2618	0.3186	0.3923	0.9861	0.9159	0.7143

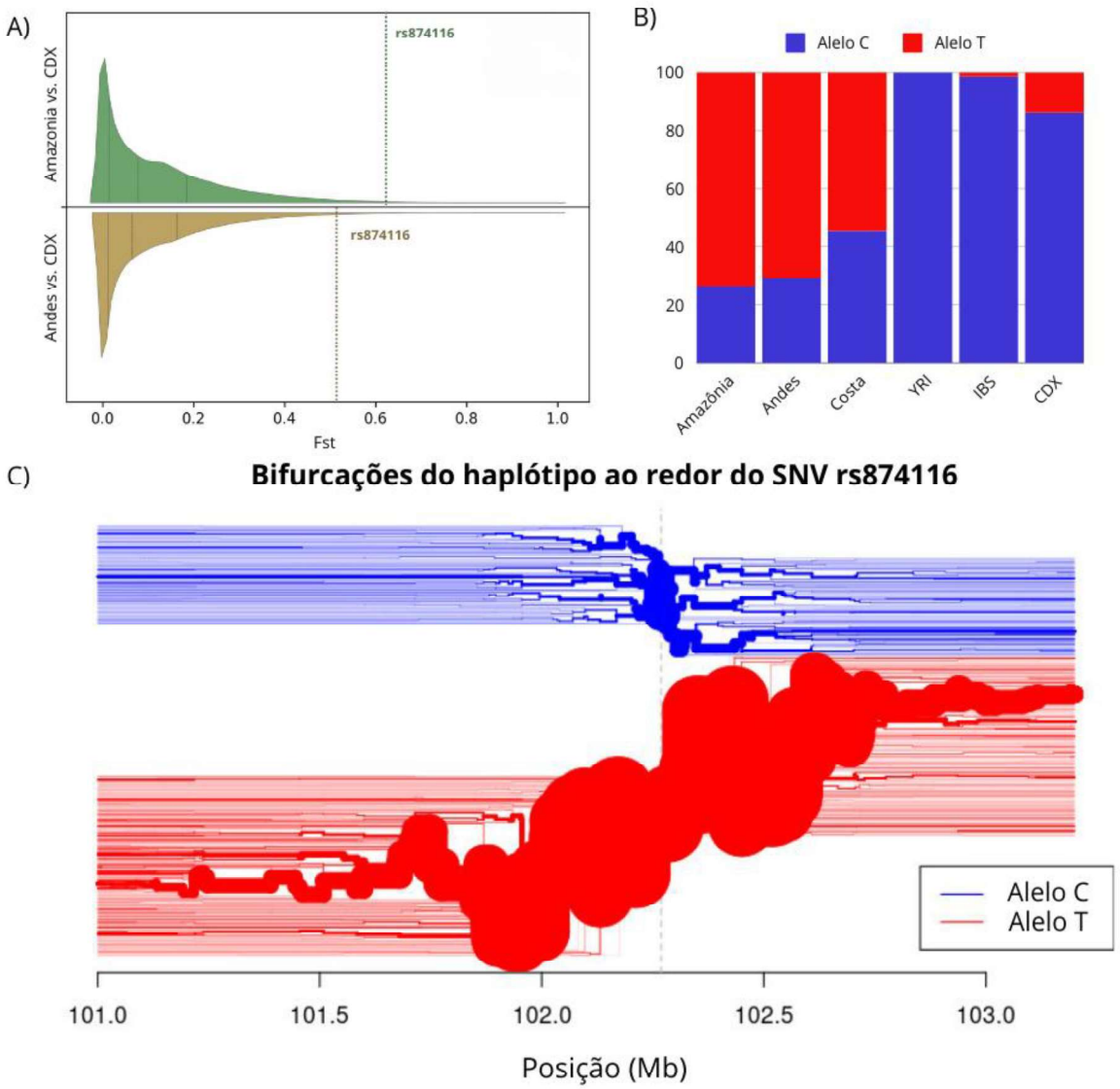


Figura 13 Resumo da variante rs874116 como potencial sinal de seleção natural. A) Plot de violino representando a distribuição dos valores de F_{ST} na comparação Amazônia versus CDX (em verde) e Andes versus CDX (em marrom). As linhas pontilhadas representam a posição do rs874116 em cada distribuição. B) Frequências alélicas da variante rs874116 em populações diversas. C) Gráfico de bifurcação do haplótipo ao redor da variante rs874116 nas populações nativas peruanas, considerando Andes, Amazônia e Costa.

2.5. Perspectivas

O estudo da seleção natural é fundamental para compreender melhor a interação genótipo-fenótipo e sua relação com o ambiente, estando suas assinaturas também associadas à importância funcional e potencialmente médica das variantes genéticas. A adaptação local exigida pelos ambientes andino e amazônico pode levar a diferenciação de genótipos e fenótipos através da seleção natural adaptativa, que pode ser detectada de formas diferentes,

dependendo da origem do alelo selecionado (mutação *de novo* ou *standing variation*) ou o tipo de fenótipo (monogênico ou poligênico) (Rees, Castellano e Andrés 2020).

Em 2020, apontamos dois sinais de assinatura de seleção natural positiva nos Andes (*HAND2-AS1* e *DUOX2*) e um na Amazônia (*PTPRC*), encontrados através dos métodos de varredura genética *Cross-Population Extended Haplotype Homozygosity* (XP-EHH) e *Population Branch Statistic* normalizado (PBSn), que detectam seleção positiva recente monogénica se baseando comparações populacionais, deixando em aberto a aplicação de métodos mais sensíveis a outras assinaturas de seleção natural, incluindo a seleção poligênica.

Recentemente, trabalhamos em uma metodologia de comparação de distribuição dos valores de F_{ST} e PBS, através do teste de Kolmogorov-Smirnov (KS). O teste KS não é capaz de apontar a direção do desvio, o que indicaria se a via tem uma proporção maior do que o esperado de SNVs com valor PBSn elevado. Para corrigir isso, usamos uma versão modificada do teste realizando o somatório de todas as distâncias entre as distribuições de proporções acumuladas da via e do banco de dados. O resultado positivo indica o deslocamento para a direita, o que indicaria um número maior de SNVs diferenciados na via do que o esperado pelo acaso. Com ele pudemos encontrar algumas funções biológicas do Gene Ontology (https://geneontology.org/) com aparente desvio em relação ao banco de dados total, como o Processo Metabólico de Xenobióticos (genes associados com o metabolismo de xenobióticos), com deslocamento positivo na população andina (valor de p do KS = 2.77×10^{-8} - Figura 14).

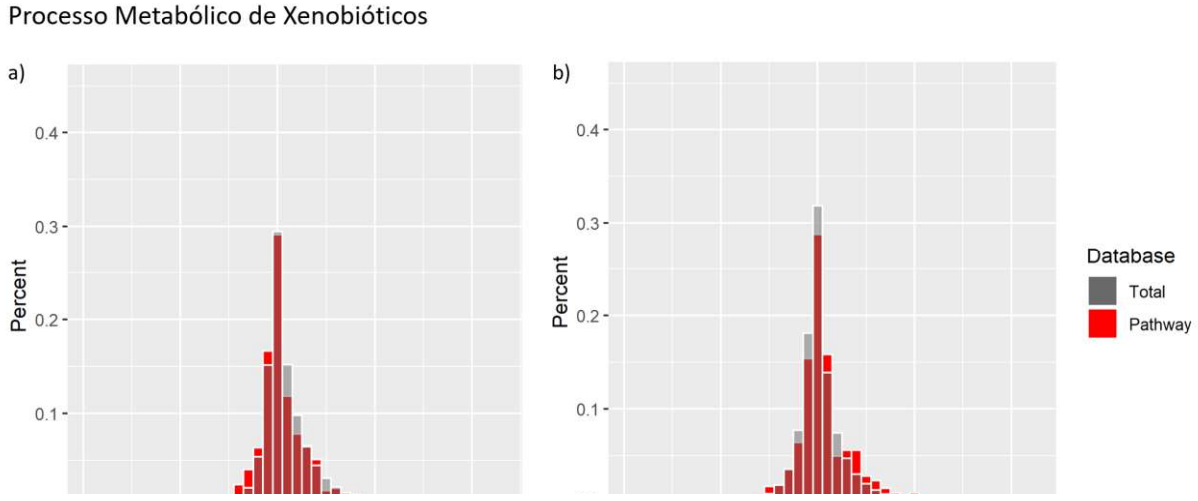


Figura 14 Distribuição dos valores de PBSn na anotação "Processo metabólico de xenobióticos". Nessa via se encontram genes como SULT1A1, CYP26B1, CYP2C18, CYP2C19, CYP2C9, CYP2E1, TBXAS, FMO3, AIP, ACSM2B e ACSM1 com SNPs com altos valores de PBSn para a população andina (PBSn > 0.1, valor de p < 0.05). a) Distribuição dos valores de PBSn calculados usando as populações amazônicas como alvo. Valor de p para KS = 2.68×10^{-11}. b) Distribuição dos valores de PBSn calculados usando as populações andinas como alvo. Valor de p para o KS = 2.77×10^{-8}.

Além da falta de exploração da seleção poligênica, o gene *DUOX2*, relacionado a produção de peróxido de hidrogênio (H_2O_2) na via metabólica dos hormônios tireoidianos e na ação antimicrobiana em mucosas intestinais e pulmonares, apesar de ter sido apontado como candidato a seleção natural por outros três estudos, como discutido anteriormente, ainda segue sem validação funcional.

Como forma de solucionar parte destas lacunas, estamos trabalhando em um projeto que possui os objetivos de:

1) Identificar sinais de seleção natural adaptativa monogênica e poligênica em populações nativas da América do Sul, identificando genes, regiões não codificantes e fenótipos poligênicos envolvidos em processos adaptativos aos ambientes contrastantes dos Andes e da Amazônia, e a sua interseção com a arquitetura genética de doenças complexas e com genes alvos de fármacos existentes ou em desenvolvimento.

2) Executar a validação funcional dos sinais de assinatura de seleção adaptativa publicados em Borda *et al.* 2020:

   a) Avaliar a correlação entre sinais de seleção natural positiva encontrados em Borda *et al.* (2020) dados de transcriptoma de uma população indígena peruana residente nos Andes e na região de Selva Alta amazônica.

   b) Testar a hipótese que os alelos do SNP candidato a seleção natural no gene *DUOX2* estão associados a diferenças na função proteica.

Para realizar este projeto, iremos focar na validação estatística da metodologia de seleção poligênica descrita e a usaremos com resultados outros testes de seleção monogênica. Ainda, aplicaremos outros métodos de detecção de seleção poligênica já publicados, como o método Sgnet, que detecta redes gênicas sob possível seleção natural (Gouy, Daub e Excoffier 2017). Avaliaremos a coerência de dados de aDNA publicados de populações nativas americanas com os sinais de seleção natural obtidos a partir de populações atuais e detectar seleção introgressiva e poligênica com o método descrito por Gouy e Excoffier 2020. Para dar força aos sinais poligênicos, analisaremos a congruência de todos os sinais encontrados com os diferentes testes.

Na parte de validação funcional, avaliaremos a correlação entre sinais de seleção natural positiva encontrados em Borda *et al.* (2020), com os níveis de expressão gênica

diferencial entre alta e baixa altitude em *Peripheral Blood Mononuclear Cells* (PBMC) em estado basal e estimulados por antígenos virais, de bactérias e de fungos de (Jaramillo-Valverde *et al*., em submissão).

Além disso, usaremos a edição gênica via CRISPR/cas9 para criar duas sub linhagens de células epiteliais foliculares primárias de tireoide humana (Nthy-ori 3-1 - Sigma-Aldrich), que expressam o gene *DUOX2* (https://gtexportal.org/home/gene/DUOX2). As linhagens se basearam no genótipo do SNV rs269868, sendo uma homozigota para o alelo C (selecionado positivamente nos Andes) e outra homozigota para o alelo T. Para testar o efeito da hipóxia, as linhagens serão expostas a diferentes concentrações de oxigênio: 2,5%, 5% e concentração atmosférica, por 4, 24 e 48 horas. Nelas iremos comparar (em triplicata): (i) o nível de expressão gênica via RT-PCR, (ii) a produção de H_2O_2 pela medida de explosão respiratória e (iii) o nível de expressão proteica através de Western Blot.

CAPÍTULO 3. Inferência da Dinâmica da Miscigenação em Latino Americanos

3.1 Introdução

No contexto da história demográfica de América Latina, caracterizada principalmente por cinco séculos de miscigenação de populações nativas americanas (autóctones), europeias e africanas, a ancestralidade local cromossômica, inferida a partir de dados de SNVs em alta densidade ao longo do genoma, nos fornece mais informações para inferir a história da miscigenação. Isso acontece graças à recombinação, que faz com que cromossomos de ancestralidades diferentes se misturem ao longo do tempo, permitindo que um mesmo cromossomo seja um mosaico de fragmentos de ancestralidades diferentes, sendo estes fragmentos chamados de *tracts*. O esperado é que os fragmentos cromossômicos de uma determinada ancestralidade sejam menores quanto maior o tempo desde o evento de miscigenação que inseriu aquela ancestralidade na população (Figura 15).

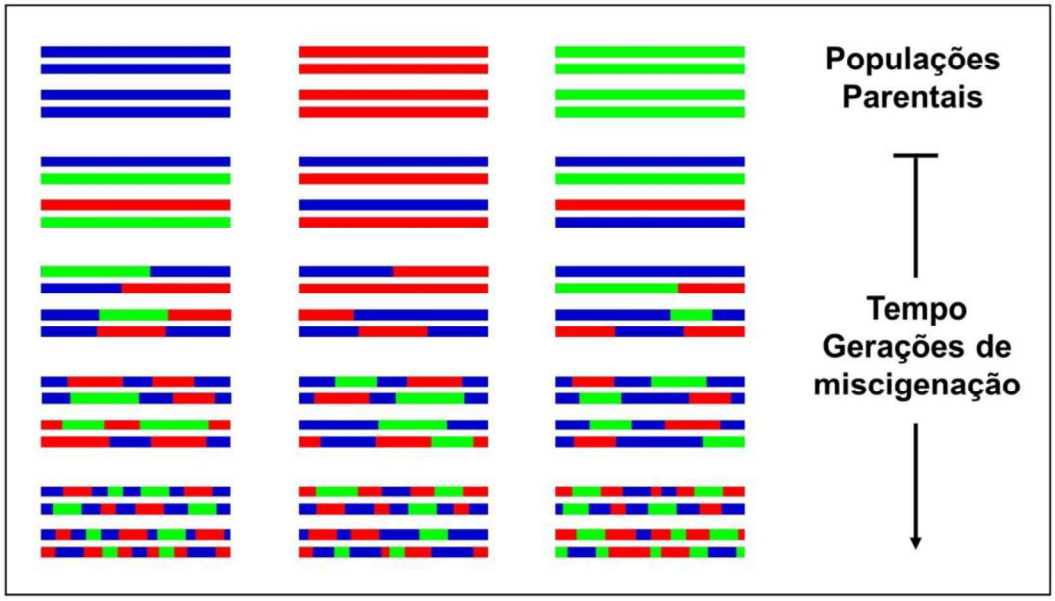


Figura 15 Padrão esquemático dos fragmentos cromossômicos de ancestralidade ao longo das gerações. Cada cor representa uma origem ancestral distinta. Devido ao processo de recombinação, os cromossomos que na primeira geração eram completos de uma única ancestralidade se quebram e formam pequenos fragmentos de ancestralidades diferentes. Fonte: Tarazona-Santos *et al*. 2016.

A nível populacional, se temos uma alta proporção de pequenos *tracts* de uma ancestralidade isso significa que a inserção dessa população parental na população miscigenada aconteceu há muitas gerações, enquanto se tivermos uma alta proporção de grandes *tracts* de uma ancestralidade isso significa miscigenação recente. Dessa forma, podemos usar a distribuição dos *tracts* de ancestralidade como um *relógio recombinatório,* termo que usamos em analogia ao termo *relógio molecular* (Zuckerandl e Pauling 1962).

Para recuperar este conhecimento através dos *tracts* de ancestralidade e tornar nosso relógio recombinatório legível, Kehdy *et al*. 2015 propuseram uma metodologia baseada num enfoque bayesiano para inferir os parâmetros de um modelo demográfico da miscigenação pós-colombiana continental nos últimos 500 anos (aproximadamente 20 gerações). No enfoque bayesiano, inferimos as distribuições a *posteriori* de parâmetros a partir de distribuições a *priori* e das informações derivadas dos dados, nesse caso dados genéticos sobre ancestralidade local cromossômica.

A metodologia de Kehdy *et al.* 2015, baseada em Computação Bayesiana Aproximada (ABC), consiste em comparar a distribuição observada dos *tracts* de diferentes ancestralidades das populações alvo (para as quais queremos inferir os parâmetros que definem sua história evolutiva) com distribuições de *tracts* geradas através de simulações de cenários de miscigenação (cujos parâmetros são amostrados de uma distribuição *a priori*), possibilitando a identificação dos cenários que formassem populações simuladas com distribuições dos tamanhos dos *tracts* próximos das observadas. Na metodologia ABC, os valores dos parâmetros utilizados para simular esses cenários que originaram distribuições de *tracts* muito próximas das observadas, têm distribuições que aproximam as distribuições a posteriori dos parâmetros que queremos inferir na população alvo (Beaumont *et al*. 2002).

Porém, a metodologia de Kehdy *et al*., publicada em 2015, possui limitações, como a perda de resolução nos intervalos de credibilidade de 90% *High Posteriori Density* (HPD) dos pulsos mais antigos, que são muito amplos. O método também se apresenta limitado a três pulsos de miscigenação, o que não permite sua generalização para outras populações com história de miscigenação diferentes da brasileira. Além disso, a metodologia ainda não passou por validação estatística. Dessa forma definimos os objetivos deste trabalho, que consistem não só em validar a metodologia, mas implementar melhorias para que a inferência seja mais acurada, incluindo a reparametrização de acordo com a população avaliada. Como objetivo futuro, visamos a implementação do método para o cromossomo X, o que permitiria inferir a dinâmica do viés sexual nas populações miscigenadas.

3.2 Modelo demográfico e genético populacional da miscigenação

As metodologias de genética de populações mais modernas são baseadas em modelos estatísticos (Johri *et al.* 2022). Assim, nós focamos em construir e aprimorar um modelo da miscigenação nas Américas nos últimos 500 anos cujos parâmetros descrevam razoavelmente a dinâmica desse processo. Inicialmente, e seguindo Kehdy *et al.* (2015) podemos conceber um modelo que divide a dinâmica migratória em três pulsos: um antigo, correspondente ao final século XVI e início do século 17 (Pulso 1), um mediano, correspondente ao século XVII (Pulso 2), e um recente, correspondente ao final do século XIX e início do século XX (Pulso 3). Este modelo permite um número maior de pulsos, mas nosso modelo inicial, que nós usamos para ilustrar o conceito, é limitado a três pulsos. Nas Américas, consideramos três populações ancestrais: Europeus (EUR), Africanos (AFR) e Nativos Americanos (NAT).

A contribuição dessas populações na formação da população alvo nos três pulsos são os nossos parâmetros. Portanto, considere *m* a proporção de migrantes pertencentes a uma população *i* que chegam à população alvo na geração *g*, o que nos leva a definição de cada parâmetro como $m_{i,g}$. Na metodologia descrita por Kehdy *et al.* 2015, considerando apenas cromossomos autossômicos, temos 9 parâmetros $m_{i,g}$ (Figura 16-A e B). Assim, a combinação de 9 valores para os 9 parâmetros $m_{i,g}$ (para três pulsos de miscigenação *g* e três populações parentais *i*) define um cenário de miscigenação. Nosso objetivo é usar dados genéticos de uma população (as distribuições dos tamanhos dos *tracts* para cada ancestralidade) para inferir as distribuições a posteriori desses parâmetros $m_{i,g.}$

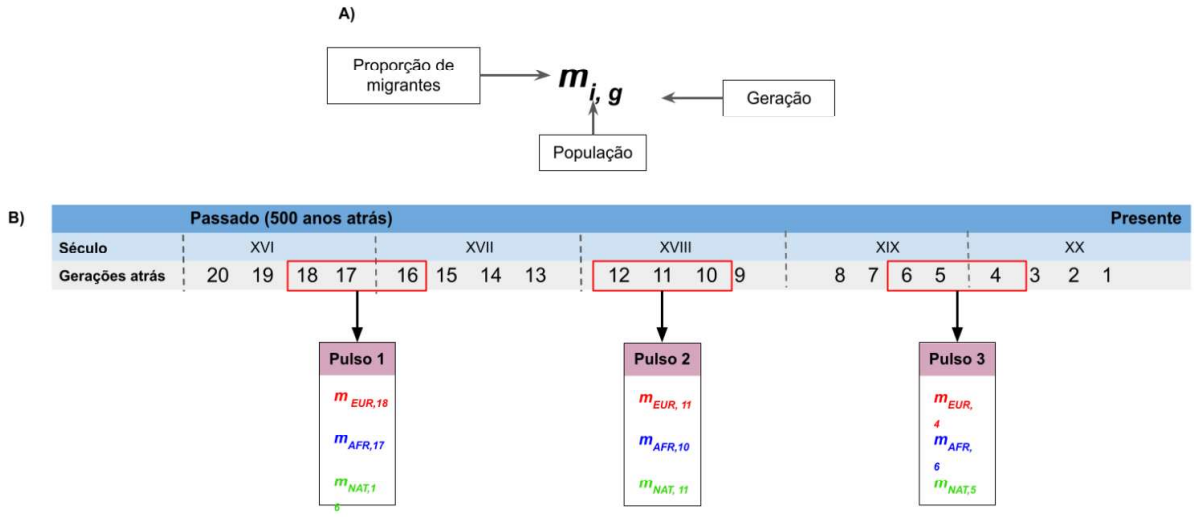


Figura 16 Modelo demográfico de miscigenação pós-colombiana e identificação gráfica dos parâmetros m de migração. A) Características dos parâmetros $m_{i,g}$ no método aplicado nos cromossomos autossômicos. B) Definição dos pulsos de migração e ao longo do tempo e dos parâmetros $m_{i,g}$ associados a cada pulso na aplicação do método nos cromossomos autossômicos.

3.2.1 Formalização matemática e generalização da dinâmica da miscigenação

Do ponto de vista teórico, a partir dos valores de contribuições de cada população parental ao longo de pulsos migracionais ($m_{i,g}$), nosso ex-aluno de Mestrado em Bioinformática (atualmente doutorando em Bioinformática, Faria 2023), derivou a fórmula para estimar os valores finais de cada ancestralidade em cada geração e analisou como cenários diferentes de dinâmicas de miscigenação, determinam as proporções de miscigenação acumulada ao longo dos pulsos ou gerações.

Para isso, definimos *M* como a proporção de ancestralidade global de uma população parental *a* na população miscigenada na geração *g*. Este valor irá variar a cada geração em que houver um fluxo migratório $m_{i,g}$ (Figura 17).

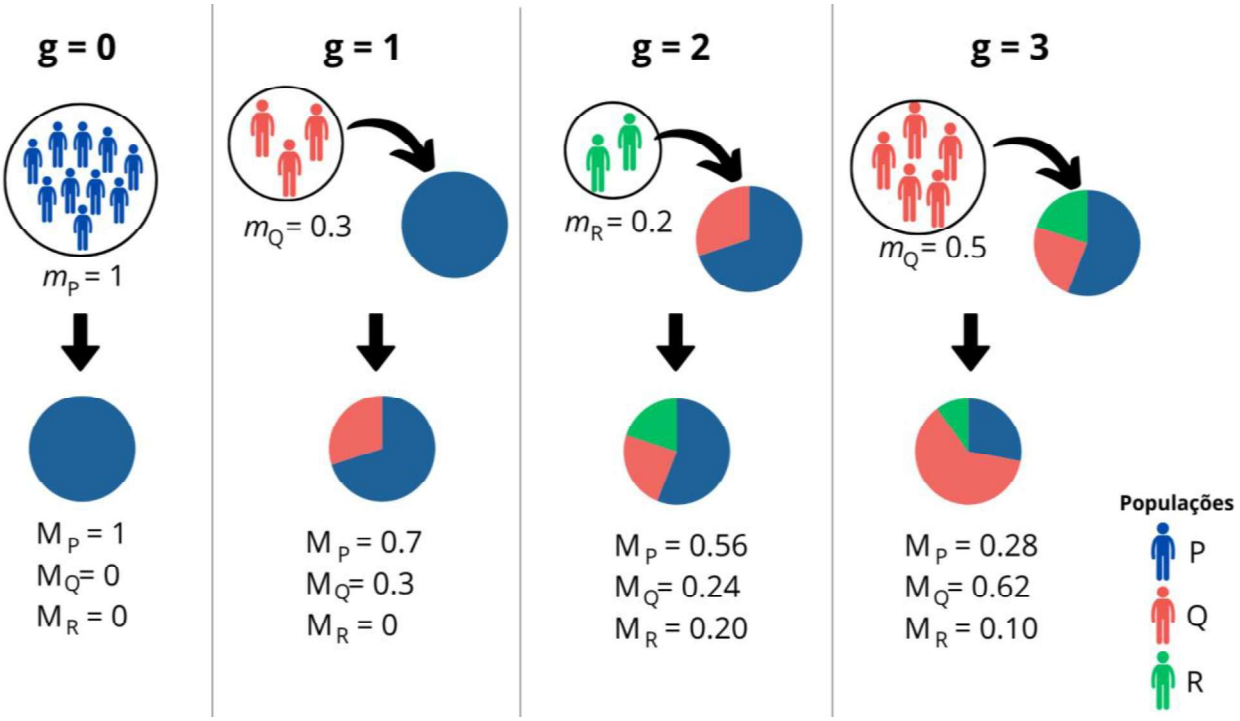


Figura 17 Ilustração da influência dos valores de $m_{i,g}$ nos valores de $M_{a,g}$. Na geração g = 0 a população é formada somente pela população P, tendo valores de $m_{P,0}$ e $M_{P,0}$ = 1 e os valores de $M_{Q,0}$ e $M_{R,0}$ são iguais a zero. Na geração g =1 há a chegada de um fluxo $m_{Q,1}$ = 0,3, o que faz com os valores de $M_{P,1}$ = 0,7, $M_{Q,1}$ = 0,3 e $M_{R,1}$ = 0. Na geração g =2 há a chegada de um fluxo $m_{R,2}$ = 0,2, o que faz com os valores de $M_{P,2}$ = 0,56, $M_{Q,2}$ = 0,24 e $M_{R,2}$ = 0,20. Na geração g =3 há a chegada de um fluxo $m_{Q,3}$ = 0,5, o que faz com os valores de $M_{P,3}$ = 0,28, $M_{Q,3}$ = 0,62 e $M_{R,3}$ = 0,10.

Para determinação dos valores de $M_{a,g}$, foi proposta uma equação de primeiro grau que usa os valores de $m_{i,g}$ em um cenário de miscigenação. Considerando $M_{a,g}$ a proporção final esperada de contribuição da população ancestral a na geração g e $m_{i,g}$ a proporção de imigrantes de ancestralidade i introduzidos na população miscigenada na geração g, a descrição de nosso modelo se inicia com a fórmula:

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

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

Equação 4 Proporção de ancestralidade da população a na geração g $M_{a,g}$

É importante ressaltar que na ausência de fluxo migratório em uma determinada geração ($m_{i,g}$= 0) os valores de $M_{a,g}$ se mantém iguais aos valores de $M_{a,(g-1)}$. Além disso, para o modelo de Faria (2023), os valores de $m_{i,g}$ devem ser menores do que 1 , pois no caso de $m_{i,g} = 1$ ocorreria a substituição completa dos outros componentes de ancestralidade na população receptora naquela geração. Outro fator a ser observado é que para esta fórmula é possível a chegada de apenas uma população *i* por geração.

Comparando com o enfoque de Kehdy *et al.* (2015), Faria (2023) generalizou um modelo de miscigenação mais geral, para qualquer número de pulsos (no limite, cada pulso pode ser uma geração) e populações ancestrais, permitindo inferir como evoluem as ancestralidades cumulativas a partir das contribuições de cada população parental em cada pulso/geração. O modelo de Kehdy *et al.* (2015) é um caso específico do modelo teórico de Faria (2023) para pulsos específicos e com 3 populações parentais.

3.3 Como inferir a dinâmica da miscigenação a partir de dados genéticos

3.3.1 ABC: como inferir a distribuição a posteriori dos parâmetros $m_{i,g}$

O método de ABC foi escolhido pelo fato de os dados possuírem uma função de verossimilhança desconhecida. A partir da distribuição dos *tracts* de cada ancestralidade inferimos uma *posteriori* que contenha os parâmetros de miscigenação condicionados em um modelo da dinâmica de miscigenação construído com base na história das populações analisadas. Para isso seguimos os seguintes passos:

1. Estimação da *priori* informativa para cada pulso migratório. Cada pulso contém 3 parâmetros.
2. Simulação dos *tracts* de ancestralidades diferentes, com base nos parâmetros na *priori* informativa.
3. Cálculo da distância entre as distribuições de *tracts* observada e simulada.
4. Estimação da *posteriori* a partir dos parâmetros gerados nos cenários simulados com distribuições de *tracts* mais próximas das observadas.

3.3.1.1 Dados simulados

3.3.1.1.1 Distribuição a priori informativa dos parâmetros $m_{i,g}$

A princípio, é necessário que tenhamos valores iniciais de $m_{i,g}$. Para isso usamos uma distribuição de Dirichlet, que cria um número limitado de valores positivos cuja soma pode chegar até 1 (Tian, Tang e Ng 2013). Dessa forma, ao amostrar uma Dirichilet de ordem 4 60 milhões de vezes, obtivemos uma distribuição a priori não-informativa que representa aproximadamente 20 milhões de cenários de miscigenação para 3 pulsos migratórios, e que respeita os seguintes requisitos:

- A geração em que cada população parental entra no pulso de miscigenação é aleatória.
- Para os pulsos 1, 2 e 3 a soma dos três valores de m correspondentes às populações parentais deve ser menor ou igual a 1.

A partir dos valores de $m_{i,g}$ presentes na priori não-informativa, usamos a fórmula descrita por Faria 2023 (Fórmula 4) para calcular a proporção total de contribuição de cada população parental para formação daquele cenário ($M_{a,g}$).

Tratando da ancestralidade a nível continental ou tri-híbrida, teremos a geração de três valores $M_{a,g}$, cada um correspondente a uma ancestralidade parental. Para a priori informativa, selecionamos apenas aqueles cenários que gerem valores de $M_{a,g}$ próximos (± 5%) aos valores de ancestralidade populacional obtidos em nosso banco de dados alvo. Este valor corresponde à média das ancestralidades individuais para cada ancestralidade parental para os dados reais.

3.3.1.1.2 Simulação dos tracts de ancestralidade

Usando os valores de $m_{i,g}$ obtidos na priori informativa usamos o simulador de Liang-Nielsen (Liang e Nielsen 2014) para gerar *tracts* de ancestralidade. Acrescentamos um Pulso 0 na geração imediatamente anterior ao Pulso 1, com o valor de $m_{i,g}$ igual a 1 para ser correspondente à população fundadora. Os cromossomos são simulados de acordo com o número de indivíduos da população estudada. Dessa forma é possível comparar a distribuição dos tamanhos dos *tracts* de ancestralidade dos dados observados com os dados simulados.

3.3.1.1.3 Cálculo da distância entre as distribuições e aproximação da *posteriori*

Usamos a estatística do teste de Kolmogorov-Smirnov para comparar as distribuições simuladas e observadas. Ela consiste no cálculo da maior distância (D) entre duas distribuições de proporções acumuladas (Massey, 1951). Em nosso modelo possuímos três distribuições de *tracts* dependentes entre si obtidas de cada cenário, sendo cada uma correspondente a uma ancestralidade. O mesmo acontece para os *tracts* simulados, portanto, comparamos as distribuições duas a duas, gerando três valores de D diferentes por cenário. Para que a aproximação da distribuição a *posteriori* aconteça de forma a considerar as três comparações de forma conjunta criamos três vetores com os valores de D para todos os cenários avaliados, sendo cada um correspondente a uma ancestralidade. Cada vetor foi ordenado de forma crescente. Selecionamos 1% dos cenários que possuem menor distância em relação às distribuições observadas.

Assim selecionamos os cenários que formam distribuições de *tracts* mais próximas das distribuições observadas, montando uma aproximação da distribuição a posteriori.

3.3.1.2 Dados Observados

3.3.1.2.1 Controle de qualidade

O banco de dados do LDGH foi limpo usando o software *MosaiQC* (https://github.com/ldgh/Smart-cleaning-public/), 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. Ele foi aplicado a cada população separadamente, sendo que a junção foi feita para iniciar as análises de ancestralidade. Durante a limpeza, após a inferência do sexo, são removidos indivíduos de sexo incerto, variantes que apresentem cromossomo 0, variantes com 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

A e T ou C e G. Além disso, durante o controle de qualidade foi adicionado o identificador das variantes de acordo com o DBSNP.

3.3.1.2.2 Banco de dados de populações miscigenadas

Inicialmente, geramos um banco de dados com **11 012** indivíduos e 1 597 279 SNVs, sendo 7 421 indivíduos miscigenados e 3 591 de populações parentais.

Tabela 2 Descrição das populações do banco de dados de 1.5M

Origem Continental	População	Fonte	Número de Indivíduos (Autossômicos)	Número de Indivíduos (X)
Europa	GBR	1000 Genomes	91	52
	Orcadian	HGDP	13	13
	IBS	1000 Genomes	107	66
	TSI	1000 Genomes	107	66
	Sardinian	HGDP	24	24
	Bergamo	HGDP	10	9
	Tuscan	HGDP	6	6
	CEU	1000 Genomes	99	60
	Basque	HGDP	21	20
	French	HGDP	24	24
África	South Africa	HGDP	4	4
	Botswana	Crawford *et al.* 2017	140	42
	GWD	1000 Genomes	113	58
	Gahna	Ghana Prostate Study (NCI-NIH) - (Chokkalingam *et al*. 2012)	968	0
	Kenya	HGDP	9	9
	LWK	1000 Genomes	99	54

	Namíbia	HGDP	2	0
	ESN	1000 Genomes	99	45
	Yoruba	HGDP	18	18
	Biaka	HGDP	20	20
	Mbuti	HGDP	8	8
	Mandenka	HGDP	18	18
	Tanzania	Crawford *et al.* 2017	43	14
	MSL	1000 Genomes	85	43
	Uganda1	EMBLEM (https://emblem.cancer.gov/)	636	344
	Uganda2	EMBLEM (https://emblem.cancer.gov/)	159	85
	YRI	1000 Genomes	108	56
	ASW	1000 Genomes	62	34
	ACB	1000 Genomes	96	49
Nativos Americanos	Peruvian Natives	LDGH/INS (Borda *et al*. 2020)	358	274
	Karitiana	HGDP	8	8
	Surui	HGDP	6	6
	Maya	HGDP	19	19
	Pima	HGDP	11	11
Miscigenados	Salvador	EPIGEN-Brasil (Kehdy *et al*. 2015)	1309	1308
	Bambuí	EPIGEN-Brasil (Kehdy *et al*. 2015)	1442	1442
	Pelotas	EPIGEN-Brasil (Kehdy *et al.* 2015)	3736	3735
	Miscigenados Peruanos	INS	391	296

	Venezuela		543	268
Total:			**11012**	**8608**

Para o as análises do cromossomo X usamos os dados das populações que possuem informação para esse cromossomo. Foram um total de 8608 indivíduos, sendo 4303 mulheres e 4305 homens. As análises foram feitas para homens e mulheres separadamente. A fim de comparação com os resultados de Kehdy *et al.* 2015, aplicamos o método ABC apenas nas coortes de Bambuí, Salvador e Pelotas, que fazem parte do projeto EPIGEN-Brasil.

3.3.1.2.3 Inferência de Ancestralidade Individual e Populacional

As inferências de ancestralidade individual e populacional foram feitas usando o software ADMIXTURE (Alexander, Novembre and Lange 2009). No presente caso, como estamos interessados na miscigenação continental indígena, africana e europeia, definimos K = 3 populações parentais. Para o cromossomo X usamos o software com a flag --haploid, desenvolvida para tratar especificamente deste cromossomo.

Para garantir independência dos SNVs, como requerido por ADMIXTURE, esse programa foi executado após o filtro de desequilíbrio de ligação que excluiu todos os SNVs com $r^2>0,4$, restando um total de 720 438 SNVs nos cromossomos autossômicos. 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 análise de ADMIXTURE requer uma ampla capacidade de processamento. Para realizá-la em tempo viável, a mesma foi executada um máquina virtual do tipo EC2 na nuvem da Amazon AWS (*Amazon Web Services*). Para isso usamos uma máquina do tipo c5a.24xlarge, que conta com 96 processadores. Ela foi montada em linguagem Linux, com sistema operacional CentOS. Para fazer as análises relacionadas à genética de populações instalamos o Plink (Purcell *et al.* 2007; versão: 1.9) e o ADMIXTURE (versão: 1.3.0). A análise ocupou as 96 CPUs. Escrevemos junto com a AWS um texto para o blog da empresa sobre nossa experiência do uso da nuvem para estudos de bioinformática (Tarazona-Santos *et al*. 2024, https://aws.amazon.com/pt/blogs/aws-brasil/a-computacao-na-nuvem-da-aws-impulsionando-avancos-em-bioinformatica/)

3.3.1.2.4 Inferência de Ancestralidade Local

As inferências de ancestralidade local são feitas a partir do software RFMix1 (Maples *et al.* 2013). As recomendações de uso do software sugerem que o número de indivíduos de cada população parental esteja balanceado. Assim, selecionamos 160 indivíduos africanos (17 ESN, 43 Gahna, 46 YRI e 54 Botswana), 167 indivíduos europeus (24 Sardinian, 60 TSI, e 83 IBS) e 156 Nativos Americanos (*Peruvian Genome Project*) que se apresentaram mais extremos na relação PC1 versus PC2 (Figura 18). Por sermos cientes que a escolha dos indivíduos mais extremos nos primeiros PCs pode gerar um viés (porque essa estratégia subestima a diversidade das populações parentais), as populações escolhidas para representar cada ancestralidade continental foram escolhidas com base na história da miscigenação da população brasileira.

Todas as populações de origem destes indivíduos representam componentes de miscigenação em indivíduos miscigenados latino-americanos (Kehdy *et al.* 2015; Gouveia *et al.* 2020; Borda *et al.* 2020).

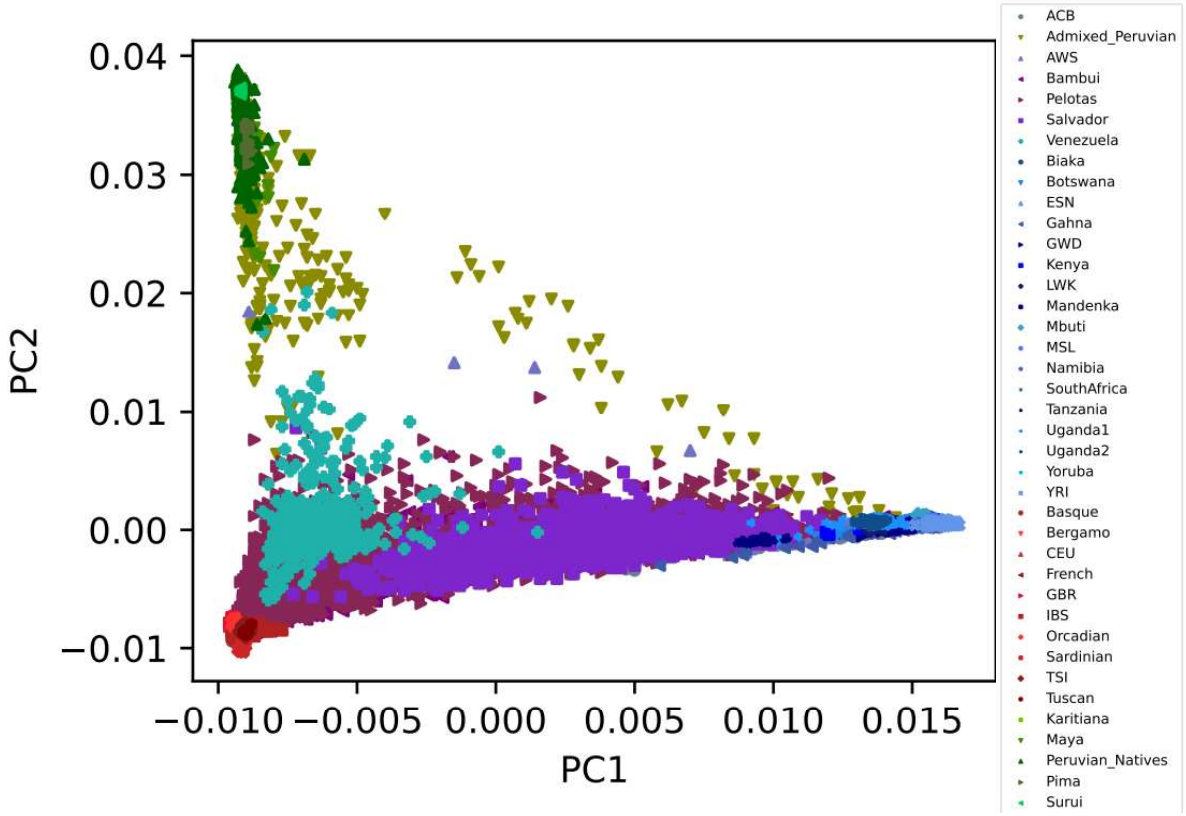


Figura 18 Análise de componentes principais obtida a partir do banco descrito na Tabela 2. PCA utilizado para seleção das amostras referência na inferência de ancestralidade local. No eixo X se encontra o PC1 e no eixo Y o PC2.

O método foi corrido para 6.878 indivíduos miscigenados para um total de 1.662.516 variantes. Os dados foram haplotipados usando o software Shapeit4 (Delaneau *et al.*, 2019), usado o banco de dados 1000 *Genomes Project* (1KGP-https://www.internationalgenome.org/) fase 3 como referência. Para o RFMix1 usamos a opção PopPhased para correção de fase e corremos 2 iterações de *Expectation Maximisation* (-e 2). Os resultados foram gerados por janela, sendo cada janela de 0.2 cM. O restante das opções usamos o padrão do software.

3.3.2 Implementação do método

O método foi implementado em python3 e R. O script foi separado em módulos para melhor organização e reparo de bugs (https://github.com/ldgh/ABC/tree/main/ABC_2015). Os módulos cumprem as etapas de I) Amostragem da Dirichlet; II) Comparação dos $M_{a,g}$ de cada cenário com a ancestralidade global da população alvo; III) Simulação dos *tracts* de ancestralidade; IV) Geração das distribuições de *tracts* simuladas; V) Cálculo do KS entre as distribuições de *tracts* observadas e simuladas; VI) Ordenação e seleção de 1% dos cenários melhor posicionados (Figura 19).

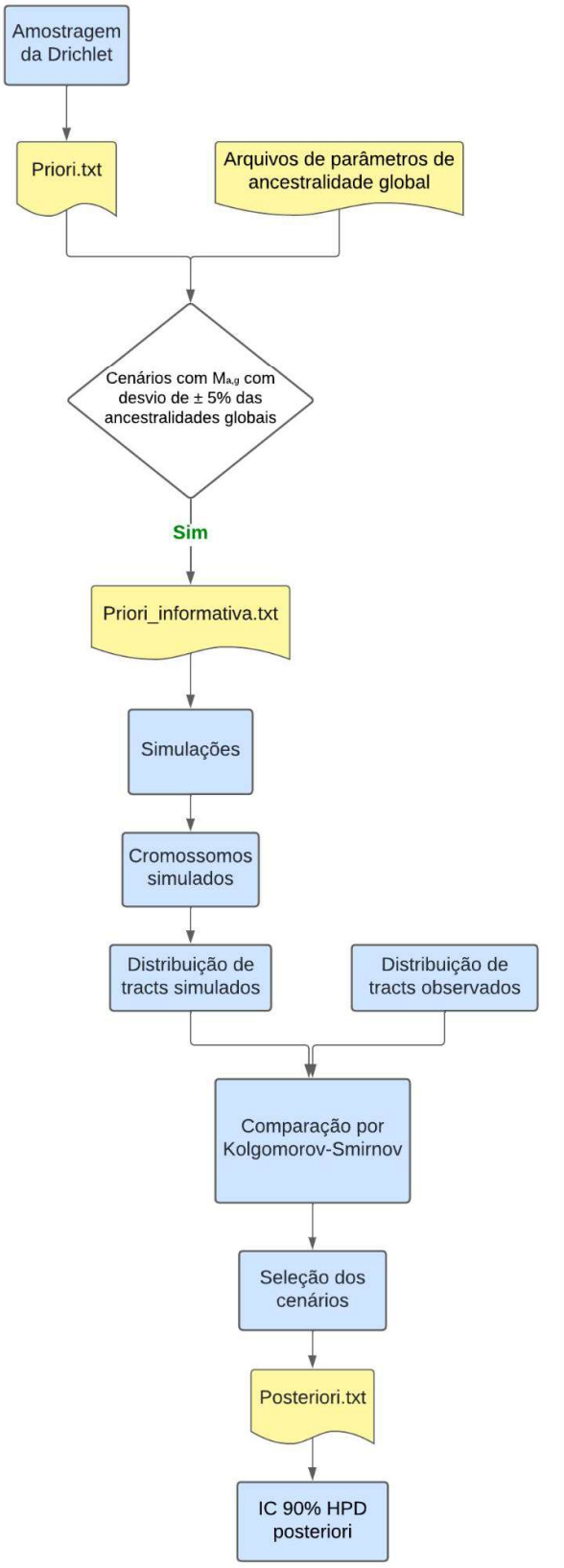


Figura 19 Fluxograma da metodologia ABC.

3.4 Resultados

3.4.1 Ancestralidade Global e Distribuição dos *tracts* de ancestralidade local nas populações alvo

Após a execução do ADMIXTURE, podemos observar que as proporções médias de ancestralidade Europeia, Nativo-Americana e Africana das coortes do projeto EPIGEN Brasil aparecem coerentes em relação ao observado em Kehdy *et al.* 2015 (Figura 5). A distribuição das ancestralidades Nativo-Americana é similar nas três coortes, enquanto as distribuições da ancestralidade Europeia e Africana sofrem variação apenas na coorte de Salvador (Figura 20-A e B). Essa observação é coerente com a maior proporção de ancestralidade Africana na coorte de Salvador.

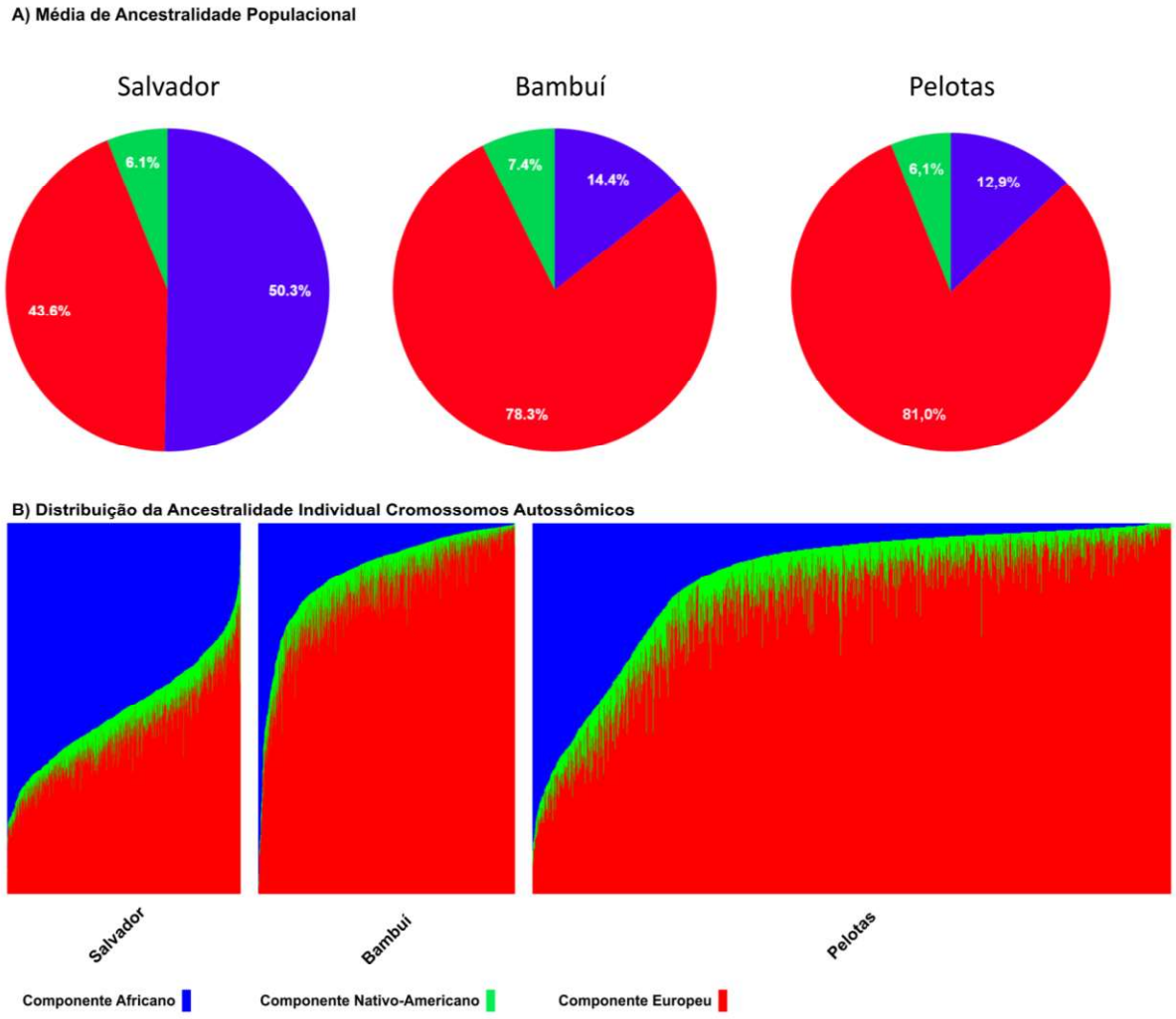


Figura 20 Ancestralidade populacional e individual nas populações miscigenadas Latino-Americanas. Em azul temos o componente de miscigenação Africano; em verde o componente de miscigenação Nativo-americano e em vermelho o componente de miscigenação Europeu. A) Em gráficos de pizza separados

vemos a proporção do componente de cada parental para formação das populações miscigenadas nos cromossomos autossômico. B) A distribuição das ancestralidades individuais calculadas com base nos cromossomos autossômicos.

Como mencionado anteriormente, as informações de ancestralidade local podem agir como um marcador temporal. Para extrair essa informação do banco de dados, nós distribuímos os *tracts* de cada ancestralidade em relação a contagem por tamanho (Figura 21). Para escalabilidade computacional e evitar simulações desnecessárias, selecionamos quatro cromossomos que juntos formam uma distribuição de *tracts* similar àquela obtida ao analisar todos os cromossomos simultaneamente (Figura 21 -A). Para isso, selecionamos os cromossomos 1, 2, 15 e 21 (Figura 21 - B).

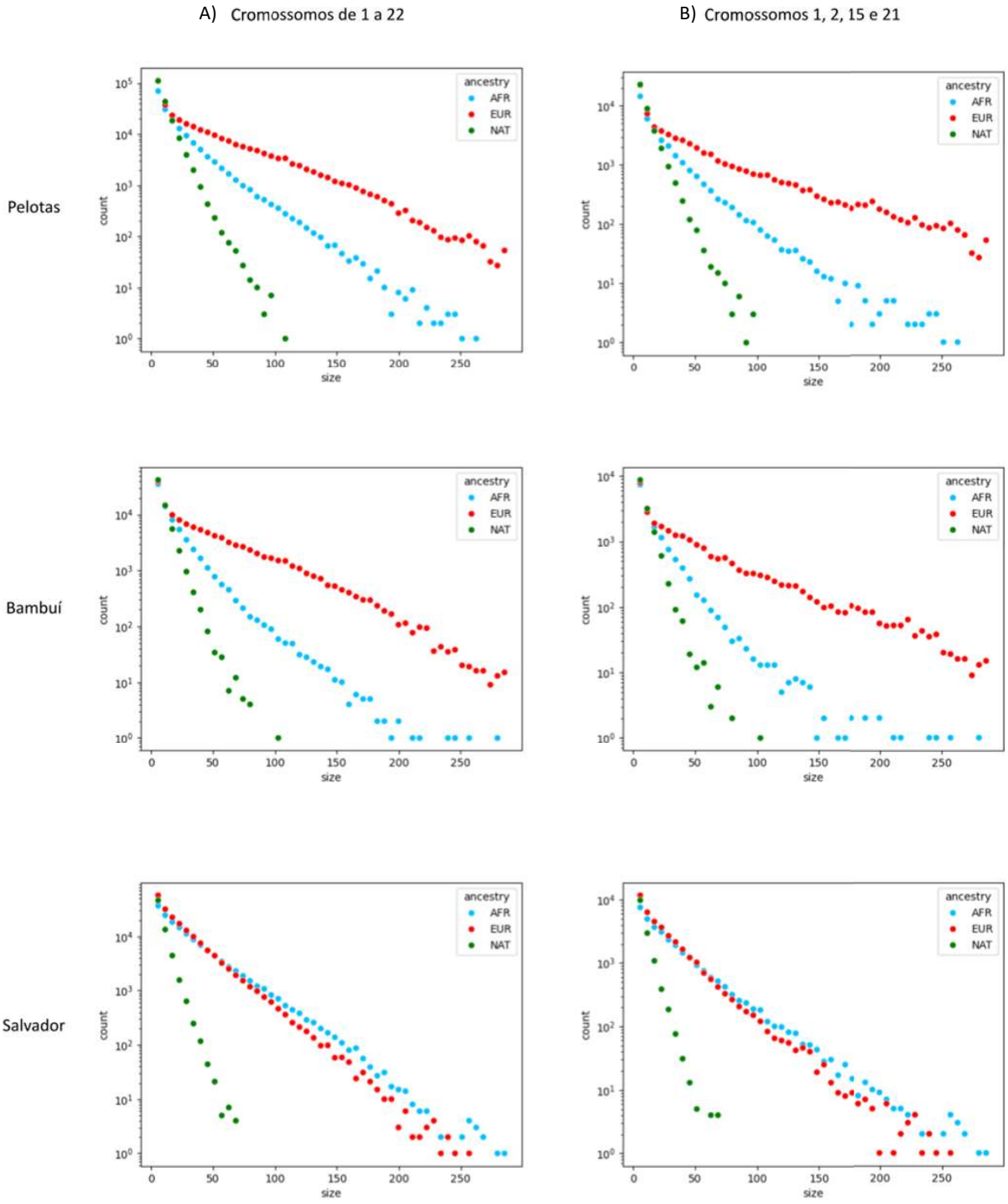


Figura 21 Distribuições de *tracts* de ancestralidade para as coortes de Pelotas, Salvador e Bambuí. A) Distribuições dos *tracts* de ancestralidade obtidos de todos os cromossomos autossômicos. B) Distribuições de *tracts* obtidas a partir dos cromossomos 1, 2, 15 e 21.

3.4.2 Parâmetros para simulação dos cenários

Partindo da priori não informativa, após a filtragem dos cenários compatíveis à ancestralidade global das populações alvo, foi possível selecionar 68 073 cenários para Salvador (0,3% dos cenários da priori não-informativa), 128 485 cenários para Bambuí (0,6% da priori não informativa) e 155 043 cenários para Pelotas (0,8% da priori não informativa).

Após a filtragem de 1% usando o KS ordenado restaram 1551 cenários para Pelotas (Figura 22 - A), 1 285 para Bambuí (Figura 23 - A) e 669 para Salvador (Figura 24 - A). Para as três coortes obtivemos os intervalos de credibilidade 90% HPD. A Tabela 3 apesenta os intervalos de credibilidade HPD para os três pulsos e as três coortes. No primeiro pulso, a menor amplitude alcançada para a contribuição europeia foi para a coorte de Pelotas (0,6247), para a contribuição africana foi a coorte de Bambui (0,5837) e para a contribuição Nativo-Americana a coorte de Salvador (0,3950). No segundo pulso, a coorte de Pelotas obteve novamente uma menor amplitude para o intervalo de contribuição europeia (0,7204), e também apresentou menor amplitude para o intervalo de contribuição da população africana (0,5542). A contribuição Nativo-Americana novamente apresentou menor intervalo na coorte de Salvador (0,1977). No terceiro pulso, as amplitudes ficam menores. A menor amplitude para contribuição europeia (0,3377) e africana (0,2150) foi na coorte de Bambui, sendo que apenas a contribuição Nativo-Americana apresentou menor intervalo na coorte de Salvador (0,0985).

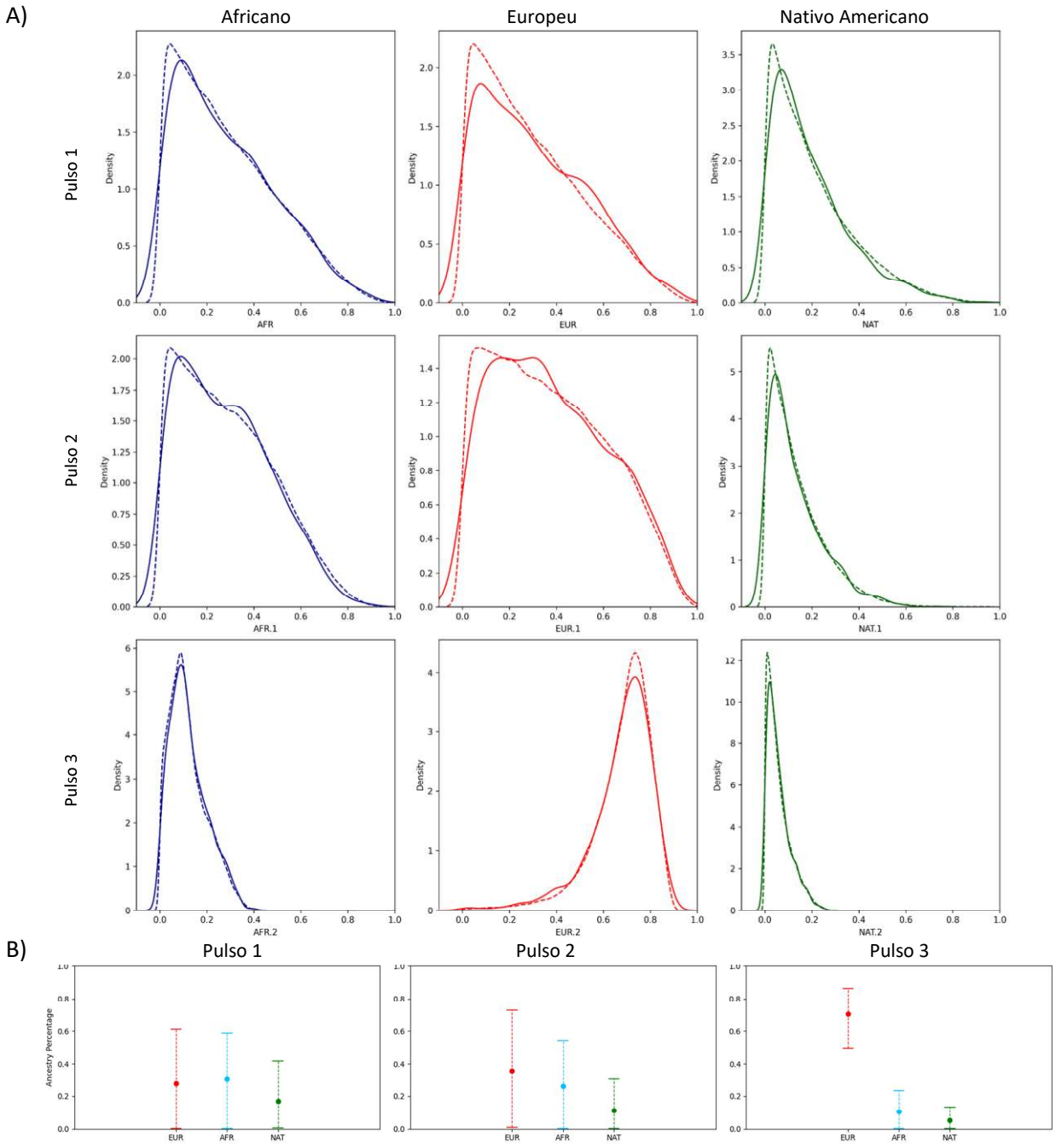


Figura 22 Distribuições dos parâmetros $m_{i,g}$ na população de Pelotas nos três pulsos de miscigenação. A) As linhas pontilhadas representam as distribuições a priori e as linhas contínuas representam as distribuições a posteriori. As cores azul, vermelho e verde correspondem às ancestralidades Africana, Europeia e Nativo-Americanos, respectivamente. B) Intervalos de credibilidade 90% HPD das distribuições a posteriori.

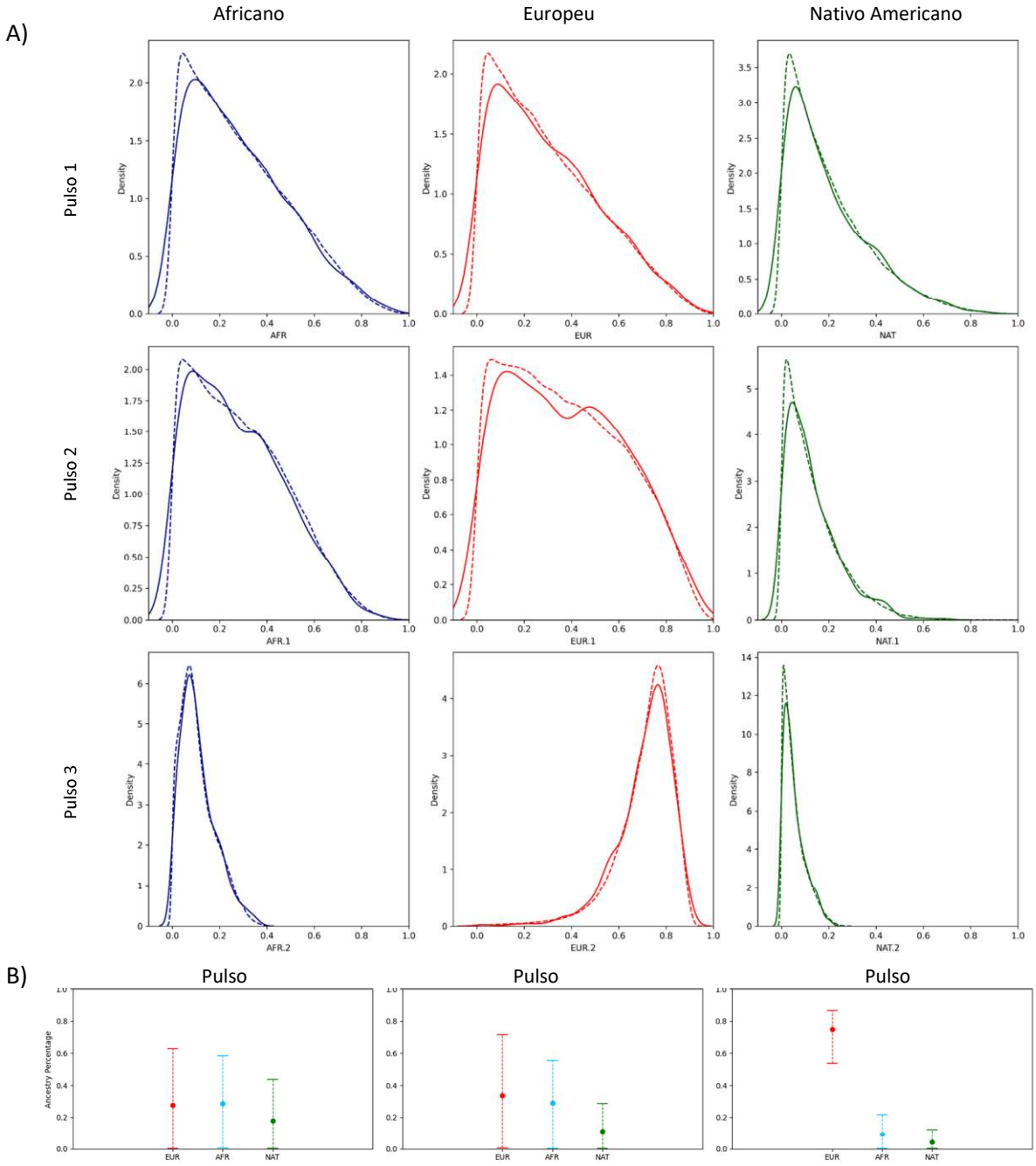


Figura 23 Distribuições dos parâmetros $m_{i,g}$ na população de Bambui nos três pulsos de miscigenação. A) As linhas pontilhadas representam as distribuições a priori e as linhas contínuas representam as distribuições a posteriori. As cores azul, vermelho e verde correspondem às ancestralidades Africana, Europeia e Nativo-Americanos, respectivamente. B) Intervalos de credibilidade 90% HPD das distribuições a posteriori.

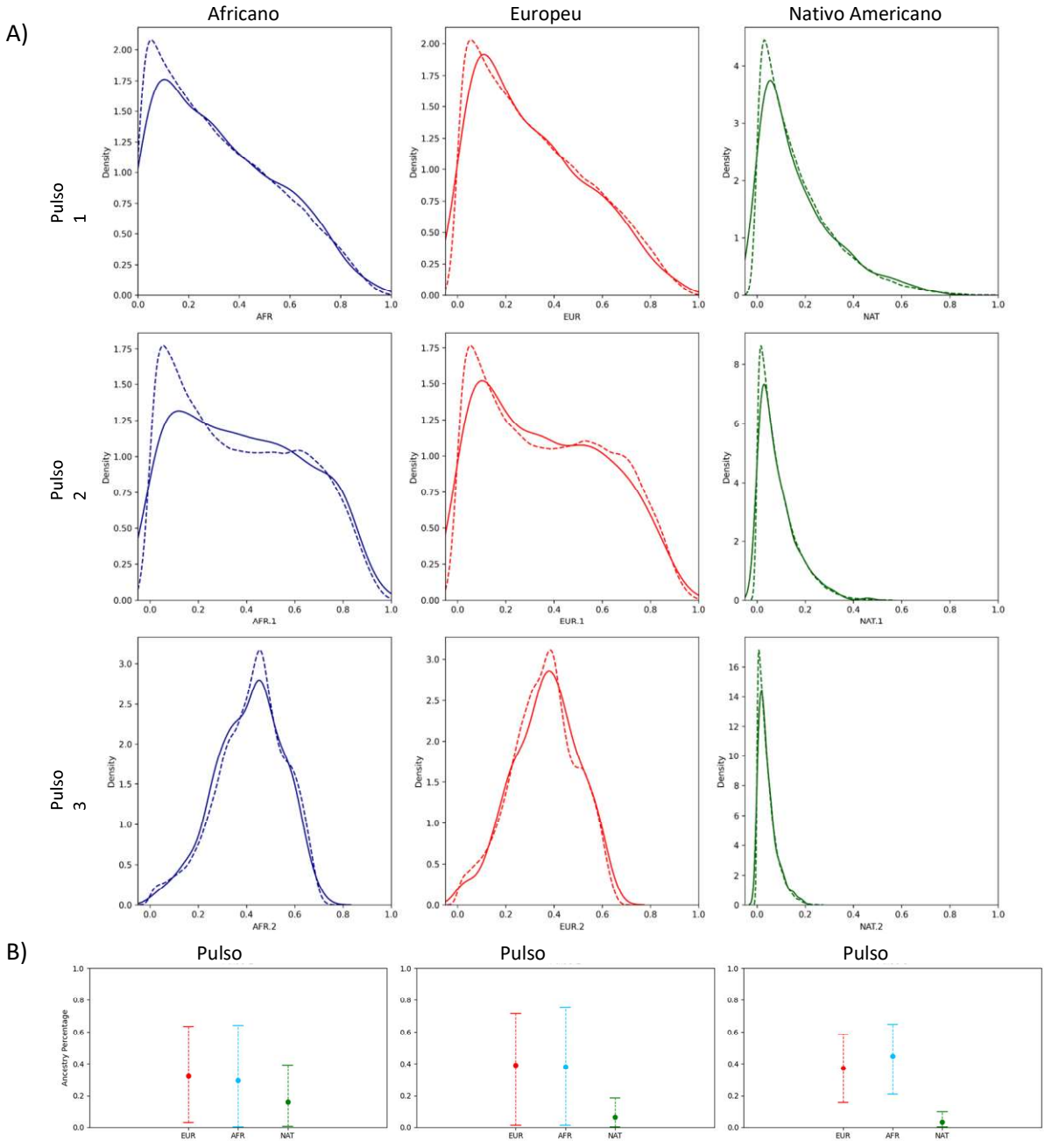


Figura 24 Distribuições dos parâmetros $m_{i,g}$ na população de Salvador nos três pulsos de miscigenação. A) As linhas pontilhadas representam as distribuições a priori e as linhas contínuas representam as distribuições a posteriori. As cores azul, vermelho e verde correspondem às ancestralidades Africana, Europeia e Nativo-Americanos, respectivamente. B) Intervalos de credibilidade 90% HPD das distribuições a posteriori.

Tabela 3 Intervalos de Credibilidade 90% HPD das distribuições a posteriori e suas amplitudes para as coorte de Bambuí, Salvador e Pelotas.

		Pulso 1		Pulso 2		Pulso 3	
	Ancestralidade	**Intervalo**	**Amplitude**	**Intervalo**	**Amplitude**	**Intervalo**	**Amplitude**
Bambuí	**Africana**	[0,0003-0,584]	0,5837	[0,0002-0,5572]	0,5570	[0,0004-0,2154]	0,2150
	Europeia	[0,0003-0,6277]	0,6274	[0,0038-0,7286]	0,7248	[0,5301-0,8678]	0,3377
	Nativo-Americana	[0,0002-0,4437]	0,4435	[0,0001- 0,2888]	0,2887	[0,0001-0,1235]	0,1234
Pelotas	**Africana**	[0,0003-0,5941]	0,5938	[0,0004-0,5546]	0,5542	[0,0002-0,2372]	0,2370
	Europeia	[0,0002-0,6249]	0,6247	[0,0071-0,7275]	0,7204	[0,4933-0,8656]	0,3723
	Nativo-Americana	[0.0007-0.4284]	0,4277	[0,000005-0,3036]	0,303695	[0,00004-0,1329]	0,1325
Salvador	**Africana**	[0,0001-0,6533]	0,6532	[0,0002-0,7237]	0,7235	[0,1464-0,5881]	0,4417
	Europeia	[0,0002-0,6668]	0,6666	[0,0001-0,7648]	0,7647	[0,2060-0,6478]	0,4418
	Nativo-Americana	[0,0001-0,3951]	0,3950	[0,0002-0,1979]	0,1977	[0,0001-0,0986]	0,0985

3.5 Discussão

O trabalho de Kehdy *et al.* 2015 descreve uma forma de recuperar informações sobre a dinâmica de migrações na formação de populações miscigenadas através dos dados de ancestralidade local. Retomar discução das limitações de Kehdy *et al.* 2015. Nós reimplementamos o método de acordo com as definições estatísticas de Kehdy *et al.* 2015, porém usamos a linguagem python e usamos o sistema de módulos, o que permitirá que nos próximos passos de desenvolvimento do projeto cada módulo seja evoluido de forma independente, diminuindo assim a chance de inserção de erros em módulos já funcionais.

Apesar disso, não obtivemos diferenças significativas entre as distribuições a *priori* informativas e a *posteriori*, para nenhum dos pulsos ou coorte. Dessa forma, a maneira como o método está implementado no presente momento, não foi possível extrair a informação de tempo das distribuições dos *tracts* de ancestralidade local, o que diverge dos resultados apresentados em Kehdy *et al.* 2015. Entre os fatores que podem ter levado a isso podemos

incluir as divergências entre os métodos utilizados nas análises do artigo, uma vez que as versões dos softwares utilizados não era a mesma disponível hoje. Como exemplo, na presente tese usamos para inferência de haplótipos o software Shapeit na versão 4, enquanto para Kehdy *et al.* foi usada a primeira versão (Delaneau, Marchini e Zagury 2012). Os artigos dos métodos apontam diversas melhorias em relação à qualidade da inferência de fase entre uma versão e outra, o que pode ter levado a um número maior de *tracts* menores de ancestralidade em 2015 do que agora. Isso se repete com o software de inferência de ancestralidade local, onde no presente trabalho foi usado o software RFMix 1, que possui módulo de correção de inferência de fase, e no artigo o software usado foi o PCAdmix (Brisbin *et al.* 2012). Além disso, usamos cromossomos maiores que os usados por Kehdy *et al.* 2015, que possuem diferênças nos padrões de recombinação.

Além das limitações apontadas para não replicação do resultado de Kehdy *et al.* 2015, também precisamos considerar as limitações da ancestralidade local usada isoladamente, uma vez que pelo acaso *tracts* diferentes da mesma ancestralidade podem se encontrar em um fragmento cromossômico, provocando uma falsa impressão de entrada recente na população, mesmo que o *tract* possua origens IBDs diferentes.

O simulador de Liang-Nielsen (Liang e Nielsen 2014), apesar de mais rápido, também possui grandes limitações. Entre elas está o fato de considerar as populações uma a uma em cada geração, impossibilitando a interação de indivíduos de outras fontes ancestrais. Além disso, o simulador disconsidera as posições de hot spot de recombinação, uma vez que usa apenas o tamanho do cromossomo em Morgan para definir a probabilidade de recombinação geral do cromossomo em cada geração. Além disso, ele seria limitado ao implementar qualquer estratégia que separem os *tracts* de acnestralidade por IBD. Estes pontos levam à discussões sobre o quanto o uso de outro simulador poderia aumentar a acurácia do método, porém é necessário que as simulações não inviabilizem o método pelo tempo computacional necessário para que as mesmas sejam feitas.

3.6 Perspectivas

O método ABC ainda precisa ser bastante trabalhado para que cumpra eficientemente seu objetivo. Cada módulo do script precisa passar por revisão e validação com dados simulados. Além disso, estamos estudando o uso de outras estratégias estatísticas como MCMC ou *Espectation Maximization* para otimização do método e redução dos intervalos de credibilide

90% HPD. Outro ponto onde prevemos aplicar melhorias é nas simulações dos *tracts* de ancestralidade. Tendo em vista as limitações do simulador de Liang-Nielsen (Liang e Nielsen 2014), planejamos testar outros simuladores.

Ao avaliar resultados de ancestralidade individual e populacional inferidos a partir do cromossomo X (Figura 24) se ressalta a presença do viés sexual, ou seja, as proporções de ancestralidade Nativo-Americana e Africanas estão maiores que quando comparadas aos cromossomos autossômicos. Além de mostrar a violência do processo de colonização pós colombiana, a história do viés sexual se modifica ao longo deste processo, como mostrado por registros as diferenças nas proporções migratórias entre homens e mulheres europeias nos dois séculos iniciais da colonização, onde as migrações europeias eram constituídas majoritariamente por homens (Priore, 2004).

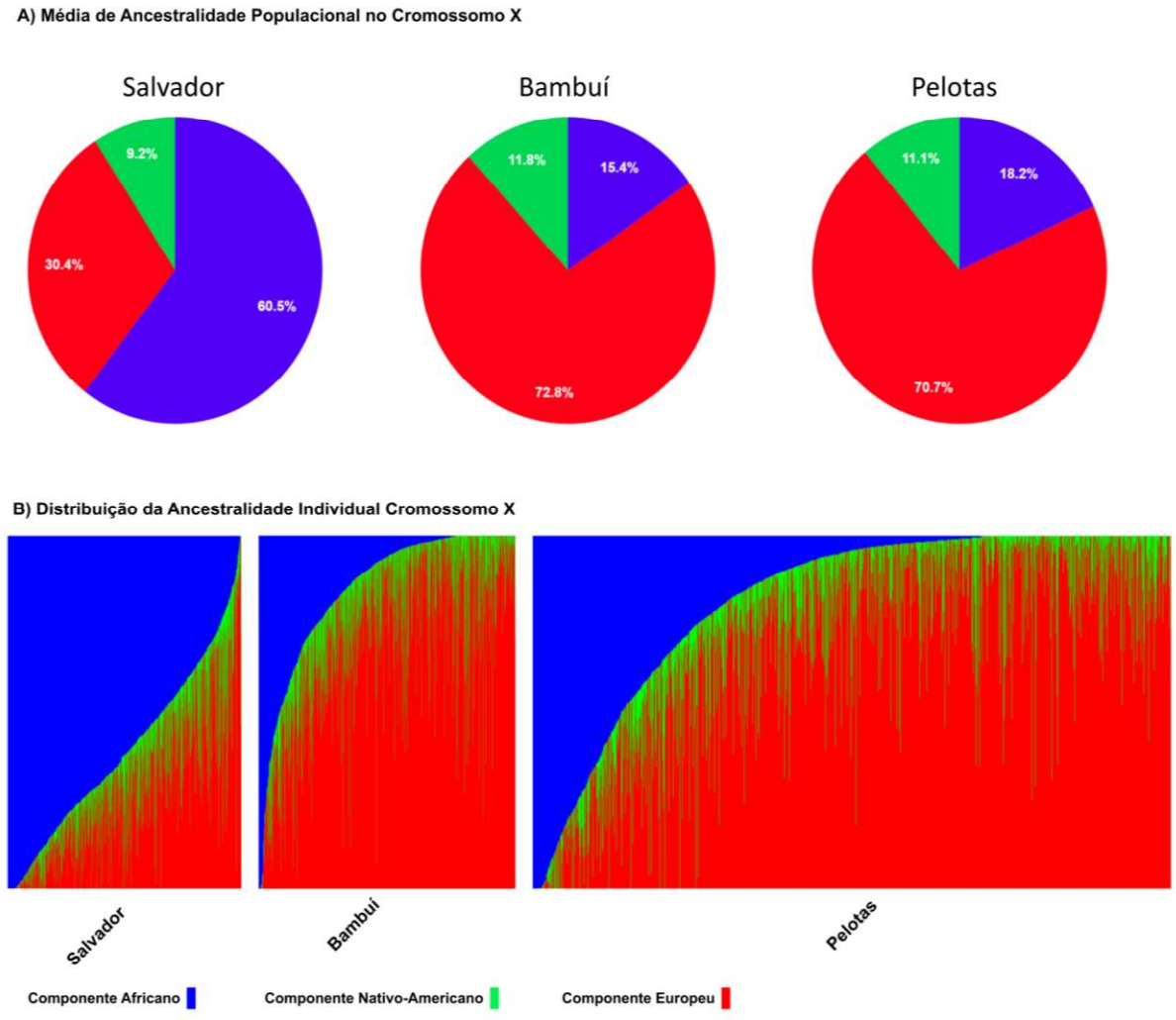


Figura 25 Ancestralidade populacional e individual nas populações miscigenadas Brasileiras inferidas com base no cromossomo X. Em azul temos o componente de miscigenação Africano; em verde o componente de miscigenação Nativo-americano e em vermelho o componente de miscigenação Europeu. A) Em

gráficos de pizza separados vemos a proporção do componente de cada parental para formação das populações miscigenadas no cromossomo X. B) A distribuição das ancestralidades individuais calculadas com base no cromossomo X.

Tendo em vista a possibilidade de inferência da dinâmica do viés sexual, nosso grupo está trabalhando na incorporação do uso de dados do cromossomo X, que tendo um mecanismo de herança diferente para homens e mulheres, permitiria discriminar entre fluxo gênico mediado por homens e por mulheres, mas essa incorporação está ainda em andamento. Em XXXX, a Dra Hanaisa Sant'Anna desenvolveu em seu doutorado uma estratégia baseada no desvio HWE do cromossomo X para determinar e quantificar o viés sexual ao longo das gerações de miscigenação (Sant'Anna 2020), método que passa atualmente por validação estatística. Para enriquecer o método de Sant'Anna (2020), pretendemos incorporar as informações de ancestralidade local do cromossomo X. Se estratificamos o fluxo gênico por sexo, teríamos 18 parâmetros no lugar de 9 (Figura 26).

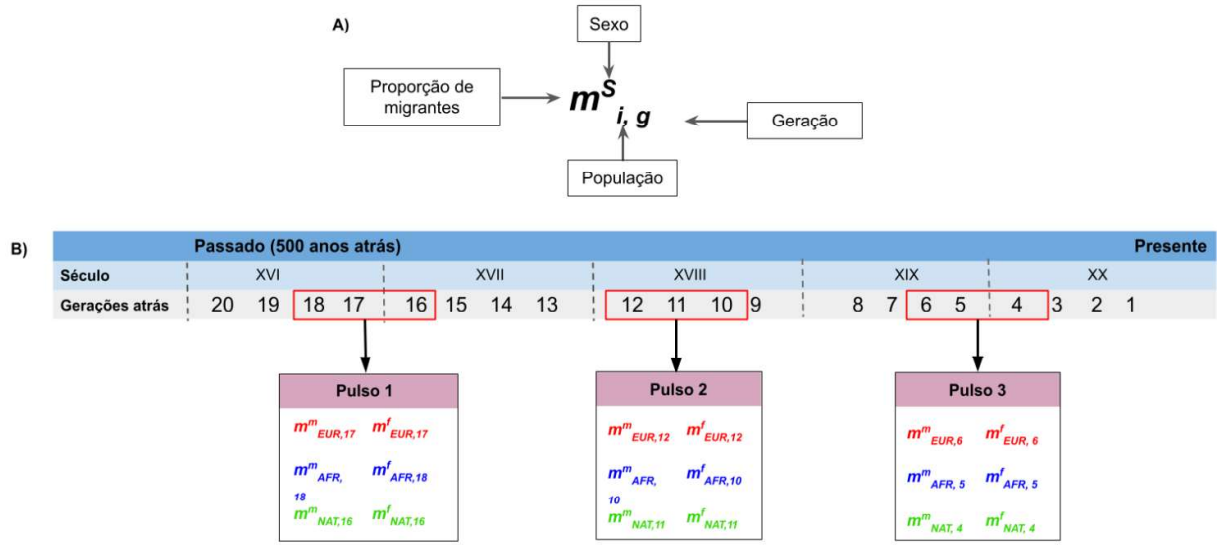


Figura 26 Modelo demográfico de miscigenação pós-colombiana e identificação gráfica dos parâmetros m de migração pelo cromossomo X. A) Características dos parâmetros $m_{i,g}$ no método aplicado no cromossomo X. B) Definição dos pulsos de migração e ao longo do tempo e dos parâmetros $m^s_{i,g}$ associados a cada pulso na aplicação do método no cromossomo X.

REFERÊNCIAS

Alexander, David H., John Novembre, and Kenneth Lange. 2009. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” Genome Research 19 (9): 1655–64.

Barreto TMM, Souza RS, São Pedro RB, *et al.* Rare genetic variants of NLRP12 in Admixed Latino-American Children with SARS-CoV-2-related Multisystem Inflammatory Syndrome. *J Infect Dis*. Published online September 27, 2024. doi:10.1093/infdis/jiae480

Borda V, Alvim I, Mendes M, Silva-Carvalho C, *et al*. 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;117(51):32557-32565.doi:10.1073/pnas.2013773117

Brisbin A, *et al*. PCAdmix: Principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum Biol. 2012;84(4):343–364. doi: 10.3378/027.084.0401.

Browning, S., Browning, B. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* **12**, 703–714 (2011). https://doi.org/10.1038/nrg3054

Browning, S. R.; Browning, B. L. Identity by Descent Between Distant Relatives: Detection and Applications. Annual Review of Genetics, v. 46, n. 1, p. 617–633, 15 dez. 2012.

Browning, B. L.; Browning, S. R. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. Genetics, v. 194, n. 2, p. 459–471, 27 mar. 2013.

Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. *Nature*. 2011;475(7355):163-165. Published 2011 Jul 13. doi:10.1038/475163a

Campbell, Lyle. 2000. American Indian Languages: The Historical Linguistics of Native America. Oxford University Press

Caro-Consuegra R, Nieves-Colón MA, Rawls E, *et al.* Uncovering Signals of Positive Selection in Peruvian Populations from Three Ecological Regions. *Mol Biol Evol*. 2022;39(8):msac158. doi:10.1093/molbev/msac158

Conomos, Matthew P. *et al.* Model-free Estimation of Recent Genetic Relatedness. The American Journal of Human Genetics, v. 98, n. 1, p. 127–148, jan. 2016.

Corpas M, Siddiqui MK, Soremekun O, Mathur R, Gill D, Fatumo S. Addressing Ancestry and Sex Bias in Pharmacogenomics. *Annu Rev Pharmacol Toxicol*. 2024;64:53-64. doi:10.1146/annurev-pharmtox-030823-111731

Couto-Silva CM, Nunes K, Venturini G, *et al*. Indigenous people from Amazon show genetic signatures of pathogen-driven selection. *Sci Adv*. 2023;9(10):eabo0234. doi:10.1126/sciadv.abo0234

Danecek, P. *et al.* The variant call format and VCFtools. Bioinformatics (Oxford, England), v. 27, n. 15, p. 2156–8, 2011.

Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2012;9(2):179–181. doi: 10.1038/nmeth.1785.

Delaneau, O. et al. Accurate, scalable and integrative haplotype estimation. Nature Communications, v. 10, n. 1, 28 nov. 2019.

De Marino A, Mahmoud AA, Bose M, Bircan KO, Terpolovsky A, Bamunusinghe V, Bohn S, Khan U, Novković B, Yazdi PG. A comparative analysis of current phasing and imputation software. PLoS One. 2022 Oct 19;17(10):e0260177. doi: 10.1371/journal.pone.0260177. PMID: 36260643; PMCID: PMC9581364.

de Ulloa, A. (1772). Noticias Americanas: Entretenimientos phisicos-historicos, sobre la América Meridional, y la Septentrional Oriental : Comparación general de los territorios, climas y produciones.

Durand, Eric Y., Nick Patterson, David Reich, and Montgomery Slatkin. 2011. “Testing for Ancient Admixture between Closely Related Populations.” Molecular Biology and Evolution 28 (8): 2239–52.

FARIA, Lucas F. Costa. Ancestralidade e Miscigenação: Estimativas e Aspectos Computacionais. 2023. 82 p. Dissertação (Doutorado em Bioinformática) - Programa Interunidades de Pós-Graduação em Bioinformática, Belo Horizonte, 2023.

FREYMAN, W. A. *et al.* Fast and Robust Identity-by-Descent Inference with the Templated Positional Burrows–Wheeler Transform. Molecular Biology and Evolution, v. 38, n. 5, p. 2131–2151, 23 dez. 2020.

Gouveia MH, Borda V, Leal TP, *et al*. Origins, Admixture Dynamics, and Homogenization of the African Gene Pool in the Americas. *Mol Biol Evol*. 2020;37(6):1647-1656. doi:10.1093/molbev/msaa033

Gouy A, Excoffier L. Polygenic Patterns of Adaptive Introgression in Modern Humans Are Mainly Shaped by Response to Pathogens. *Mol Biol Evol*. 2020;37(5):1420-1433. doi:10.1093/molbev/msz306

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, *et al*. 2010. "A Draft Sequence of the Neandertal Genome." Science 328 (5979): 710– 22

Guimarães Alves AC, Sukow NM, Adelman Cipolla G, et al. Tracing the Distribution of European Lactase Persistence Genotypes Along the Americas. *Front Genet*. 2021;12:671079. Published 2021 Sep 22. doi:10.3389/fgene.2021.671079

Guio, H. *et al.* (2024) *The Peruvian Genome Project: Expanding the Global Pool of Genome Diversity from South America* [Preprint]. doi:10.1101/2024.05.05.24306840.

Harris, Daniel N., Wei Song, Amol C. Shetty, Kelly S. Levano, Omar Cáceres, Carlos Padilla, Víctor Borda, *et al.* 2018. "Evolutionary Genomic Dynamics of Peruvians Before, During, and after the Inca Empire." Proceedings of the National Academy of Sciences of the United States of America, June, 201720798.

Hedrick, P. W. Genetics of populations. Boston: Jones And Bartlett Publishers, 2005.

Helgi Hilmarsson *et al.* High Resolution Ancestry Deconvolution for Next Generation Genomic Data. bioRxiv (Cold Spring Harbor Laboratory), 21 set. 2021.

Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine, v. 12, n. 3, p. e1001779, 31 mar. 2015.

Hrdlička A. (1928) The Origin and Antiquity of Men in America. Bull. N. Y. Acad. Med. 4, 802 .

Jacovas VC, Couto-Silva CM, Nunes K, *et al.* Selection scan reveals three new loci related to high altitude adaptation in Native Andeans. *Sci Rep*. 2018;8(1):12733. Published 2018 Aug

24. doi:10.1038/s41598-018-31100-6

Jaramillo-Valverde L. *The effect of altitude on the expression of immune-related genes in Peruvian rural indigenous*. Submetido.

JIN, Y. *et al*. Quickly identifying identical and closely related subjects in large databases using genotype data. PLOS ONE, v. 12, n. 6, p. e0179106, 13 jun. 2017.

Johri, Parul, *et al*. "Recommendations for Improving Statistical Inference in Population Genomics." *PLOS Biology*, vol. 20, no. 5, 31 May 2022, p. e3001669, https://doi.org/10.1371/journal.pbio.3001669.

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.

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

Lea AJ, Garcia A, Arevalo J, *et al.* Natural selection of immune and metabolic genes associated with health in two lowland Bolivian populations. *Proc Natl Acad Sci U S A*. 2023;120(1):e2207544120. doi:10.1073/pnas.2207544120

Leal TP, Rao SC, French-Kwawu JN, *et al.* X-Chromosome Association Study in Latin American Cohorts Identifies New Loci in Parkinson's Disease. *Mov Disord*. 2023;38(9):1625-1635. doi:10.1002/mds.29508

LEE, H.; CHEN, L. Inference of kinship using spatial distributions of SNPs for genome-wide association studies. BMC Genomics, v. 17, n. 1, 20 maio 2016.

Liang M, Nielsen R. The lengths of admixture tracts. *Genetics*. 2014;197(3):953-967. doi:10.1534/genetics.114.162362

Lumbreras, Luis Guillermo. 2015. Los orígenes de la civilización en el Perú. IAEAS, Instituto Andino de Estudios Arqueológico-Sociales.

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

Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, *46*(253), 68–78. https://doi.org/10.2307/2280095

Mills MC, Rahal C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat Genet*. 2020;52(3):242-243. doi:10.1038/s41588-020-0580-y

Naslavsky MS, Scliar MO, Yamamoto GL, *et al*. Whole-genome sequencing of 1,171 elderly admixed individuals from São Paulo, Brazil [published correction appears in Nat Commun. 2022 Mar 30;13(1):1831. doi: 10.1038/s41467-022-29575-z]. *Nat Commun*. 2022;13(1):1004. Published 2022 Mar 4. doi:10.1038/s41467-022-2864

Naslavsky MS, Suemoto CK, Brito LA, *et al.* Global and local ancestry modulate APOE association with Alzheimer's neuropathology and cognitive outcomes in an admixed sample. *Mol Psychiatry*. 2022;27(11):4800-4808. doi:10.1038/s41380-022-01729-x

Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. *Trends Genet*. 2009;25(11):489-494. doi:10.1016/j.tig.2009.09.012

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly MJ, Reich D. Methods for high-density admixture mapping of disease genes. Am J Hum Genet. 2004 May;74(5):979-1000. doi: 10.1086/420871. Epub 2004 Apr 14. PMID: 15088269; PMCID: PMC1181990.

Patterson, Nick, Alkes L. Price, and David Reich. 2006. “Population Structure and Eigenanalysis.” PLoS Genetics 2 (12): e190.

Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature*. 2016;538(7624):161-164. doi:10.1038/538161a

Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*. 2010;11(7):459-463. doi:10.1038/nrg2813

Priore, M. D. *História das mulheres no Brasil*. São Paulo: Contexto, 2006. 8 ed.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of Population Structure

Using Multilocus Genotype Data. Genetics, v. 155, n. 2, p. 945–959, 1 jun. 2000.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559-75. doi: 10.1086/519795.

Rees JS, Castellano S, Andrés AM. The Genomics of Human Local Adaptation. *Trends Genet*. 2020;36(6):415-428. doi:10.1016/j.tig.2020.03.006

Rostain S, Dorison A, de Saulieu G, *et al.* Two thousand years of garden urbanism in the Upper Amazon. *Science*. 2024;383(6679):183-189. doi:10.1126/science.adi6317

Sant'Anna HP. Ancestralidade, mapeamento de doenças complexas e dinâmica da miscigenação na população brasileira. Tese de doutorado em Bioinformática. Universidade Federal de Minas Gerais. 2020.

Sant'Anna HP *et al*. Genomic Ancestry, F8 Variants, and Immune Tolerance in Hemophilia A Patients with Inhibitors: Exome Sequencing Insights. Submetido.

Scliar MO, Sant'Anna HP, Santolalla ML, *et al.* Admixture/fine-mapping in Brazilians reveals a West African associated potential regulatory variant (rs114066381) with a strong female-specific effect on body mass and fat mass indexes. *Int J Obes (Lond)*. 2021;45(5):1017-1029. doi:10.1038/s41366-021-00761-1

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, Giordano B. Novas abordagens para integração de bancos de dados e desenvolvimento de ferramentas bioinformáticas para estudos de genética de populações. 2014. Tese (Doutorado em Bioinformática) – Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Minas Gerais.

Solis, R.S., Haas, J., and Creamer, W. (2001). Dating Caral, a preceramic site in the Supe Valley on the central coast of Peru. Science 292, 723–726.

Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful?. *Nat Rev Genet*.

2015;16(1):33-44. doi:10.1038/nrg3821

TARAZONA-SANTOS, Eduardo. A computação na nuvem da AWS impulsionando avanços em bioinformática. Blog da AWS, [*S. l.*], 5 jun. 2024.

Tarazona-Santos, E., D. R. Carvalho-Silva, D. Pettener, D. Luiselli, G. F. De Stefano, C. M. Labarga, O. Rickards, C. Tyler-Smith, S. D. Pena, and F. R. Santos. 2001. "Genetic Differentiation in South Amerindians Is Related to Environmental and Cultural Diversity: Evidence from the Y Chromosome." American Journal of Human Genetics 68 (6): 1485–96.

Thornton, T. A. *et al*. Estimating Kinship in Admixed Populations. American Journal of Human Genetics, v. 91, n. 1, p. 122–138, 13 jul. 2012.

Tian G, Tang M e Ng K, 2013. *Dirichlet and related distributions*. Hoboken, N.J.: Wiley.

Zhou D, Udpa N, Ronen R, *et al*. Whole-genome sequencing uncovers the genetic basis of chronic mountain sickness in Andean highlanders. *Am J Hum Genet*. 2013;93(3):452-462. doi:10.1016/j.ajhg.2013.07.011

ZHOU, Y.; BROWNING, S. R.; BROWNING, B. L. A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. The American Journal of Human Genetics, v. 106, n. 4, p. 426–437, 2 abr. 2020.

ANEXOS

Anexo 1 - Artigos em colaboração

Anexo 1.1 - Resumo do artigo "*Rare genetic variants of NLRP12 in Admixed Latino-American Children with SARS-CoV-2-related Multisystem Inflammatory Syndrome*", publicado em colaboração com a Fiocruz PE

JOURNAL ARTICLE ACCEPTED MANUSCRIPT

Rare genetic variants of *NLRP12* in Admixed Latino-American Children with SARS-CoV-2-related Multisystem Inflammatory Syndrome

Thaís M M Barreto, Roberta S Souza, Raquel B São Pedro, Isadora M Paiva, Andréia S Silva, Ana L Nogueira, Ana P N Bellinat, Nathália L S Dias, Sara Nunes, Gabriela S G Britto ...
Show more
Author Notes

The Journal of Infectious Diseases, jiae480, https://doi.org/10.1093/infdis/jiae480
Published: 27 September 2024 **Article history**

PDF Split View Cite Permissions Share

Abstract

Multisystem Inflammatory Syndrome in Children (MIS-C) is a rare, potentially fatal complication of SARS-CoV-2 infection. Genetic defects in inflammation-related pathways have been linked to MIS-C, but additional research is needed, especially in diverse ethnic groups. The present study aimed to identify genetic variants underlying MIS-C in Brazilian patients. Whole-exome sequencing was performed, focusing on genes involved in the host immune response to SARS-CoV-2. Functional assays assessed the impact of selected variants on NF-κB signaling. Nine rare, potentially deleterious variants were found in eight of 21 patients, located in *IL17RC*, *IFNA10*, or *NLRP12* genes. Unlike the wild-type NLRP12 protein, which inhibits NF-κB activation in HEK 293T cells, the mutant NLRP12 proteins have significantly reduced inhibitory properties. In conclusion, our results indicate that rare autosomal variants in immune-related genes may underlie MIS-C, highlighting the potential role of NLRP12 in its predisposition. These findings provide new insights for the appropriate management of MIS-C.

Keywords: MIS-C, SARS-CoV-2, Genetics, *IL17RC*, *IFNA10*, *NLRP12*, Inflammation

Topic: signal transduction, inflammation, immune response, child, genes, hispanics or latinos, genetics, severe acute respiratory syndrome, gene abnormality, whole exome sequencing, sars-cov-2, covid-19, pediatric inflammatory multisystem syndrome, racial/ethnic diversity

Issue Section: Major Article

Anexo 1.2 - Resumo do artigo "*The Peruvian Genome Project: Expanding the Global Pool of Genome Diversity from South America*", em submissão em colaboração com o Projeto Genoma Peruano

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES
CSH Cold Spring Harbor Laboratory BMJ Yale

Follow this preprint

The Peruvian Genome Project: expanding the global pool of genome diversity from South America

Heinner Guio, Omar Caceres, Cesar Sanchez, Carlos Padilla, Omar Trujillo, Victor Borda, Luis Jaramillo-Valverde, Julio A. Poterico, Carolina Silva-Carvalho, Mary Horton, Cristina M Lanata, Alessandra Carnevale, Sandra Romero-Hidalgo, Samuel Canizales-Quinteros, Víctor Acuña-Alonzo, Marco Machacuay-Romero, Pedro Novoa, Roberto Frisancho, Ruth Shady, Pedro Flores-Villanueva, Timothy D. O'Connor, Manuel Corpas, Eduardo Tarazona-Santos

doi: https://doi.org/10.1101/2024.05.05.24306840

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

0 0 0 0 0 0 18

Abstract Full Text Info/History Metrics Preview PDF

SUMMARY

The process of inhabiting the Americas by ancestral native American populations involved many individuals settling in the Peruvian Andes and Amazonian regions. Due to Latin American countries representing less than 1% of the human genome data available in public reference databases, the evolution and migration processes involved in adapting to this unique geography have not yet been fully explained. The Peruvian Genome Project is an initiative, started in 2011, to address the underrepresentation of genomic data from native South American populations. This project has collected 1,149 samples from 17 traditional native and 13 mestizo (mixed of native, African, and European ancestry) communities. Currently, 150 whole genomes and 873 array-genotyped individuals have been sequenced from across the geography of Western South America, including coastal, Andes, and Amazonian regions. We discovered 1.6 million novel genetic variants with varying frequencies, indicative of local environmental adaptations and population drift. These novel variants allow us to infer local evolutionary traits and population-specific allele frequencies for people living at different altitudes, as well as varying adaptations to pathogens and living conditions. The Peruvian Genome Project is the result of over a decade of work in sample selection, logistics, and approved regulatory community engagement, designed to enhance the human genome pool of diversity of native Americans. The data collected here enable the targeted characterization of endemic diseases, trait adaptations, and new variants of clinical significance in South America. The Peruvian Genome Project represents a step forward in international and multidisciplinary efforts to make precision medicine more inclusive and accessible for underrepresented communities in Latin America, offering significant potential for drug development and diagnostics in a neglected continent.

Anexo 1.3 - Resumo do artigo "*The effect of altitude on the expression of immune-related genes in Peruvian rural indigenous*", em submissão em colaboração com o Projeto Genoma Peruano

The effect of altitude on the expression of immune-related genes in Peruvian rural indigenous

Authors: Luis Jaramillo-Valverde[1,2+], Gilderlanio Santana de Araújo[1+], Julio A. Poterico[1,2], Catalina Martinez-Jaramillo[1], Vicky Roa-Linares[1], Sandra Alvites-Arrieta[1], Nelis Pablo-Ramirez[1], Milward Ubillus[1], Diana Palma-Lozano[1], Carolina Silva-Carvalho[3], Luca Vasconcelos-da-Gama[3], Lucas F Costa[3], Eduardo Tarazona-Santos[3], Soumya Raychaudhuri[4], Heinner Guio[1,2*]

Universidad de Huánuco, Huánuco, Peru

[2]INBIOMEDIC, Research and Technological Center, Lima, Peru

[3]Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Belo Horizonte, Universidade Federal de Minas Gerais, MG, Brazil

[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[+]Equal contribution

*Corresponding Author:

E-mail: heinnerguio@gmail.com

AUTHOR CONTRIBUTIONS

Study design: LJ-V, SR and HG. Performed the experiments: LJ-V, GSA,JAP,CMJ,VRL,SAA,NPR,MU,DPL. Analyzed the data: LJ-V, GSA, CSC, LVG, LFC,ETS,SR and HG All authors have read and approved the final manuscript.

ABSTRACT

Background: Some genes associated with immune response have been elucidated in some Andean populations, which may explain part of the immune system adaptation and pathogen response in high-altitude residents. We performed differential expression analysis and mainly focused on genes with high fold-change between groups and those genes with differential

expression related to the immune system. Methods: We performed a transcriptome analysis of Peruvian individuals, primarily from rural areas, with high genetic backgrounds from ancient indigenous people, exposed to different living altitudes (high vs low). We collected peripheral blood mononuclear cells (PBMC) from 62 volunteers and exposed them to bacterial lipopolysaccharide (LPS), Pam3CSK4 (a synthetic triacylated lipopeptide responsible for bacterial components), and R848 (an imidazoquinoline compound related to viral nucleic acids). Results: Differential expression analysis results and data integration analysis with immune system pathway data exhibit a set of 17 genes associated with the immune system, nine down-regulated genes (*ITGAX, CCL22, CSF1, CXCL8, IL12A, MMP9, CSF2, PTGS2, ENSG00000138685.17*), and eight up-regulated genes (*HLA-DPB1, FN1, CD36, MMP2, HLA-DRB1, FCGR1A, CCL17, HLA-DRB5*). These genes are notably enriched in antigen-presenting cells, such as dendritic cells and macrophages. Conclusion: The differential expression of immune system genes in native inhabitants with a prominent Andean genetic background highlights the active involvement of both the innate and adaptive immune systems. Remarkably, our study revealed a distinct gene expression signature in our population, setting it apart from other transcriptomic studies conducted thus far.

Keywords: altitude, expression, immune-related genes, rural indigenous

Anexo 2 - Materiais suplementares de artigos em primeira autoria

Anexo 2.1 Material Suplementar do artigo: "The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture"

PNAS
www.pnas.org

Supplementary Information for

The genetic structure and adaptation of Andean highlanders and Amazonian dwellers is influenced by the interplay between geography and culture

Victor Borda[1,2,3,19], Isabela Alvim[1,19], Marla Mendes[1,19], Carolina Silva-Carvalho[1,19], Giordano B Soares-Souza[1], Thiago P Leal[1], Vinicius Furlan[1], Marilia O Scliar[1,4], Roxana Zamudio[1], Camila Zolini[1,5,6], Gilderlanio S Araujo[7], Marcelo R Luizon[1], Carlos Padilla[3], Omar Cáceres[3,8], Kelly Levano[3], César Sánchez[3], Omar Trujillo[9], Pedro O. Flores-Villanueva[3], Michael Dean[10], Silvia Fuselli[11], Moara Machado[1,10], Pedro E. Romero[12], Francesca Tassi[11], Meredith Yeager[10], Timothy D O'Connor [13,14,15], Robert H Gilman [12,16], Eduardo Tarazona-Santos[1,17,20], Heinner Guio[3,8,18,20]

Eduardo Tarazona-Santos, Heinner Guio.

Email: edutars@gmail.com, heinnerguio@gmail.com

This PDF file includes:

Supplementary text

Figures S1 to S30

Legends for Datasets S1 to S10

SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S10

Supplementary Information Text

Section 1: Sampling, quality control and Datasets

1.1. Sampling: The present work is part of the Peruvian Genome Diversity Project (PGDP), of the Peruvian Institute of Health (Instituto Nacional de Salud - INS). This is a genomic initiative to explore the genetic composition of native and admixed Peruvians. The protocol for this study was approved by The Research and Ethics Committee (OI003-11 and OI-087-13) of the INS. The PGDP was funded by the Ministry of Health of Peru and involves the collaboration of Tarazona-Santos's Laboratory of Human Genetic Diversity (LDGH) of Universidade Federal de Minas Gerais (UFMG) and INS. Fifteen Native American populations (INS data) were sampled as part of the PGDP including populations from the South Pacific Peruvian Coast (SPPC), Andean and Amazon regions (Dataset S1 and Fig. 1).

Populations from the SPPC region were sampled in northwestern Peru and involved two native communities: Moche and Tallanes. In the Andes, four communities were sampled, two Quechua-speaking communities, Qeros and Chopccas, from South Central Andes, and two Aymara-speaking populations, Jaqarus and Uros. The Amazon populations were sampled in two different ecological areas, Amazon Yunga and lower Amazon. The Amazon Yunga corresponds to a cloudy forest which is a transition between the Andean mountains and the Lower Amazon. Six populations were sampled from Amazon Yunga: Chachapoyas, Lamas, Awajun, Candoshi, Ashaninkas and Matsiguenkas. Chachapoyas comprises several groups that played an important role for the Andean people as an open door to Amazon resources (1). The Lamas population, that lives in the Upper Huallaga river, started as a "reduction" (a forced concentration of Native American groups) during the XVIII century (2). Currently, Both Chachapoyas and Lamas groups speak Quechua. The Awajun population belongs to the Jivaroan linguistic family and its presence in the Amazon Yunga is dated back to around 1,200 years ago (3, 4). This Jivaroan population was involved in several cultural trades with South Pacific Coast populations from Peru and Ecuador but with conflicted interaction with the Inca empire (Andes). The Candoshi group is settled along the tributaries of the Pastaza River and its language has a controversial origin by being considered by some scholars as part of the Jivaro group or even independent (5). For the Arawakan linguistic group, three individuals were collected in the Sepahua district from the Ucayali region that belong to the Matsiguenkas tribe. For the Lower Amazon, three populations of the Panoan linguistic family were collected: Shipibo-Conibo, Matses and Nahuas. Finally, all native participants were required to be over 18 years old, for whom all four grandfathers were born in the selected ancestral native population.

We also included four Native American populations (LDGH data) collected by Universidad Peruana Cayetano Heredia and includes two Andean (one Quechua and one Aymara-speaking) and two Amazonian (Ashaninkas and Matsiguenkas from Shimaa community) (Dataset S1 and Fig. S1). This sampling was conducted under approval of the Institutional Reviews Boards from the Universidad Peruana Cayetano Heredia, Asociación Benéfica PRISMA, Universidade Federal de Minas Gerais and Johns Hopkins University. The Quechua-speaking population was sampled in a large area comprising two continuous regions, Ayacucho and Apurimac, for this reason these individuals were grouped as Quechuas_AA. In the case of the Aymara-speaking population, individuals were collected near the Titicaca lake shore in Puno region. The two Amazonian populations inhabit the Amazon Yunga area and belong to the Arawakan linguistic family.

1.2. DNA sampling: For INS data, after collecting the 10ml blood sample from participants (we only collect blood) we proceed to extract the DNA using commercial kits (Qiagen, USA). For this

procedure, we travel to the community with our supplies and equipment to perform the DNA extraction, when we obtain the DNA purified we aliquoted, send to our Laboratory (Laboratorio de Biotecnología y Biología Molecular) in Lima, Perú to be stored in -80 °C until we continued with the next project phase genotyping. Specifically, for LDGH data, we extracted genomic DNA from blood samples using the Gentra Puregene blood kit (Qiagen, USA) in the LDGH.

1.3. Genotyping and Quality Control: A total of 289 individuals were genotyped using the Illumina Human Omni array 2.5M at the INS. The total number of genotyped SNPs was 2,391,739. Quality control was performed using the PLINK 1.7 software (6) and in-house scripts (7). We removed SNPs and individuals with high levels of missing data (>10%), loci with 100% of heterozygous, non chromosomal information and A/T-C/G genotypes. LDGH data was genotyped by the Illumina facility using the HumanOmni2.5–8v1 array for 127 individuals. Quality control for LDGH data was the same as for the INS data. We merged the INS data with LDGH data (Dataset S1 and Fig. S1). populations:.The merged data, INS and LDGH individuals, contain a total of 2,077,858 SNPs for 418 individuals organized in a total of 19 populations (Dataset S1). Both groups, INS and LDGH datasets, include independent samples of the Ashaninka population from the same region, for this reason we merge these individuals in a unique Ashaninka sample and a total of 18 populations.

Before filtering by relatedness, we removed SNPs that were in high linkage disequilibrium (LD) for each population, as it affects the inferences of relatedness, with PLINK 1.7 using the flag --indep-pairwise with the following parameters: 200 25 0.1. The first parameter indicates a window of 200 SNPs, the second indicates that the window steps of 25 SNPs between consecutive windows and the third indicates the LD threshold (r^2).

Family structure affects the analysis of population structure as a familiar cluster can be confounded with a discrete population (8). To overcome this issue, we estimated the kinship coefficients (Φij) for each pair of individuals for each population using autosomal SNPs. For each population, we estimated the kinship coefficients using the option –genome in PLINK 1.7. We considered a thresholds of Φij $\geq$0.25 to define relatedness or not. A pair of individuals with Φij above 0.25 is defined as first-degree relatives (Parent-offspring pair and full sibling). We used a network approach to identify which individuals should be removed preserving a maximum number of unrelated individuals (9). After applying the kinship filter, we kept 358 individuals (Dataset S1) for an unrelated dataset (UDataset).

1.4. Merging datasets: We merged the UDataset with the following datasets:

- 1000 Genomes project (10).
- Human Genome Diversity Project (HGDP) (11).
- Native Americans previously genotyped by Reich *et al.* (unmasked data) (12).
- Native individuals from Guatemala (Kaqchikel population) from Michael Dean-Lab (National Cancer Institute).
- Native American individuals from two public datasets (Simons Genome Diversity Project (13) and Raghavan *et al.,* 2015 (14)).

From the 1000 Genomes Project, we selected individuals of European (IBS, CEU), African (YRI and LWK) and East Asian (CHS, CDX, CHB) ancestries. The unmasked dataset from Reich *et al.* (12) included individuals from HGDP: Yakut, Karitiana, Surui, Pima, Maya, Piapoco, Papuan and Melanesian. From the Simons Genome Diversity Project and individuals generated by Raghavan *et al.* (14), we included all Native Americans. The available dataset of Raghavan *et al.* (14) included the ancient genome of Anzick-1 individual from the Clovis complex (hereafter Clovis). Before merging individuals from Reich *et al.* (12), we applied a relatedness filter. We removed 58 individuals with kinship coefficient above 0.1 using the same procedure employed for our samples. We generated three datasets (Datasets S1-S3, Fig. S1) considering the density of SNPs and sample size:

a) **Natives 1.9M Dataset (**1,927,769 SNPs/673 individuals**):** Dataset with maximum number of genotyped SNPs. This dataset includes just Peruvian Native individuals from INS and LDGH and 107 Iberian (IBS), 108 Yoruba (YRI) and 100 East Asian individuals (CDX) from 1000 Genomes Project (Dataset S1).
b) **Natives 500K Dataset (**567,718 SNPs/849 individuals**)**: This dataset includes individuals from **Natives 1.9M Dataset,** Native American, Siberian, South Asian (Onge) and Oceanian (Bougainville and Papuan), individuals from the Simons Project (13), Raghavan *et al.,* 2015 and 79 individuals from Guatemala of Michael Dean NCI lab (Dataset S2) genotyped for 600K SNPs. The Guatemalan sample includes individuals from the Kaqchikel native population and non-native individuals with more than 99% of Native American ancestry.
c) **Natives 230K Dataset** (235,352 SNPs/1,286 individuals): Dataset with maximum number of individuals. This data includes individuals from Natives 1.9M dataset and all Native Americans from the unmasked data of Reich *et al.* (2012) (~300K SNPs), which includes HGDP individuals (Dataset S3).

The East and South Asian, Siberian and Oceanian populations were used only for population history analysis of the masked data and genotype based methods and not for the population structure analyses in order to avoid any confounding signal.

Use of datasets masked for Non-Native American local ancestry: For *D* statistics analysis (15) and Admixture Graphs (16), we used a dataset where regions of European and African ancestries were masked. These regions were identified using RFMix software (17) and then masked. Using masked datasets and methods based on allele frequency correlation, we inferred genetic affinity among South American Natives. RFMix identifies regions of a specific ancestry in the genome of admixed individuals using reference panels of individuals of European, African and Native American ancestries. For this purpose, we used the phased **Natives 500K** (Dataset S2) and **Natives 230K** (Dataset S3) **datasets**. We used 100 African (YRI and LWK) and 100 European (CEU and IBS) individuals from 1000 Genomes project as parentals. For the Native American reference panel, we selected individuals with less than 0.002% of Non-Native ancestry (European + African ancestries) using the ADMIXTURE results (see Section 2) for 3 ancestry clusters (K=3). All other Native American individuals that have some level of European or African ancestry were used as targets. We ran RFMix with the option PopPhased to enable the phase correction option. We also used two rounds of the expectation-maximization (EM) algorithm. All other settings were used as default. After running RFMix, we used the forward-backward probability output to set all local ancestry inferences that have less than 0.95 posterior probability of being Native American as missing data. Finally, the genomic regions in each sample that did not contain homozygous high quality Native American ancestry inferences were set as missing data.

In this paper, we will apply several methods on our three datasets to explore the following four

scientific questions:

Question 1 (Section 2): whether the between-population homogenization of Western South America, and the dichotomy Arid Andes/Amazonia extends to the northward Fertile Andes?

Question 2 (Sections 2 and 3): whether gene flow accompanied the cultural and socioeconomic interactions between Andean and Amazon Yunga populations?

Question 3 (Section 4): when this between-population genetic homogenization started in the context of the arid Andean chronology.

Question 4 (Section 5): were there episodes of genetic adaptation to the Arid Andes and the Amazonian tropical forest?

Section 2: Genetic relationships in Western South America

To address our scientific questions:
Question 1: whether the between-population homogenization of Western South America, and the dichotomy Arid Andes/Amazonia extends to the northward Fertile Andes?

and

Question 2: whether gene flow accompanied the cultural and socioeconomic interactions between Andean and Amazon Yunga populations?

We performed population structure analysis using two approaches: genotype based (ADMIXTURE and PCA) and haplotype based (CHROMOPAINTER and fineSTRUCTURE) methods.

2.1. Methods

2.1.1. Population Structure using genotype based methods

We applied genetic clustering analysis and Principal Component Analysis (PCA). For the genetic clustering analysis, we ran ADMIXTURE (18). The ADMIXTURE algorithm assumes that the genetic composition of each individual is made up of up to K parental populations or ancestry clusters, where K is defined by the user. ADMIXTURE estimates the fraction of each K population that contributes to an individual, as well as the allele frequencies of each of the K populations, by fitting the Hardy Weinberg equilibrium in each of the K populations/clusters. We ran ADMIXTURE in unsupervised mode for different values of K and used a cross validation (CV) test to determine the K value with the best model fitting. The ADMIXTURE results are represented as a bar plot where each individual is represented by a vertical bar in which each color corresponds to the ancestry proportion of a specific cluster. The PCA is a non-model based method that reduces a complex data (i.e. genotypes and individuals) to few dimensions (19).

ADMIXTURE analysis and PCA assume independence among SNPs, for this reason we pruned all datasets for linkage disequilibrium (LD). We removed highly linked SNPs using PLINK 1.7 with the option indep-pairwise 200 25 0.4 for each dataset. We generate three datasets pruned by LD:

- **Natives 1.9M dataset_LDpruned** (625,736 SNPs)
- **Natives 500K dataset_LDpruned** (229,895 SNPs)
- **Natives 230K dataset_LDpruned** (136,797 SNPs)

We ran 50 replicates of ADMIXTURE in unsupervised mode with different random seeds for each K value and calculated the cross validation error for each run. We ran ADMIXTURE considering from K=2 ancestral clusters until cross validation error started to increase for each dataset. We plot all ADMIXTURE runs with the higher log likelihood for each K value. We ran the PCA using EIGENSOFT 4.21 (19) for the three LD pruned datasets.

Natives 1.9M dataset

ADMIXTURE results are displayed on Fig. S2. The lower CV error was obtained for the run with five ancestry clusters (K=5). ADMIXTURE run with K=3 infers clusters related to continental ancestry: Native American (green), European (IBS, red) and African (YRI, blue) clusters. This result showed some Native American individuals (Quechuas_AA, Chachapoyas and Moche populations) with European ancestry (~10%). Specifically, for the result with the lowest cross validation error (K=5), we observed the Andean populations as a homogeneous group (brown cluster). On the other hand, we observed an ancestry cluster (light green) predominant in Northern Peruvian populations that is shared between SPPC and Chachapoyas population (Amazon Yunga).

For the PCA (Fig. S3), we excluded Africans (YRI) due to its high level of differentiation that masks the relationships in Native Americans. The first principal component (PC1, Variance explained=2.36%) showed an axis of differentiation between the European and Native American groups. We observed that some Andean, Moches, and Chachapoyas individuals have some degree of European ancestry. The PC2 (Variance explained=1.2%) separated Western (Andean and SPPC populations) and Eastern (Amazon) South American natives. Chachapoyas showed affinity with SPPC populations. Jivaroan populations (Awajun and Candoshi), were intermediate in the axis Western-Eastern. Furthermore, the PC2 showed a cline for the genetic diversity of the Amazon populations, from North (Matses) to South (Matsiguenkas). As in ADMIXTURE, both Matsiguenkas groups, Shimaa (Matsiguenkas 1) and Sepahua (Matsiguenkas 2), showed high genetic affinity.

For this dataset, ADMIXTURE analysis and PCA showed high differentiation between populations within the Amazonia and high genetic affinity among Central Arid Andean groups. Chachapoyas showed a close genetic relationship with SPPC populations. Moreover, North Amazon populations (Awajun and Lamas) share ancestry with SPCC as well as with other Amazonian populations.

Natives 500K dataset

ADMIXTURE results are presented on Fig. S4. For bar plot representation, we grouped Surui and Karitiana as Tupian. Mesoamerican individuals were divided into Guatemalan and Mexican (Mixe, Mixtec, Pima, Zapotec and Mayan), and we grouped the Clovis individual, two Greenland, two Aleutian and two Athabascan individuals as North America. Our ADMIXTURE runs showed the lowest CV error for eight clusters (K=8). Our description was focused on patterns not observed on the 1.9M dataset for the lowest cross validation.

ADMIXTURE run K=8 showed 6 clusters associated with Native American groups, associated with Andes (brown), Mesoamerica (purple), SPPC (pink) and three Amazon related clusters (shades of green). SPPC populations showed a predominant pink ancestry that is also predominant in Chachapoyas population. The Andean populations have a predominant brown cluster. Moreover, central Andean populations (Jaqarus, Quechuas_AA and Chopccas) showed ~10% of SPPC related ancestry. Matsiguenkas individuals were observed as a highly differentiated population since it has a specific ancestry cluster (darkgreen) which is not shared with other populations of the same linguistic group (Ashaninkas). Panoan populations (Shipibo, Matses and Nahua) showed a predominant ancestry associated with the Ashaninkas population. Jivaroan groups showed a specific ancestry which was predominant in the Awajun population.

For the Principal Component Analysis (Fig. S5), the PC1 separated Native Americans from Europeans. Some Chachapoyas, Quechuas_AA, 1 Moche, 1 Mixtec and 1 Shipibo individuals showed affinity to Europeans due to admixture. The PC2 separated Amazon from non-Amazon populations; Jivaroan and Tupian individuals were observed as intermediate between these groups. The PC3 separates a group that includes Andean and Matsiguenkas individuals from other natives. PC4 showed the separation between a group including Mesoamericans and Tupian individuals from other natives. Higher PC values showed population specific differentiation and genetic variation in the IBS population.

For this dataset, both ADMIXTURE and PCA support the similarity between SPPC and Chachapoyas individuals. Awajun and Candoshi were intermediate between Andean-Amazon axis of genetic diversity.

Natives 230K dataset

The following description was focused on clusters related to South American natives. The lower CV error was obtained for 18 ancestral clusters (K = 18, Fig. S6). The ADMIXTURE plots (Fig. S6) from K=3 to K=5 inferred continental ancestry clusters. The ADMIXTURE plot K=5, five ancestry clusters related to each continental region were identified: Africa, Europe, Asia, Oceania (light purple, pops 90-91) and America. ADMIXTURE K=6 showed a cluster (dark green) associated with Arawakan groups (Ashaninkas and Matsiguenkas). ADMIXTURE K=8 identified a cluster (light pink) for Costa Rican natives (Bribri, Cabecar, Chorotega, Guaymi, Huetar, Maleku, Teribe). ADMIXTURE K=9 inferred an ancestral cluster (black) related to Mesoamericans, predominantly in Pima. ADMIXTURE K=10 showed an ancestry cluster (light green) related to the Awajun population which represented most of

non Andean natives. ADMIXTURE K=12 identified an ancestry cluster (gray) associated with Tupian populations (Surui and Karitiana) also observed in Mesoamericans and in non Andean natives. Furthermore, a brown cluster is predominant in the Andean populations. ADMIXTURE K=13 showed a light blue cluster associated with Pima natives. ADMIXTURE K=15 showed an ancestry cluster, darkgreen and forest green, for each Peruvian Arawakan individuals (Ashaninkas and Matsiguenkas). ADMIXTURE K=17 detected a cluster (light blue) associated with the SPPC population, and which is also present in Chachapoyas and Lamas. The ADMIXTURE K=18 with lowest cross validation showed differentiation in Asian natives. ADMIXTURE K=21 showed an ancestry cluster related to Lamas sample.

In the Principal Component Analysis (Fig. S7), the PC1 showed the differentiation between the Old world from the Native American ancestry. Old world ancestry included Europeans (IBS), Oceanians, Asians (Russian and Mongolians) and admixed Native Americans. Greenland natives were shown as intermediate between the Old and the New world groups. In the Native American axis, the Matsiguenkas individuals were the most differentiated. Considering PC1 and PC2, we discriminated three blocks of Native ancestry, one Eskimo-Aleut (Greenland), the second Athabaskan and Algic (North American group), and the third including all other Native Americans. This pattern is consistent with Reich *et al. (12)*.

In summary, our genotype frequency approach supports the dichotomic model between the Arid Andes and the adjacent Amazonia. However, this is not valid in Northern Peru in which SPPC populations were closely related to Amazon Yunga Chachapoyas, Moreover Jivaroan populations were observed as more similar to SPPC and Chachapoyas than to other Amazon populations Arawakan and Panoan. Furthermore, The Fertile Andes populations (SPPC and Chachapoyas) and Jivaroan populations, conditioning on K=17 ancestry clusters, shared some level of genetic ancestry with Mesoamericans.

2.1.2. Population Structure using haplotype based methods

We used haplotype-based methods (CHROMOPAINTER and fineSTRUCTURE algorithms) that explore the patterns of haplotype similarity among individuals (20). First, we phased our datasets using shapeit2 software (21). For the phasing process, we used the complete dataset (without LD pruning). To increase the accuracy of the phasing process, we used 200 conditioning states and 30 main iterations of Markov chain Monte Carlo (MCMC).

The haplotype-based methods are based on the identification of the LD patterns along the genome of individuals in order to infer the number and length of DNA chunks (CHROMOPAINTER) shared among them. Then, this information is exploited in the identification of clusters of individuals based on the pattern of genetic similarity at a fine scale (fineSTRUCTURE). The identification of shared DNA chunks between individuals is called chromosome painting and is performed by CHROMOPAINTER (20). In this process each haplotype (recipient) is reconstructed based on chunks shared (or “donated”) with (by) other individual haplotypes (donors) (20). For this inferences CHROMOPAINTER requires phased data and two scalar parameters (inferred in a previous CHROMOPAINTER run): 1) the recombination scaling and 2) mutation parameters. The result of the chromosome painting is summarized in two interindividual matrices called coancestry matrices. These matrices of putative donor-recipient similarity contain as their elements the total number (chunkcounts) and the length (chunklengths) of DNA chunks shared among individuals.

After the chromosome painting, we used the chunkcounts co-ancestry matrix to infer the population structure using the model-based approach fineSTRUCTURE (20). Using a reversible-jump MCMC, fineSTRUCTURE assigns individuals into clusters that may resemble their populations. Like other MCMC algorithms, fineSTRUCTURE is dependent on the number of MCMC iterations so it uses a previous burn-in stage and then several iterations (i.e 2 millions).

Since our main question is about the history of Native Americans, we excluded individuals with >5% of Non-Native ancestry in the Natives 230K dataset except for Chachapoyas individuals. We did this because Chachapoyas population has almost all individuals with more than 5% of Non-Native ancestry, so we maintained all individuals (see ADMIXTURE results). We removed slightly admixed individuals just in the 230K dataset because being the most numerous dataset, the exclusion of 202

individuals did not represent a considerable loss. We determined which individuals had to be removed based on ADMIXTURE results (K=3). For Natives 1.9M and Natives 500K datasets we maintained the complete dataset.

To define two scalar parameters (recombination scaling and mutation) for the entire dataset analysis, we ran the Expectation-Maximization CHROMOPAINTER algorithm for a subset of individuals and chromosomes for each data set, we obtained the following values for the parameters. The recombination scaling and mutation were respectively: 220.324 and 0.00018 for the Natives 1.9M dataset, 144.362 and 0.0002 for the Natives 500K dataset, and, finally, 150 and 0.00045 for the Natives 230K dataset. After obtaining the co-ancestry matrix, we ran fineSTRUCTURE for all datasets considering 1,000,000 of burn-in steps, 2,000,000 of MCMC iterations and 100,000 of sampling. After the MCMC calculations, we construct the tree using 100,000 additional steps of hill-climbing steps. We represented the fineSTRUCTURE results as a tree and the chunklengths coancestry matrix as a heatmap.

Natives 1.9M dataset

The fineSTRUCTURE tree clusterize the native individuals in three main groups: natives with some European admixture (Fig. S8A), Amazon (Fig. S8B), SPPC and Andean individuals (Fig. S8C). Almost all Native Americans (~95%) were grouped in clusters containing individuals of the same population label. Most of the SPPC individuals have a close relationship with Andean populations. In the Amazon, we observed that Jivaroan populations (Candoshi and Awajun) were not closely related among them as is observed in other amazon linguistic groups (Fig. S8B). Chachapoyas population showed a close relationship with the admixed Native Americans.

Natives 500K dataset

The arrangement of the clusters was similar to the resulting tree of Natives 1.9M dataset except (Fig. S9): 1) the Moche cluster and the Chachapoyan cluster were more similar to Andean clusters (Fig. S9A), 2) West Mesoamerican natives (Mixe, Mixtec, Zapotec and Pima) clustered together with some Amazon Tupian (Surui and Karitiana) individuals (Fig. S9B), 3) Peruvian Arawakan natives (Matsiguenkas and Ashaninkas) were shown as the most differentiated clusters (Fig. S9D).

Natives 230K dataset

The fineSTRUCTURE tree (Fig. S10A and S10B) showed a more external cluster that contains IBS, Chipewyan and Greenland natives, reflecting the high level of admixture of these North Native American populations. The second more external clusters include Arawakan speakers as the most differentiated populations (Fig. S10B). Other Native Americans were organized in two macro clusters, Andean (Fig. S10B) and non-Andean clusters (Fig. S10A and S10B). The Andean cluster (Fig. S10B) showed no major differences from the clustering of the other datasets, showing Uros as the most differentiated Andean population. Costa Rican populations (Fig. S10B) showed close affinity to Northern South American populations (Embera, Wayuu, Waunana and Kogi). Moreover, Mayan populations (Mayan and Kaqchikel individuals) showed close affinity to Northern Peruvian (Lamas and SPPC) and Inga individuals (Fig. S10A). Panoan individuals (Matses and Nahua) clusterize with other Eastern populations (Fig. S10A).

Summarizing the fine-scale population structure analyses, we inferred that Mesoamericans (Maya and Kaqchikel) share more ancestry with SPPC natives than with Arid Andes populations. The Arid Andes populations showed high similarity among them. Moreover, Chachapoyas populations were highly similar to SPPC individuals.

2.1.3. Ancestry profiles inferred by GLOBETROTTER and SOURCEFIND

We performed a Chromosome painting inference for each Native American population, setting a population as recipient from all other populations (CHROMOPAINTER “-f” switch). This inference result in a chunklengths matrix that summarizes the contribution (shared DNA) from the donor populations to the recipient. Then, with this matrix, we applied two approaches to infer the ancestry proportions: a regression model (non-negative least squares) implemented on GLOBETROTTER software (MIXTURE MODEL) (22–24) and a Bayesian model implemented in SOURCEFIND (25). To

generate the chunklengths output for each Native American groups, we used the same two scaling parameters inferred in the last subsection. Furthermore, since fineSTRUCTURE and ADMIXTURE showed no differences among Matsiguenkas individuals, Matsiguenkas 1 (Shimaa) and Matsiguenkas 2 (Sepahua), for GLOBETROTTER inferences we considered these two groups as one single Matsiguenkas group.

Natives 1.9M dataset

The MIXTURE MODEL (Fig. S11) revealed a particular pattern in Andean populations. Each Andean population shares a high proportion of DNA with other Andean populations showing a homogeneous pattern. Specifically, the ancestry profile of the Uros population showed that 95% of its DNA is shared with the Puno population indicating lower genetic variability of Uros. Moreover, Aimara-speaking Jaqarus showed high genetic affinity with Quechua-speaking groups. The SPPC populations have more affinities with northern Amazon populations. These affinities are related to the similar ancestry proportions and the ancestry related to Chachapoyas. Also, the three North Amazon populations (Candoshi, Awajun, Lamas and Chachapoyas) showed a significant proportion of ancestry related to SPPC populations (>10%). Furthermore, we observed that SPPC and North Amazon populations have a significant proportion of shared DNA with Andean population (>20%). Simulations suggest that SOURCEFIND tends to eliminate contributions that could be due to background noise (25). For this particular dataset, SOURCEFIND identifies the SPPC ancestry in north Amazon populations and that most of the Andean ancestry is explained by Quechua related ancestry (Fig. S11). For this dataset, that has the advantage of being the more dense in terms of number of SNPs, it is important to consider the poor representation of other native populations, Mesoamericans and South Americans. The Amazon populations showed a common pattern of genetic composition among them. Only Quechuas_AA, Jaqarus, Moche and Chachapoyas showed European ancestry.

Natives 500K Dataset

The MIXTURE MODEL (Fig. S12) showed the Andean populations with the same pattern as was observed with the Natives 1.9M dataset, they share a high proportion of DNA with other Andean populations showing a homogeneous pattern. Furthermore, it is possible to observe Mesoamerican related ancestry in almost all Peruvian natives. The ancestry composition of SPPC populations showed more similarity to Chachapoyas population. Moreover, Jivaroan and Lamas populations showed very similar patterns of ancestries. We observed that Moche and Chachapoyas populations have a significant proportion of shared DNA with Andean population (~30%). All SPPC and Amazon populations (except Ashaninkas and Matsiguenkas) showed more ancestry related to Mesoamerican (>18%) than to the Andean populations. Matsiguenkas showed more ancestry related to Ashaninkas, indicating a close genetic affinity among Arawakan populations. SOURCEFIND (Fig. S12) showed that the Mesoamerican related ancestry is particularly high in SPPC and Karitiana populations. The Andean contribution in the Amazon was reduced and restricted to the Chachapoyas population. Moreover, Jivaroan (Awajun and Candoshi) populations showed some contribution from the SPPC group.

Natives 230K Dataset

Using fineSTRUCTURE results (Fig. S10A and S10B), we organize Native American populations in the following clusters:

- Chopccas (n=17)
- Quechuas_Per (n=13) [Quechuas_AA and Quechua_R2]
- Aimaras_PB (n=29)
- Qeros (n=12)
- Uros (n=13)
- Quechuas_Bol (n=10)
- Ashaninkas (n=35)
- Matsiguenkas (n=26)
- Matses (n=11)
- Nahua (n=2)
- Awajun (n=23)
- Lamas (n=21)
- Chachapoyas (n=9)
- Moche (n=25)
- Tallanes (n=34)
- Pima (n=21)
- Tepehuano (n=21)
- Maya (n=29)
- Kaqchikel (n=6)
- Mixe (n=17)
- Zapotec1 (n=6)

- Zapotec2 (n=21)
- Mixtec (n=5)
- Karitiana (n=8)
- Surui (n=10)
- Guahibo (n=6)
- Palikur (n=3)
- Piapoco (n=6)
- Ticuna (n=4)
- Embera (n=4)
- Kogi (n=3)
- Waunana(n=3)
- Wayuu (n=2)
- Toba (n=4)
- East_Amazon_Brazil [Arara (n=1) - Parakana (n=1)]
- Inga (n=2)
- 1 Chaco cluster [Guarani (n=3) - Chane (n=2)]
- Wichi (n=4)
- Maleku (n=2)
- Cabecar (n=16)
- Guaymi (n=5)
- Teribe (n=3)
- Chipewyan (n=3)
- East Greenland (n=3)
- IBS cluster (n=107)
- CDX cluster (n=93)

The number in parenthesis indicates the number of individuals after filtering the ones with more than 5% of European ancestry. We did not include Chono and Huilliche as donors for two reasons 1) Due to the low probability that they were involved in gene flow events with Central Andes or Amazonia (12, 26), and 2) almost all individuals have higher levels of European ancestry (>5%). Furthermore, the Quechuas individuals from our dataset that clustered with Quechuas R2 from Reich *et al.* (2012) (12) were included as Quechuas Per and all Quechuas R1 were considered as Quechuas Bol (from Bolivia). Although Jamamadi samples form a cluster with Arara and Parakana, these individuals are geographically distant from Arara and Parakana, for this reason we excluded them from the GLOBETROTTER analysis. After the contribution inferences, and just to improve the visualization of the results, we made an arbitrary merge of clusters as follow:

- North Amazon 1: Inga and Ticuna.
- North Amazon 2: Guahibo, Palikur and Piapoco.
- Caribbean: Embera, Kogi, Wayuu and Waunana.
- Chaco natives: Wichi, Guarani and Chane.
- Pampas: Toba.
- Central America: Cabecar, Maleku, Guaymi and Teribe.
- Mayan: Maya1 and Maya 2.
- West Mesoamericans: Mixe, Mixtec, Zapotec1 and Zapotec2.
- North American: East Greenland and Chipewyan.

The MIXTURE MODEL (Fig. S13) showed a contrasting pattern between Western (SPPC and Arid Andes) and Amazon groups. The SPPC populations as well as Amazonian groups showed shared haplotypes with Mesoamerican groups. Differently, SOURCEFIND (Fig. S13) only detected sharing with Mesoamerican haplotypes for Chachapoyas, Awajun and Inga groups. Moreover Chachapoyas and Inga groups showed some level of Andean related ancestry.

2.2. Conclusions

Considering our Question 1, we conclude that the genetic dichotomy between populations living on the Arid Andes and adjacent Amazonia **does not extend** to the Fertile Andes. Furthermore, Chachapoyas, an Amazon Yunga population, is more genetically similar to SPPC populations than to Arid Andes or other Amazon populations. Regarding our Question 2, GLOBETROTTER results suggest longitudinal (west/east) gene flow in the northern of Peru.

Section 3: Cultural interactions were accompanied by gene flow events across the Andes

3.1. Introduction

Further evidence of commercial interaction between populations of the Coast, Andes and Amazon of fertile Northern Peru: Archaeological evidence suggests an intensive interaction across

the Fertile Andes in contrast to the Arid Andes. Cultural and commercial interactions involve the South Pacific Coast, Andean and Amazonian populations (3, 4, 27, 28). Archaeological data point out that the Fertile Andes (involving Southern Ecuador and Northern Peru) was a main crossroads for the Andes-Amazon interaction (27), which could be facilitated since the Andean mountain chain has its lowest altitude in this region (29, 30). These cultural and commercial interactions involved the trade of Spondylus shells from the Coast and medicinal plants and herbs from the Amazon. Particularly, Chachapoyas, an Amazon Yunga, was part of a significant interregional exchange network during the Early Intermediate period (around 2300 YBPto 1400 YBP) and that extended to the Inca period (around 1500 AD) (1). Also, it was suggested a socioeconomic exchange between Moche (Coast) and Awajun (Amazon Yunga) natives (3). Nowadays, Chachapoyas and Lamas populations both living in the Amazon Yunga speak Quechua, which could be adopted as a lingua franca in the last centuries. Specifically, Lamas population is an intriguing case since it was attributed to have an Andean Origin (related to Chanka population). A recent study demonstrated that Lamas population has a closer genetic affinity to surrounding Amazon populations than with Chanka population, suggesting a possibly Amazonic origin instead of Andean (31), which is consistent with our results in this paper (Figs S2-S13). In the latter section, we showed that Jivaroan populations were more similar to Fertile Andes populations (SPPC and Chachapoyas) than to Arid Andes populations. This genetic proximity could be related to the historical interactions. Here we address:

Question 2: whether gene flow accompanied the cultural and socioeconomic interactions between Andean and Amazon Yunga populations?.

To address this question we performed Patterson's statistics analyses: *D* statistics and Admixture graphs analyses on the masked data. The *D* statistics determine if two populations have an excess of alleles sharing due to gene flow. The admixture graphs explore the best model of relationships among populations taking into account gene flow events in a statistical framework.

3.2. Methods

3.2.1. *D* statistics

The *D* statistics (15, 32), or ABBA-ABAB test, is a method to detect gene flow among closely related populations. It evaluates four populations: P_1, P_2, P_3 and an *Outgroup*:

$$Outgroup\ (P_3\ \ (P_1,\ \ P_2))$$

This treeness test considers, as a null hypothesis, that P_1, P_2, P_3 and an outgroup have a tree relationship. In this null hypothesis, P_1 and P_2 diverged earlier from the ancestor of P_3, with no gene flow between P_3 and P_1 or P_2 after the divergence. The alternative hypothesis is that P_3 was involved in gene flow with P_1 or P_2 after the divergence. This analysis is restricted to biallelic sites and considered an allele "A" as the ancestral allele of the outgroup (O) and an alternative allele "B" in P_3. Considering the order O-P_3-P_1-P_2, the *D* test is focused on the ABBA or ABAB pattern. The first pattern (ABBA) corresponds to the total number of sites with the alternative allele (B) shared only by P_1 and P_3. The second configuration (ABAB) is the total number of sites with the alternative allele (B) shared only by P_2 and P_3. The *D* statistic is calculated by the relationship of: the difference of ABBA-ABAB counts (numerator) and the total count ABBA+ABAB (denominator). The numerator of this relationship indicates the signal of the statistic which is interpreted as the direction of gene flow. If the *D* statistic is not significantly different from zero, we accept the null hypothesis of treeness. If the result differs significantly from zero, we reject the null hypothesis and consider the possibility of gene flow between P_3 with P_1 or P_2. For *D* statistics estimation we used ADMIXTOOLS (16). The results are interpreted as follows: negative values of *D* are interpreted as gene flow between P_1 and P_3 and positive values indicate a gene flow between P_2 and P_3. A *D* value is considered statistically significant if the absolute value of the relationship between the *D* value and its standard deviation is equal or above 3 ($|Z\text{-score}| \geqslant 3$). We applied D statistics to the masked Datasets 500K and 230K. Considering the huge number of combinations for *D* inferences and even obtaining highly significant values, we construct Q-Q plots to analyze the Z-score distribution, which is expected to be approximately normal under the hypothesis that our results could be expected by chance (H_0) (33).

Considering the divergence between Western (including Andean, SPPC) and Eastern (Amazon) groups, we tested the following configurations:

- (Outgroup, Eastern (Western$_{2,}$ Western$_{1}$))
- (Outgroup, Western (Eastern$_{1}$, Eastern$_{2,}$))

Outgroup: For both masked datasets, we used Africans (YRI).

Natives_500K Dataset (Masked)

Configuration **(Outgroup, Western (Eastern$_{2}$, Eastern$_{1}$)):**
We explored the gene flow among populations in South America. This configuration explores if one western population shares more alleles with one population from the Amazon (Amazon Yunga or Lower Amazon) than another. Deviation from the diagonal of the Q-Q plot indicates that some populations in this configuration are involved in gene flow events. We observed that the major signals of gene flow involved Chachapoyas, Lamas and Jivaroan populations (Awajun and Candoshi) with SPPC (Fig. S14A).

Configuration **(Outgroup, Eastern (Western$_{2}$, Western$_{1}$)):**
When we explored the possible signal of gene flow from Eastern (Amazon yunga and Lower Amazon) to Western, we found results similar to the first configuration, a significant deviation from the diagonal involving SPPC, with Chachapoyas, Lamas and Jivaroan populations (Fig. S14B).

Natives_230K Dataset (Masked)

For this dataset, both configurations showed congruent results with the Dataset 500K (Fig. S15A - B) and no new signals of gene flow appeared.

In conclusion, regarding **Question 2: whether gene flow accompanied the cultural and socioeconomic interactions between Andean and Amazon Yunga populations**?, *D* statistics show evidence of longitudinal gene flow between the north Coast of Peru (part of the so-called northern Fertile Andes) and Amazonian populations of similar latitudes.

3.2.2. Admixture graphs

Rationale: Population Structure Analysis (Section 2) and *D* statistics showed high genetic affinity among Native American groups of Fertile Andes. Strong signals of *D* statistics involved SPPC and Jivaroan populations, Chachapoyas and Lamas suggesting gene flow among these groups (Fig S14-S15). In order to determine the direction and parameters (contributions) of the gene flow or admixture events we applied admixture graphs using qpGraphs in ADMIXTOOLS (16). qpGraph evaluates the fit of *a priori* suggested tree and the *f*-statistics applied on a set of populations included in the tree.

We test two contrasting hypotheses of gene flow. **The first one** (W→E), Amazon groups (Eastern) receives a contribution from the Coast (Western) and **the second one** (E→W), Coast (Western) receives a contribution from Amazon (Eastern). As a result, qpGraph offers a log-likelihood of the fit, *Z*-score values for the *f*-statistics and the parameters involved in admixture events (contributions) if these are tested. We accept an hypothetical tree if the absolute value of the highest *f*-statistic is less than 3 (|Z-score| < 3). Moreover, if we obtain trees with similar Z-score values, we select the tree with fewer zeroed branches and lower log-likelihood.

First, we selected a small group of populations from the masked Natives 500K dataset and we merged it with the MA1 sample (34). Our final dataset included 235,402 SNPs. The MA1 sample is important due its relationship with one of the dual ancestries that gave origin to Native Americans (34). To create a draft tree in which admixture events will be fitted, we select the following populations:

African:	YRI
Siberian:	MA1

South Asian:	Onge
East Asian:	CDX
North America:	Clovis
Mesoamerican:	Mixe
Amazon:	Ashaninkas and Matses
Andes:	Uros

The second step was to add the Tallanes population from the Peruvian Coast (Western) and a Northern Amazon population (Awajun, Candoshi, Lamas, Chachapoyas). We test each of these populations as unadmixed and admixed. We selected Tallanes since this population showed the most of the highest values of *D* statistics and therefore, was likely involved in gene flow. We tested our hypotheses for 4 combinations:

-Tallanes-Awajun
-Tallanes-Candoshi
-Tallanes-Lamas
-Tallanes-Chachapoyas

Results: All graphs (data not shown) that fit Tallanes and Northern Amazonas as unadmixed showed poor fit. When we allowed for gene flow considering our two contrasting hypotheses, the best fit was for the hypothesis 1 (W→E) (Fig S16A,S17A,S18A,S19A). Admixture graphs for hypothesis 1 and 2 showed the same value for the worst *f*-statistic (Z-score), but hypothesis 2 (E→W) included zeroed length branches, except for the Tallanes-Chachapoyas combination. In this last combination the hypothesis 1 has the best fit with the lower Z-score (Fig S19A). These admixture graphs support a predominant contribution from Coast related populations to the North Amazon.

3.3. Conclusion

- Populations around the Fertile Andes were involved in gene flow events. Specifically, North Coast populations showed a significant contribution to the genetic ancestry or North Amazon populations.

Section 4: Dating the between-population homogenization of the arid Andes

Question 3 (Section 4): when the Andean between-population genetic homogenization started in the context of the arid Andean chronology.

4.1. Methods

4.1.1 Identical-by-descent segment analysis

We analyzed the pattern of segments identical-by-descent (IBD) to infer the relationship among populations across the time. If two DNA segments are identical and have the same ancestral origin they are considered Identical-by-descent (35, 36). From one generation to another, large segments of DNA are inherited, but in successive generations recombination events break these regions (37). The relationship between the size of an IBD segment found between two individuals and the time in generations until coalescence have the following approximation (38, 39):

$$E \backsimeq 3/2L$$

Where:
E: time in generations to the most recent common ancestor.
L: length of IBD segments (in units of Morgan).

To infer the pattern of gene flow along the time, we ran RefinedIBD software (40) with the **Natives 1.9M Dataset** and **Natives 500K Dataset**. To analyze the demographic evolution in Central Andes, we used IBDne software (41) with the **Natives 1.9M Dataset**, both approaches are described below.

RefinedIBD

To infer IBD segments, we used RefinedIBD (40). This software performs two steps: first, it uses the GERMLINE algorithm (42) for IBD detection, and second, a refinement step, that calculates the probability of each segment to be IBD (40). We removed all missing data in the specific dataset selected for this analysis using PLINK (--geno parameter). We used the genetic map GRCh37 from HapMap and we restricted our analyses to segments larger than 3.2cM. We organized the IBD segments in four intervals that could be related to historical periods, considering one generation as a period of 28 years:

1) 3.2 to 4.2cM (50 to 36 generations before present)
2) 4.2 to 7.8cM (36 to 19 generations before present)
3) 7.8 to 9.3cM (19 to 16 generations before present)
4) all segments greater than 9.3cM (16 generations before present to present day)

The first interval is related to pre-Inca times, more specifically to the Middle Horizon and Late Intermediate, that correspond to the Wari-Tiwanaku Empire. The second interval involves the rise and fall of the Inca Empire. Finally, the last interval is related to colonial times until the present day (Fig. 2, Fig. S20).

We calculated the average amount of shared DNA between two individuals from the same (aaIBD) or different populations (abIBD), for each interval(40). Considering a specific pair of populations (a and b), we calculated the total amount of shared DNA between one sample from “a” and another from “b”. After that, we sum all pairwise values and divided by the number of pairs between a and b:

$$\text{abIBD} = \frac{\Sigma_{ij} L_{ij}}{Npairs}$$

Where:
abIBD: average of the total shared IBD length between two individuals from different populations (or the same population if it is aaIBD).
i and *j*: the two individuals.
L: total IBD length shared between each pair of individuals
Npairs: Na * Nb (for different populations), and Na(Na-1)/2 (For same population). Where N is the number of individuals in the respective population (a or b).

The representation of IBD relationships was presented as a similarity heatmap constructed with the log of the abIBD values (Fig. 2, Fig. S21).

Natives 1.9M dataset

In the first interval (3.2 to 4.2 cM, Fig. 2B) it is possible to observe homogeneous patterns among Andean populations. We did not observe differences between the intra and interpopulation sharing in the arid Andes. In a temporal view, this interval coincides with the Middle Horizon (43) that included the expansion of Tiwanaku-Wari and its falling. This society, which dominated the political landscape of the central highlands of the Andes, was probably an ancestor of Quechua-speaking populations (44), which may be related to the fact that 3 of the 6 Andean studied populations speak this language today. Posteriorly, the difference between intra and interpopulational sharing ratio for Andean populations gradually increased until the most recent interval, but remains smaller than other groups. However, the hypothesis that the Andean homogenization already existed before the Incas was evidenced by the visualization of the high degree of sharing of IBD segments between these groups during the Tiwanaku-Wari expansion.

Natives 500K dataset

In the earliest interval (Fig. S21A), the Andean region already appears homogeneous, corroborating the **Natives 1.9M dataset** results. The SPPC populations showed high internal affinity degrees. The Amazon group in general has some relations with other groups, but the diagonal is very intense, evidencing its high degree of intrapopulation IBD. In the second interval (Fig. S21B), corresponding to the period between the falling of Tiwanaku-Wari Empire and the beginning of the Inca Empire, the arid Andes stay homogeneous. The next period (Fig. S21C) comprises the entire duration of the Inca Empire, which remains, in general, homogeneous. In the last interval (Fig. S21D), after the Europe conquest, Andes is apparently more structured. SPPC populations remain connected since the first interval, as do Matses and Lamas. Like the first dataset, the genetic affinity between Chachapoyas and Andean and SPPC is constant along the intervals.

4.1.2. IBDne

To understand the demographic dynamics of populations in the arid Andean, we calculated the pattern of effective population size (Ne) with software **IBDne** (41). This algorithm infers the pattern of Ne along the generations, allowing us to study how demographic changes make the genetic diversity of a population vulnerable to genetic drift. This method has some particularities that need to be taken into account: 1) it tends to smooth over sudden changes in Ne, 2) it assumes a closed population, 3) it assumes a homogenous population. For this reason we performed the analysis just for the arid Andean group. To avoid the underestimate of effective population size, we restricted the analysis to segments larger than 4cM, as suggested by the authors for array data (41). We inferred this parameter (Ne) only between 4 and 50 generations before the present, because segments related to the last 3 generations are not informative for the dynamic of the population. As our arid Andean populations are genetically homogeneous, we grouped as a unique population for the IBDne inference, which would not be acceptable for the other groups. Moreover, as the density of SNPs is also an important factor for these inferences, we only applied this method for **Natives 1.9M dataset**.

Natives 1.9M dataset

In the earliest heatmap interval, approximately between 50 to 36 generations before present, we can see an expansion period in the population effective population size. After approximately 27 generations, the Ne decreased (Fig. S22), which can mean a bottleneck or continuous population structuring, this reduction stopped in the last 10 generations.

4.2. Conclusions

- The Andean homogenization already existed before the Late Intermediate. Probably related to the Tiwanaku-Wari expansion.

.

- Inferences on the dynamics of the effective population size based on IBD suggest that the decline in population size that followed the European conquest (~1500 AD) affected the genetic diversity of Andean populations, making it more vulnerable to be affected and lost by genetic drift.

- The effective population size (Ne), estimated from IBD segments, shows the dynamics (Fig. S22) characterized by a Post-Contact decline to around one-third the level observed around 1250 years ago (Middle Horizon), when it was rising likely due to an increase in population size in the arid Andean regions.

Section 5: Genetic Differentiation and Natural Selection in the Andes and Amazon

5.1. Introduction

The evolutionary mapping of genetic variants is an efficient approach to identify functional genomic regions that have played an essential role in survival, and possibly have consequences for human health (45, 46). The evolutionary history of modern humans is marked by major migration events for environments with different climates, diets, and diseases (47). These factors compose the selective pressures that act on variants that affect biological mechanisms that influence the adaptation process (48, 49). The process of natural selection leaves genetic signatures that can be detected, making it possible to identify regions of the human genome related to these mechanisms.

In the following section, we applied statistical methods based on population differentiation (Population Branch Statistic - PBS) and linkage disequilibrium (cross population extended haplotype homozygosity - xpEHH) to identify genomic regions under natural selection in Andean and Amazon populations. For this purpose, we used the **Natives 1.9M dataset** considering only the following populations organized in **two groups**:

1) **Arid Andean group:** Chopccas, Quechuas_AA, Qeros, Puno, Jaqarus, and Uros.We excluded 2 Quechua individuals who had more than 10% non-native ancestry according to the ADMIXTURE analysis.

2) **Amazon group :** Ashaninkas, Matsiguenkas (including Matsiguenkas 1 and 2), Matses and Nahua. We did not include Awajun, Candoshi, Lamas and Chachapoyas in this analysis because our previous results (Section 2 and 3) demonstrated that these populations were involved in gene flow and this may mask differentiation signals.

5.2. Methods

Natural Selection Candidate SNPs

5.2.1. Population Branch Statistic

PBS is a statistical test to identify changes in the allele frequencies of a target population since its divergence from an ancestral population. PBS is based on the comparison of differentiation (F_{ST}) values among three groups: 1) the target population; 2) a population closely related to the target, and 3) an outgroup (50).

Before the PBS analysis we applied a MAF (Minimum Allele Frequency > 0.05) filter with PLINK. Since we are searching for evidence of differentiation between Andes and Amazon, we considered only SNPs with low differentiation inside these groups (F_{SC}<0.15) (51). The F_{SC} for each SNP for each group was estimated with varcomp function from the hierfstat R package (52). 4P software (53) was used to calculate F_{ST} for each SNP. The F-statistics estimated through varcomp function and 4P rely on the Weir and Cockerham (1984) algorithm (54). Subsequently, the F_{ST} values were transformed as following (55):

$$F_{ST}T = -\log(1-F_{ST})$$

To the transformed F_{ST} values, we applied the PBS formula (50):

$$PBS = (F_{ST}T1+F_{ST}T2-F_{ST}T3)/2$$

Where:
$F_{ST}T1$: transformed F_{ST} between the target population and the closely related population.
$F_{ST}T2$: transformed F_{ST} between the target population and the distant population.
$F_{ST}T3$: transformed F_{ST} between the close population and the distant population.

To avoid spurious outliers when the branches were long or short in all groups, we applied a normalized version from PBS (56):

$$PBSn = PBS1 / (1 + PBS1 + PBS2 + PBS3)$$

Where:
PBSn: normalized PBS.
PBS1: estimated PBS when the PBS is calculated for the target population.

PBS2: estimated PBS when PBS is focused on the closely related population.
PBS3: estimated PBS when PBS is focused on the distant population.
Our final result is based on the PBSn.

We performed PBS with the following configurations: 1) Andes as a target group, Amazon as a closely related group; and 2) Amazon as a target group, and Andes as a closely related group; in both approaches the CDX (Chinese Dai in Xishuangbanna, China), a population from 1000 Genomes (57) was used as an outgroup. We analyzed the results in windows of 20 SNPs with 5 SNPs of overlap. To determine the probability that a PBS value occurs under the null hypothesis of genetic drift, we simulated 10,000 chromosomal regions of 1Mb under the neutral model for the three populations involved (Andes, Amazon and CDX) (Fig. S23) using the Recosim program to simulate the recombination maps and Cosi2 (58) to simulate the genetic data under a neutral model as described:

```
######################## NEUTRAL MODEL ########################

#DETAILS: In this model the split in Native Americans is Andean (source) and Amazon (new population)
#Andean and Coast events are based on the inference of Ne performed on IBDNe based on IBD segments
#split <label> <source pop id> <new pop id> <T>  516 generations ~ 12900 Andes-Amazon 12700 AndesCosta years
estimated by Harris et al. 2018 (1 generation = 25 years)

gene_conversion_relative_rate 0.0000000045

# mu,
mutation_rate         1.5e-8
length 1000000
# population info
# for each population, include a line:
# pop_define pop-index pop-label

pop_define 1 amazon
pop_define 2 andean
pop_define 3 asian
pop_define 4 coast

#init sample pops
# for each sample set, include
# pop_size pop-label pop-size
# sample_size pop-label sample-size

#amazon
pop_size 1 2749
## Ne is the mean of three values obtained for Matses population in Harris et al. 2018 (N1=2848,N2=2881,N3=2518)
sample_size 1 206
## 206 Considering 103 samples

#andean
pop_size 2 8064
## Ne is the mean of six values obtained for Chopccas  (N1=7774,N2=7070,N3=9348), populations in Harris et al. 2018
sample_size 2 166
## 166 considering 83 diploid samples

#asian
pop_size 3 7700
sample_size 3 240
# 240 considering 120 diploid samples

#coast
pop_size 4 6975
sample_size 4 62
# 62 considering 31 diploid samples

pop_event exp_change_size "Andean second expansion" 2 4 9 8064 2500
pop_event bottleneck "Andean bottleneck due to European conquest" 2 29 0.067
pop_event bottleneck "Coast bottleneck due to European conquest" 4 29 0.067
pop_event bottleneck "Amazon bottleneck" 1 479 0.067
pop_event exp_change_size "Andean expansion" 2 30 450 7000 2426
pop_event exp_change_size "Coast expansion" 4 4 9 6975 1500
pop_event split "andean and amazon split" 2 1 516
pop_event split "andean and coast split" 2 4 508
pop_event bottleneck "native bottleneck" 2 959 0.067
```

```
pop_event split "asian and native split" 3 2 960
pop_event bottleneck "asian bottleneck" 3 1998 0.067

random_seed 2022747205
######################## END OF FILE ########################
```

After this, we estimated the PBSn values for the simulated data with the same methodology used for empirical data. For each observed PBSn result, we calculated the p value as a proportion of simulated PBSn values that are equal or greater than the observed value (50). We considered as candidates for natural selection those SNPs in the 0.05% higher values of PBSn (PBSn > 0.150 for the Andes and PBSn > 0.191 for the Amazon) that were encompassed in the windows in the 0.05% higher PBSn mean values (PBSn mean > 0.095 for the Andes and PBSn mean > 0.116 for the Amazon). We found 142 signals comprising 16 genes in the Andes and 137 signals comprising 15 genes in the Amazon (Tables S1, S2; Fig. 3).

5.2.2. xpEHH: cross population extended haplotype homozygosity

Positive selection events increase the frequency of a genetic variant and, consequently, the frequency of the variants around it (59). This process occurs faster than the haplotypes are broken down by recombination, leading to the emergence of an unusual high frequency long-range haplotype. Considering this, we decided to perform an extended haplotype homozygosity (EHH) test to select the most likely candidates for natural selection.
Sabeti *et al.* (60, 61) developed methods to detect natural selection signatures calculating the EHH, defined as the probability of finding homozygosity of all SNPs around the haplotype of interest choosing two random chromosomes containing this haplotype in a population:

$$EHH\left(x_i\right)=\sum_{h\in C(x_i)}\frac{\binom{n_h}{2}}{\binom{n}{2}}$$

Where $C(x_i)$ is the number of all possible distinct haplotypes considering the extension from de core SNP to the i-th SNP, and n_h is the number of observed haplotypes of a specific type h (62).

The method xpEHH defines a core SNP and calculates the EHH for all SNPs in 1MB of distance forwards and backwards considering the chromosomes of two target populations, A and B. When the EHH decays to 0.03-0.05 before reaching 1MB of distance this point is defined as SNP X, if this score is not reached in this range the core SNP is discarded from the next analyzes. Next, the populations are separated and the EHH parting from the core SNPs selected in the first step is calculated again until it reaches the value of 0.03-0.05 (SNP X) in each population. Then the integral of the EHH in respect to the distance from the core SNP to the SNP X is calculated giving the results called I_A (for population A) and I_B (for population B). The xpEHH log ratio is defined as $\ln(I_A/I_B)$. The results are genome-wide normalized. Extreme positive values are indicative of selection in population A, and negative values in population B. The xpEHH analysis was performed with the software Selscan [62].

We considered as positive signals for natural selection the SNPs representing the 99.5 percentile of the xpEHH results of an Andean vs Amazon comparison (xpEHH > 2.97 for the Andes and xpEHH < -3.34 for the Amazon). For each observed xpEHH result, we calculated the empirical p value as the proportion of values that are equal or greater than the observed value. Only concordant results between PBS and xpEHH were considered as strong candidates for natural selection, with this approach we found 22 candidate SNPs comprising 3 genes in the Andes and 21 SNPs comprising 1 gene in the Amazon (Tables S3, S4; Fig. S26, S27).

To assess whether the results obtained from xpEHH analysis between Andean and Amazon populations are corroborated when comparing each group with an outgroup, we performed xpEHH analysis between each group and East Asian populations from 1000 genomes ($xpEHH_{ANDvsEAS}$ and $xpEHH_{AMZvsEAS}$). The results for Andean populations show a high score for gene HAND2-AS1 (higher

$xpEHH_{ANDvsEAS}$=1,59, p-value=0.001), but not for gene RARS (higher $xpEHH_{ANDvsEAS}$=0,79 p-value=0,105). For Amazon populations, the highest signal was from an intergenic region near the gene PTPRC (rs1326288 higher $xpEHH_{AMZvsEAS}$=-1,23 p-value=0.026)(Tables S3, S4).

The candidate loci for natural selection were annotated with MASSA (Multi-agent Annotation System) (63), that mines the following datasets for SNPs (based on rs code): dbSNP (64), OMIM (65) [Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 2020. World Wide Web URL: https://omim.org/], Reactome (66) , HGNC (HGNC Database (67)), GWAS Catalog (68), PolyPhen2 (69), Provean (70), SIFT (71); and the following datasets for genes: UCSC (72), Gene Ontology (73, 74) , PharmGKB (75).

It is interesting to note that the strongest signal for PBS analysis for the Andean populations was from the *DUOX* genes (Fig. S24), previously suggested as a candidate gene for natural selection by Jacovas *et al*. (76). However, this signal does not appear in the xpEHH results.

5.2.3 Linkage Disequilibrium Patterns in natural selection signals

To find other possible candidate SNPs under natural selection, we look for linkage disequilibrium (LD) patterns associated with our strongest natural selection signals. We had access to sequencing data from 60 Andean native individuals (30 Chopccas and 30 Uros) for the genes *DUOX2* and *HAND2-AS1* from Harris et al. (77) . The Amazon sample (12 individuals) was not large enough to allow LD inferences. We calculated LD using the software Haploview 4.2 version (7821). We only consider in LD those variants with $r^2 \geq 0.80$. We found 37 no genotyped SNPs in LD with our two missense signals of natural selection (rs269868 and rs57659670) in the *DUOX2* region in chromosome 15 including three missense mutations: one in the *DUOX2* gene (rs2001616: G>A,T: Pro138Leu), one in the *DUOXA2* gene (rs2252371: C>T: Pro126Leu), and one in the *DUOXA1* gene (rs61751061: C>G,T: Arg478Pro). The SNP rs2001616 is located in the Peroxidase Homologue Domain (aa 26-601) of DUOX2 protein, and rs2252371 is located in an extracellular strep of the DUOXA2 protein (aa 78-183). In these domains occur disulfide bridges between specific cysteines that are essential to the stability and function of the DUOX complex (7989).

In the gene *HAND2-AS1* we found 23 no genotyped SNPs in LD with our 4 natural selection signals, all of them in intronic regions, including rs3775587, mapped within a putative enhancer (Fig. S28). These SNPs were used in the analysis of regulatory elements described below.

5.2.4 Identification of regulatory elements located around HAND2-AS1 locus

GeneHancer track available in the UCSC Genome Browser (90 80) was used to identify active regulatory elements (enhancers and promoters) that may target *HAND2-AS1* (Fig. S29). GeneHancer database was created by integrating >1 million regulatory elements from seven genome-wide databases: ENCODE project Z-Lab Enhancer-like regions (version v3); Ensembl regulatory build (version 92); functional annotation of the mammalian genome (FANTOM5) atlas of active enhancers; VISTA Enhancer Browser; dbSUPER super-enhancers; Eukaryotic Promoter Database (EPDnew) promoters; and UCNEbase ultra-conserved noncoding elements. Genes were linked to enhancers by GeneHancer using five methods: eQTLs from GTEx (v6p); Capture Hi-C promoter-enhancer long range interactions; FANTOM5 co-expression of enhancers in the form of noncoding enhancer RNA; transcription factor co-expression; and gene target distance. For this analysis a “double elite” dataset was considered, which is composed of regulatory elements derived from more than one database (elite enhancers) that are associated to genes from more than one method (elite association).

Highly Differentiated Variants Between Andean and Amazon Populations and its Medical Relevance

5.2.5 F-statistics

We searched for functionally relevant SNPs differentiated between the Arid Andes and Amazon populations of similar latitudes (i.e. the same groups of populations tested for natural selection) using

the classical F statistics for each SNP as defined by Weir and Cockerham (54). These SNPs are differentiated between the two groups not necessarily due to the action of natural selection, but their sharp differences in frequencies in the Andean vs. the Amazon environments may be biomedically relevant. We found 1,985 highly differentiated SNPs between the two groups of populations (0.1% highest values of FCT distribution: > 0.318, estimated with the 4P software19), but relatively homogeneous within the groups (FSC<0.15, estimated with hierfstat software20). We annotated the 1,985 SNPs usIng our bioinformatics MASSA platform (63), that mines the following datasets based on SNPs rs code): dbSNP (64), Ensembl (81), GWAS Catalogue (68), PharmGKB (75), SIFT (71), PolyPhen (69). SNPs that are GWAs hits, related to drug response and missense mutations, are listed in Tables S5-7.

5.3. Conclusions

- We have confirmed a natural selection signal from a gene previously reported in Andean populations, *DUOX2* (76) (PBSn=0.22 p-value=0.002, xpEHH=-2.647 p-value=0.991).

- We identified Natural selection signals Andeans in genes related to (Tab. S4):
    - High altitude adaptation: ***SULT1A1***: PBSn=0.167 p-value=0.007; ***RARS***: PBSn=0.15 p-value=0.010, xpEHH=2.980 p-value=0.0025 (82, 83),
    - Heart development: ***HAND2-AS1***: PBSn=0.21 p-value=0.003, xpEHH=4.481 p-value<2e-5 (84),
    - Immune response: ***UBQLN4***: PBSn=0.17 p-value=0.007, xpEHH=-0.217 p-value=0.607; ***SSR2***: PBSn=0.17 p-value=0.007, xpEHH=-0.215 p-value=0.606; ***DUOX2***: PBSn=0.22 p-value=0.002, xpEHH=-2.647 p-value=0.991 (85–87).

- We identified Natural selection signals in Amazon populations related to (Tab. S5):
    - Immune response: ***PTPRC***: PBSn=0.265 p-value=0.004, xpEHH=-4.222 p-value=0.0003 (88),
    - Food intake regulation: ***MCHR1***: PBSn=0.26 p-value=0.004 (89),

    - Lipid transport: ***ABCA9***: PBSn=0.21 p-value=0.008, xpEHH=-1.570 p-value=0.060, ***ABCA6***: PBSn=0.19 p-value=0.011, xpEHH=-1.362 p-value=0.084 (90, 91).

Supplementary Figures

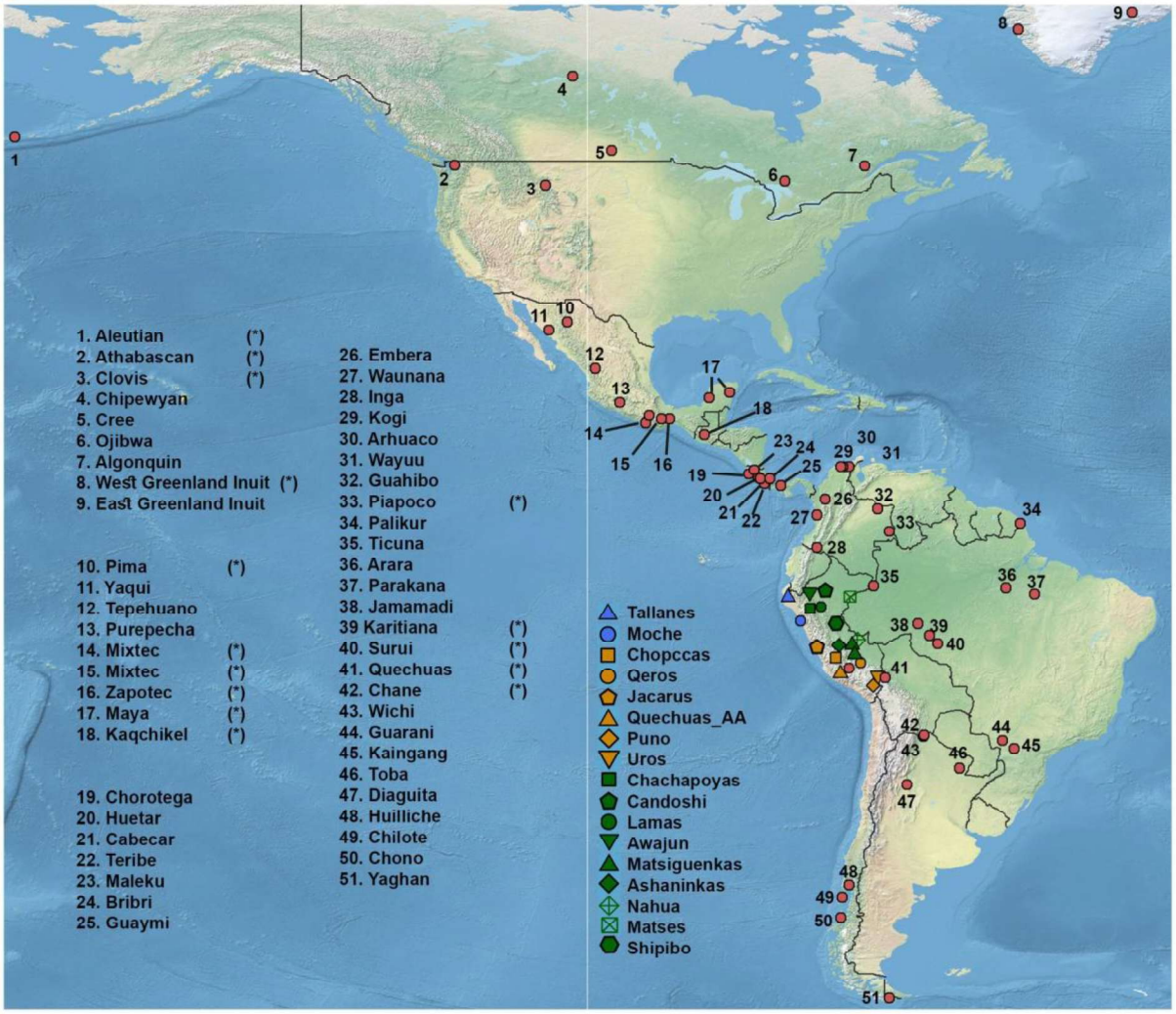


Figure S1. Geographical distribution for the 18 Peruvian Native populations sampled, plus the 65 sampled Native American populations and public data sets (Mallick *et al.* 2016, Raghavan *et al.* 2015, Reich *et al.* 2012). All samples except Clovis and Athabascan were included in a data set of ~ 230,000 SNPs. Peruvian samples and (*) were included in a data set of ~ 500,000 SNPs.

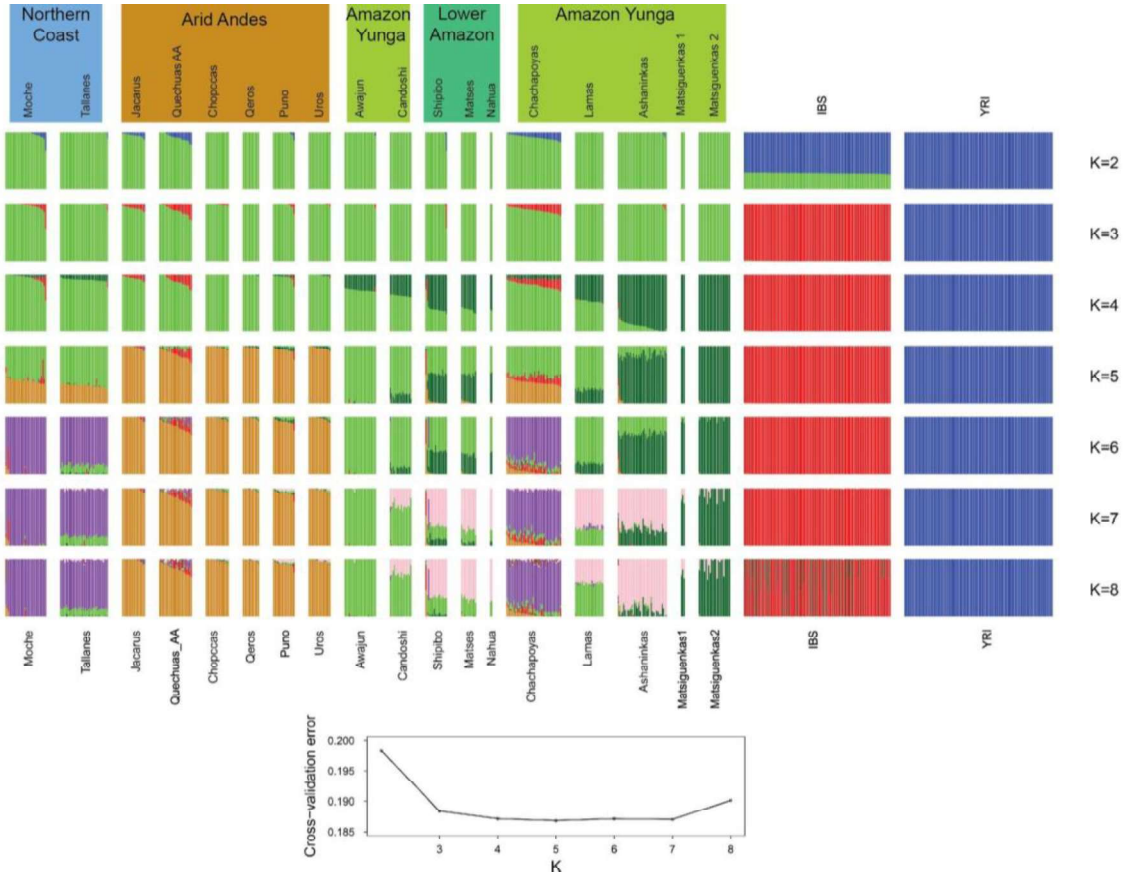


Figure S2. ADMIXTURE analysis for 18 Native American populations, as well as Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 1.9M Dataset). Figure shows results for 2 to 8 ancestral clusters (K) and a plot (Bottom) with the ADMIXTURE cross-validation errors as a function of K. The lowest cross validations error corresponds to K=5 in which we observed four Native American, one European and one African cluster.

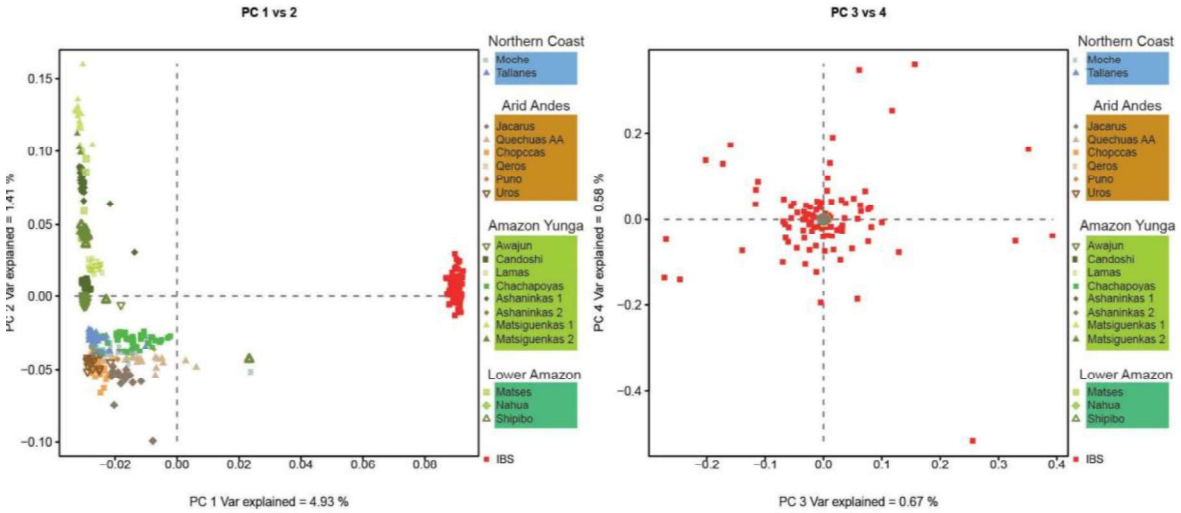


Figure S3. Principal Component Analysis for 18 Native American Peruvian populations and Iberian individuals (IBS) from 1000 Genomes Project (Natives 1.9M Dataset). Shades of blue are related to Coast populations. Orange-brown colors are related to Andean populations and green colors are related to Amazon.

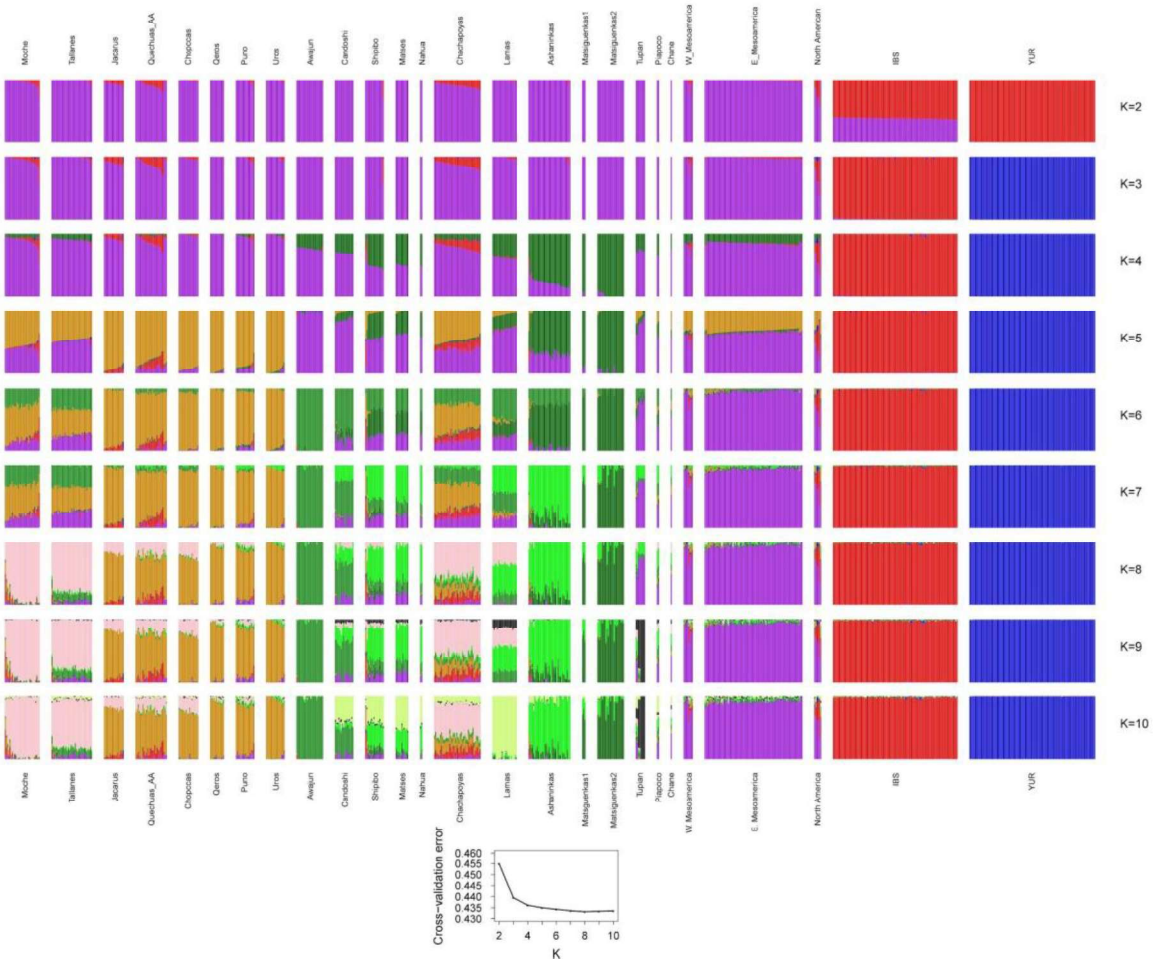


Figure S4. ADMIXTURE analysis for 18 Natives American Peruvian populations, Guatemala samples, Native Americans from Raghavan *et al.* 2015 and the Simons Project (Mallick *et al.* 2016) Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 500K Dataset). Figure shows results for 2 to 10 ancestral (K) clusters and a plot (Bottom) with the ADMIXTURE cross-validation errors as a function of K. The lowest cross validation error corresponds to K=8.

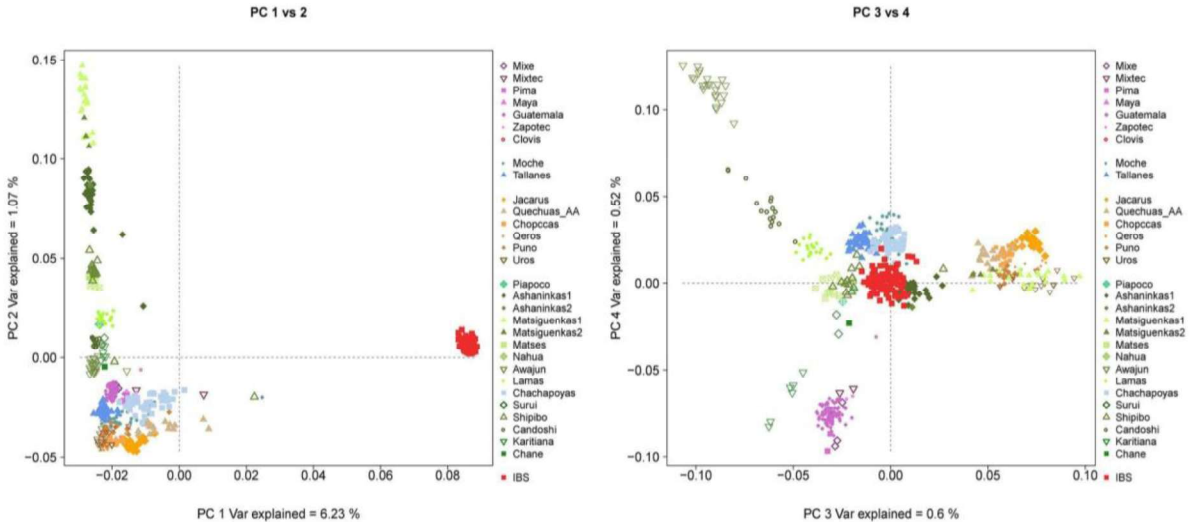
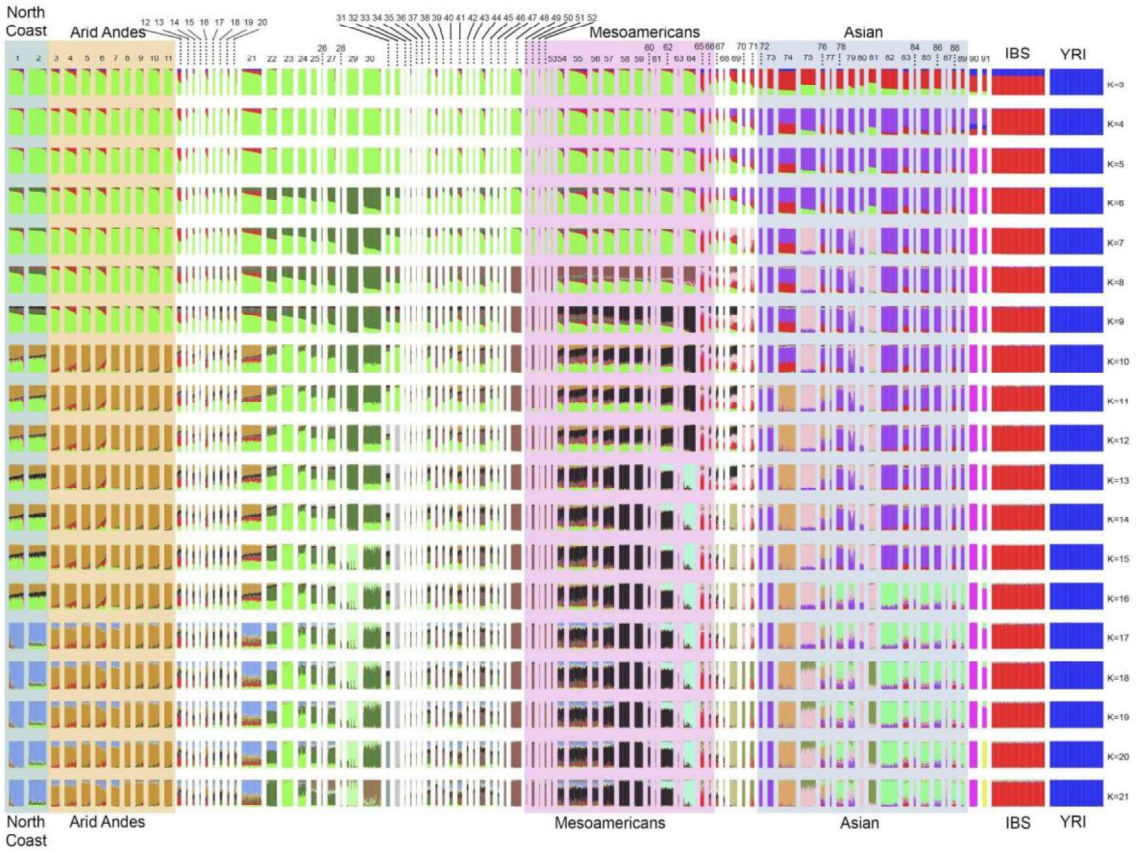
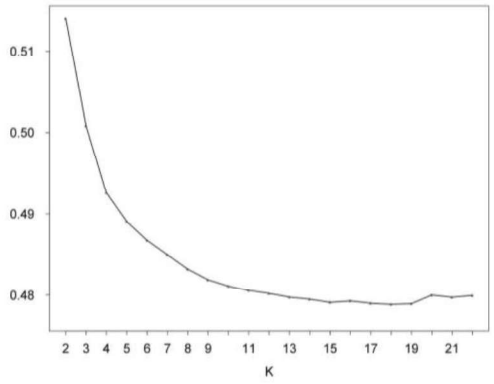


Figure S5. Principal Component Analysis for 18 Native American Peruvian populations, Guatemala samples, Native Americans from Raghavan *et al.* 2015 and the Simons Project (Mallick *et al.* 2016) and Iberian (IBS) populations from 1000 Genomes Project (Natives 500K Dataset). Shades of blue are related to Peruvian Coast populations. Orange-brown colors are related to Andean populations and green colors are related to Amazon. Shades of purple are related to Mesoamericans. Shades of beige are related to North American natives.



1. Moche
2. Tallanes
3. Jacarus
4. Quechuas
5. Quechua_R1
6. Quechua_R2
7. Chopccas
8. Qeros
9. Aimaras
10. Aimaras_R
11. Uros
12. Chilote
13. Chono
14. Hulliche
15. Chane
16. Diaguita
17. Toba
18. Wichi
19. Yaghan
20. Guarani
21. Chachapoyas
22. Lamas
23. Awajun
24. Candoshi
25. Matses
26. Nahua
27. Shipibo
28. Matsiguenkas
29. Shimaa
30. Ashaninkas
31. Karitiana
32. Surui
33. Arara
34. Kaingang
35. Parakana
36. Jamamadi
37. Piapoco
38. Arhuaco
39. Embera
40. Guahibo
41. Inga
42. Kogi
43. Waunana
44. Wayuu
45. Ticuna
46. Palikur
47. Bribri
48. Cabecar
49. Chorotega
50. Guaymi
51. Huetar
52. Maleku
53. Teribe
54. Kaqchikel
55. Maya1
56. Maya2
57. Zapotec1
58. Zapotec2
59. Mixe
60. Mixtec
61. Purepecha
62. Tepehuano
63. Yaqui
64. Pima
65. Aleutian
66. Cree
67. Ojibwa
68. Algonquin
69. Chipewyan
70. EastGreenland
71. WestGreenland
72. Nganasan1
73. Nganasan2
74. Khanty
75. Chukchi



76. Selkup
77. Tundra_Nentsi
78. Tuvinians
79. Yukaghir
80. Koryak
81. Naukan
82. Yakut
83. Altaian
84. Dolgan
85. Buryat
86. Evenki
87. Ket
88. Mongolian
89. Mongola
90. Papuan
91. Melanesian
92. IBS
93. YRI

Figure S6. (Top) ADMIXTURE analysis for 90 worldwide populations including 71 Native American populations, 18 Asian, 2 Oceanian, Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 230K Dataset). Figure shows results for 3 to 21 ancestral clusters (K). (Bottom) ADMIXTURE cross-validation errors as a function of K and list of populations included. The lowest cross validation corresponds to ADMIXTURE K=18.

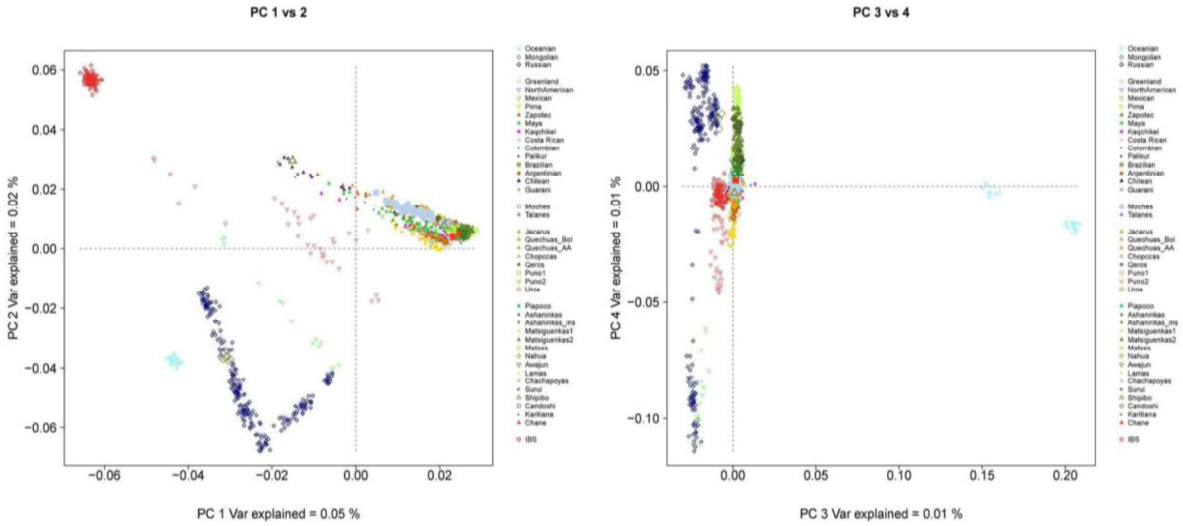


Figure S7. Principal Component Analysis for 89 worldwide populations including 68 Native American populations, 18 Asian populations, 2 Oceanian populations, Iberian (IBS) populations (Native 230K Dataset).

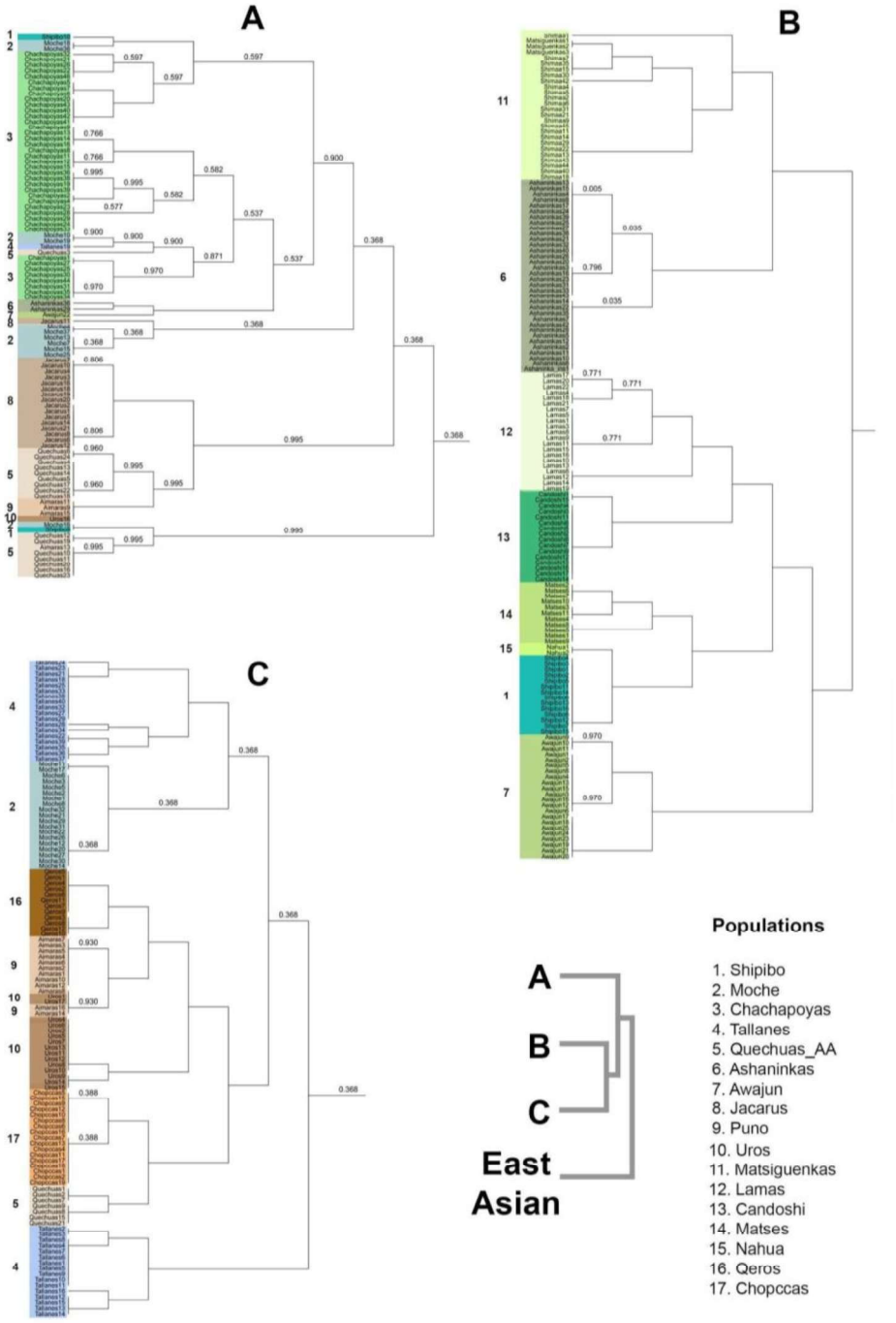


Figure S8. fineSTRUCTURE clustering analysis for the Dataset 1.9M dataset. The tree shows the haplotype sharing between Native Americans and East Asian samples. Figures A, B, and C represent the clusters A, B, and C, respectively, in the tree on the right. East Asian clusters grouped all Asian samples. Shades of blue are related to Peruvian Coast populations. Orange-brown colors are related to Andean populations and green colors are related to Amazon. Shades of purple are related to Mesoamericans. Shades of beige are related to North American natives.

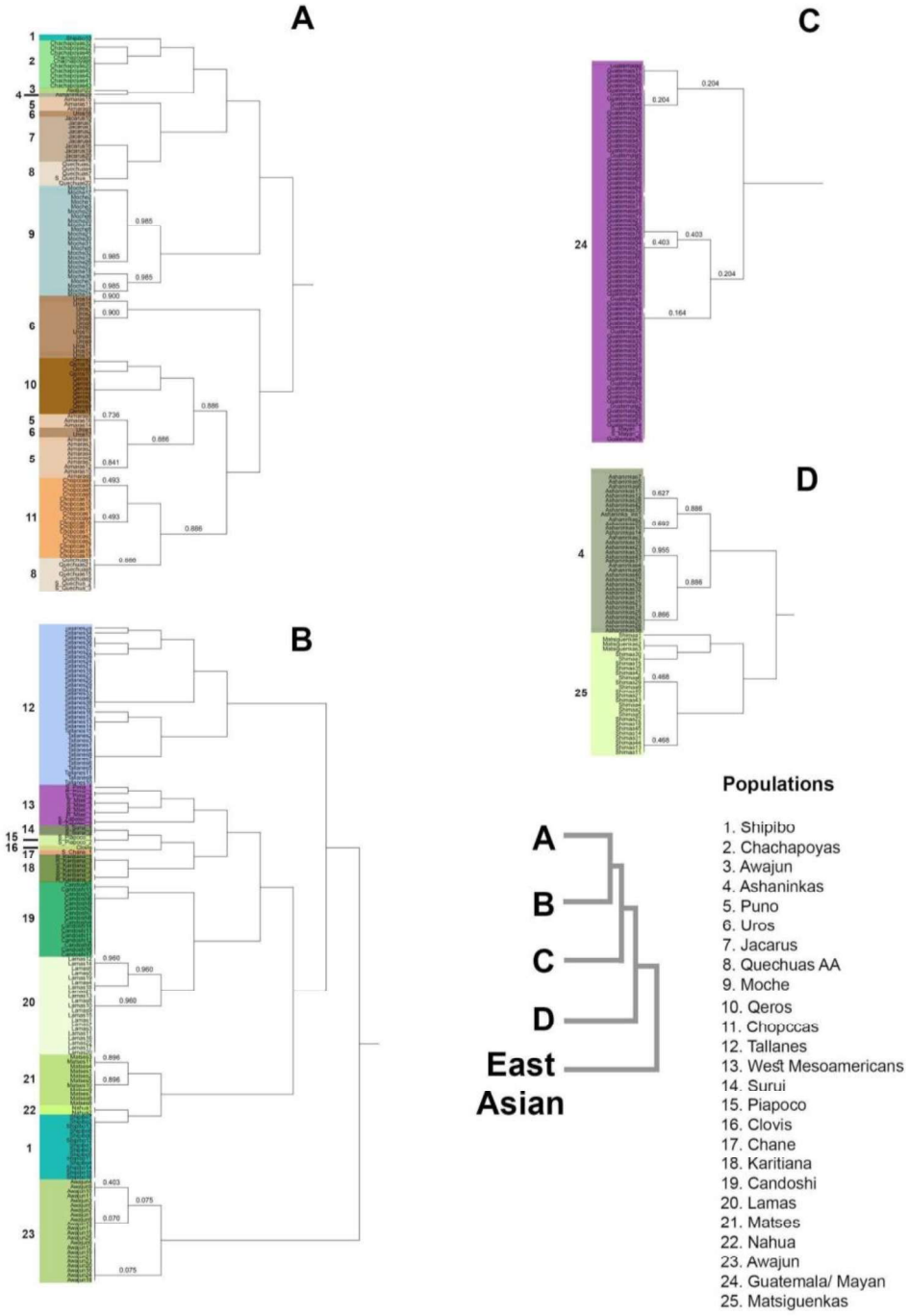


Figure S9. fineSTRUCTURE clustering analysis for the Dataset 500K dataset. The tree shows the haplotype sharing between Native Americans and East Asian samples. Figures A, B, C and D represent the clusters A, B, C and D, respectively, in the tree on the right. East Asian clusters grouped all Asian samples. Shades of blue are related to Peruvian Coast populations. Orange-brown colors are related to Andean populations and green colors are related to Amazon. Shades of purple are related to Mesoamericans. Shades of beige are related to North American natives.

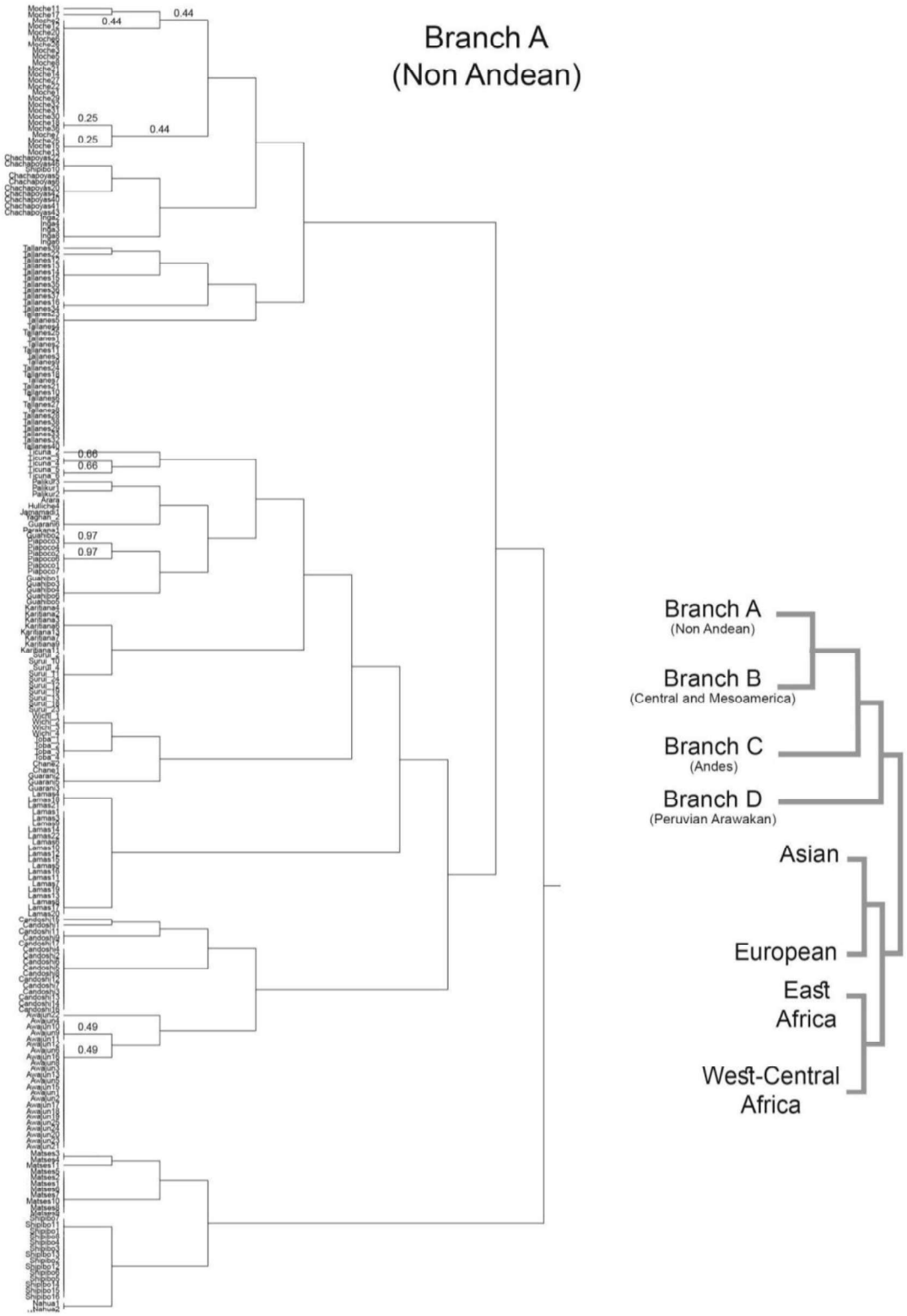


Figure S10. fineSTRUCTURE clustering analysis for the Dataset 230K dataset. A) Branch A of the tree showing the clustering of the Non Andean populations of South America. B) Branches B (Central Americans and Mesoamericas), C (Andean populations) and D (Peruvian Arawakan Ashaninkas and Matsiguenkas).

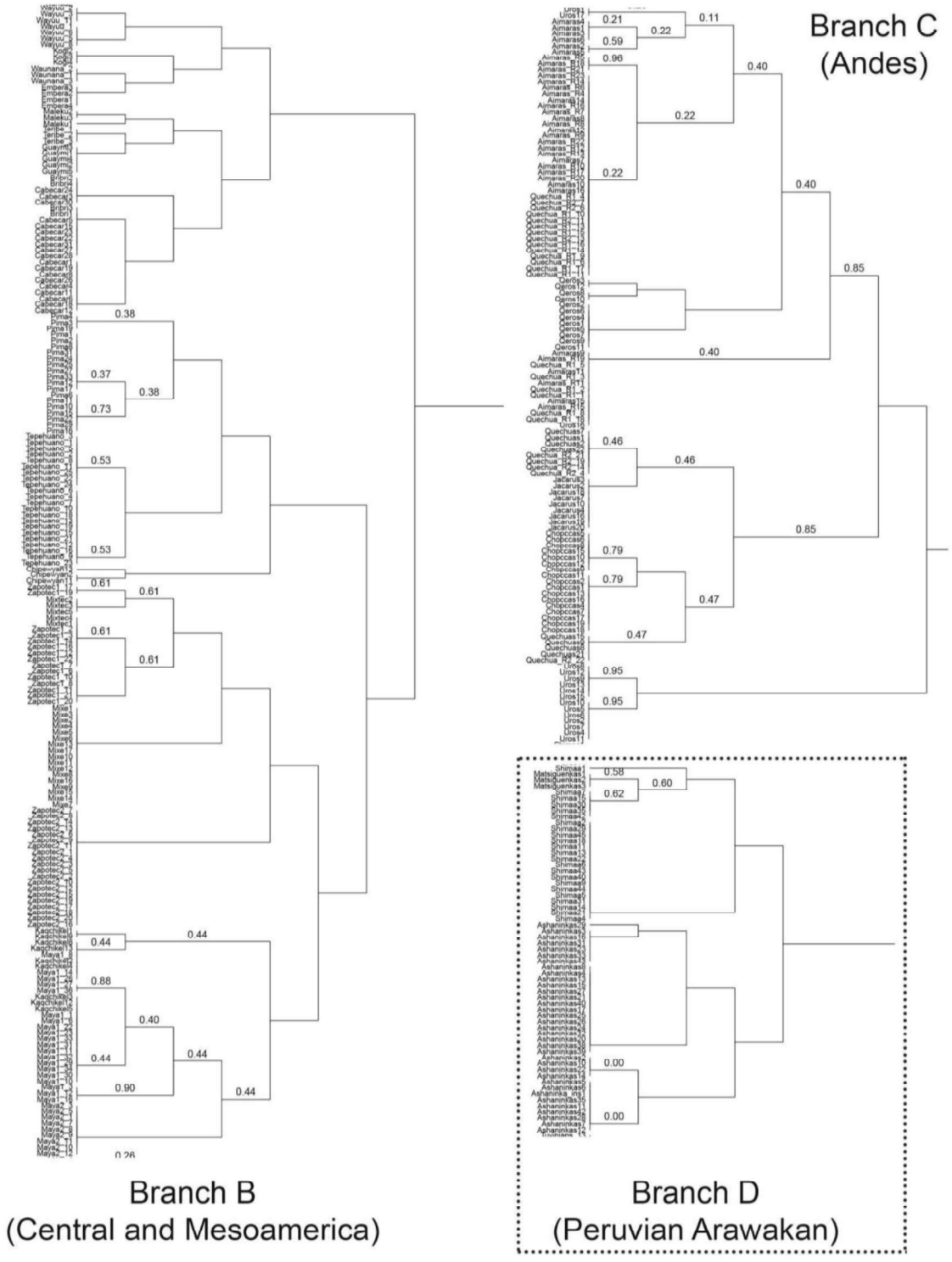


Figure S10 (Continued). fineSTRUCTURE clustering analysis for the Dataset 230K dataset. A) Branch A of the tree showing the clustering of the Non Andean populations of South America. B) Branches B (Central Americans and Mesoamericas), C (Andean populations) and D (Peruvian Arawakan Ashaninkas and Matsiguenkas).

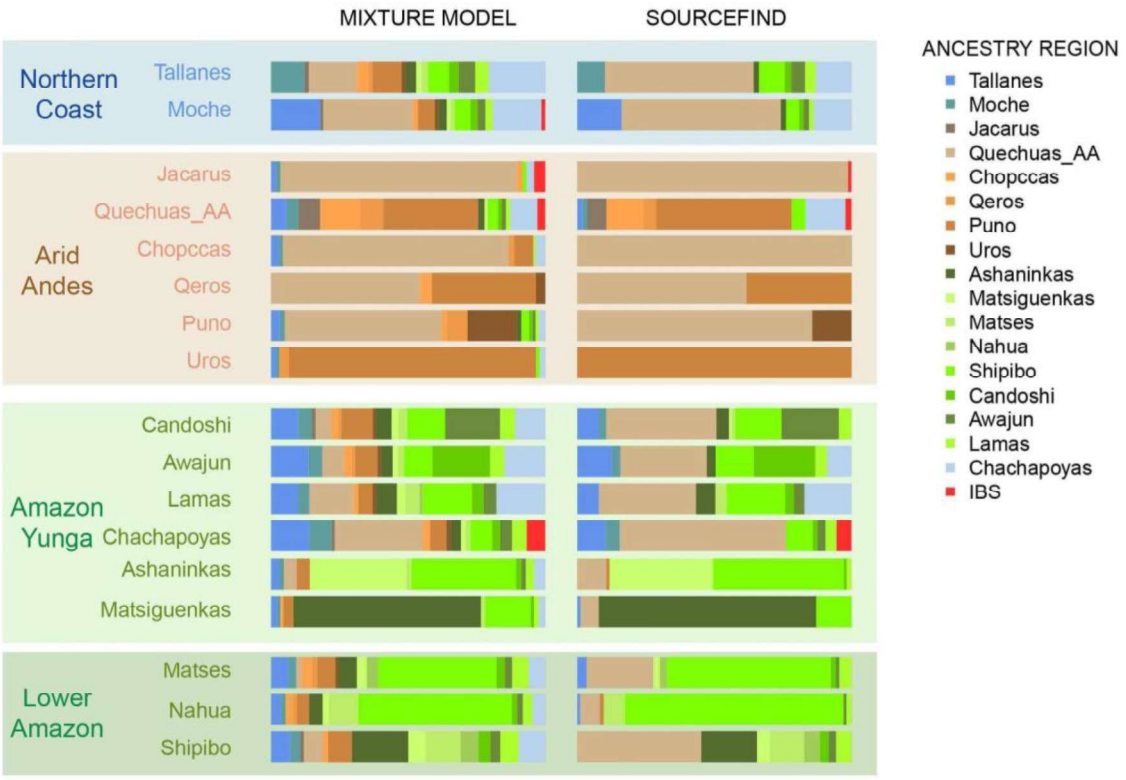


Figure S11. Proportions of haplotype sharing for each target population respect to Native Americans, Europeans and Africans donors populations, for the Dataset Natives 1.9M, inferred by two approaches: A non negative regression (MIXTURE MODEL) and a Bayesian approach (SOURCEFIND). Colored bars indicate a proportion of shared haplotypes shared DNA between the target population and a specific donor.

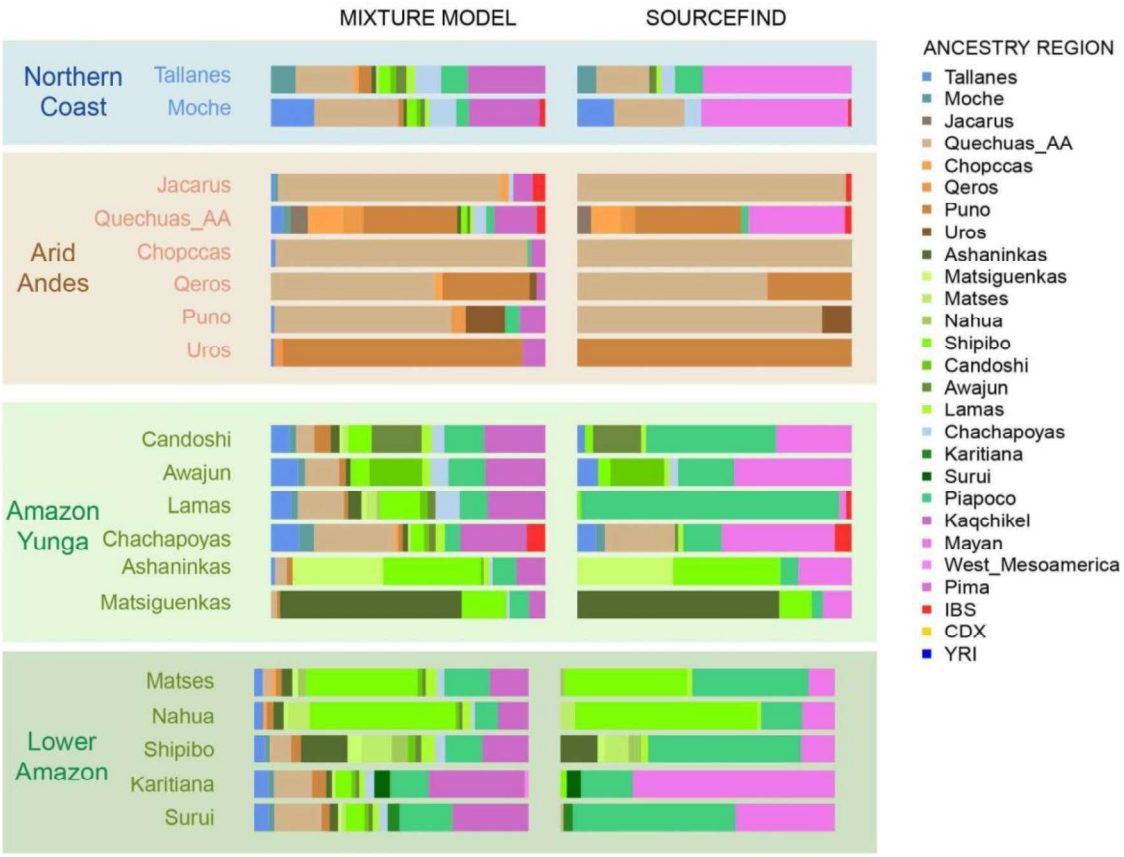


Figure S12. Proportions of haplotype sharing for each target population respect to Native Americans, Europeans and Africans donors populations, for the Dataset Natives 500K, inferred by two approaches: A non negative regression (MIXTURE MODEL) and a Bayesian approach (SOURCEFIND). Colored bars indicate a proportion of shared haplotypes shared DNA between the target population and a specific donor.

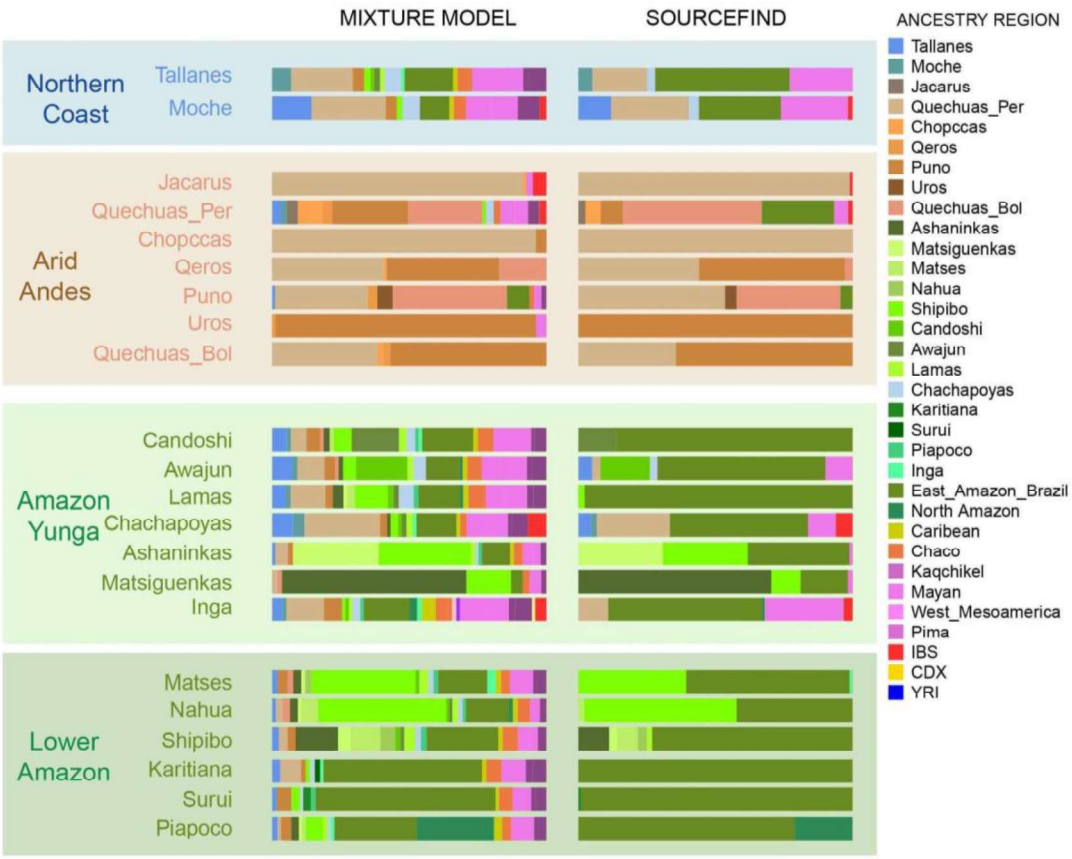


Figure S13. Proportions of haplotype sharing for each target population respect to Native Americans, Europeans and Africans donors populations, for the Dataset Natives 230K, inferred by two approaches: A non negative regression (MIXTURE MODEL) and a Bayesian approach (SOURCEFIND). Colored bars indicate a proportion of shared haplotypes shared DNA between the target population and a specific donor.

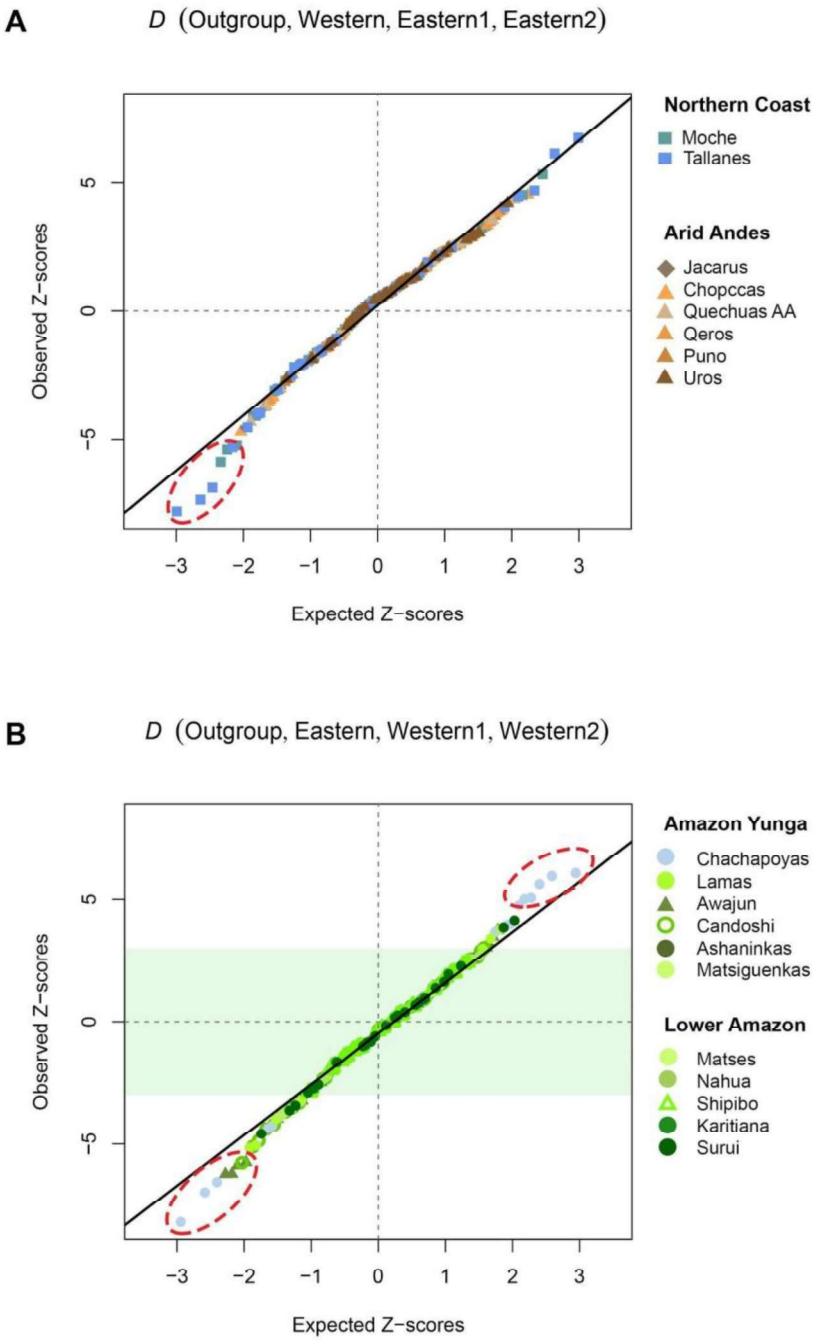


Figure S14. Quantile-quantile plot comparing Z-scores from *D*-statistics relating Western (Northern Coast and Arid Andes) and Eastern Andean slope (Amazon Yunga and Amazon Yunga) populations to those expected under a normal distribution (green diagonal) for the Dataset 500K. Red dashed circles show the Eastern populations with significant values of *D* statistics . **A)** We tested the configuration (outgroups (Western (Eastern1, Eastern2))). We detected evidence of gene flow between the Northern Coast and Amazon Yunga populations. **B)** We test the configuration (outgroups (Eastern (Western1, Western2))). We detected strong genetic affinity between Awajun, Candoshi, Lamas and Chachapoyas with Western populations.

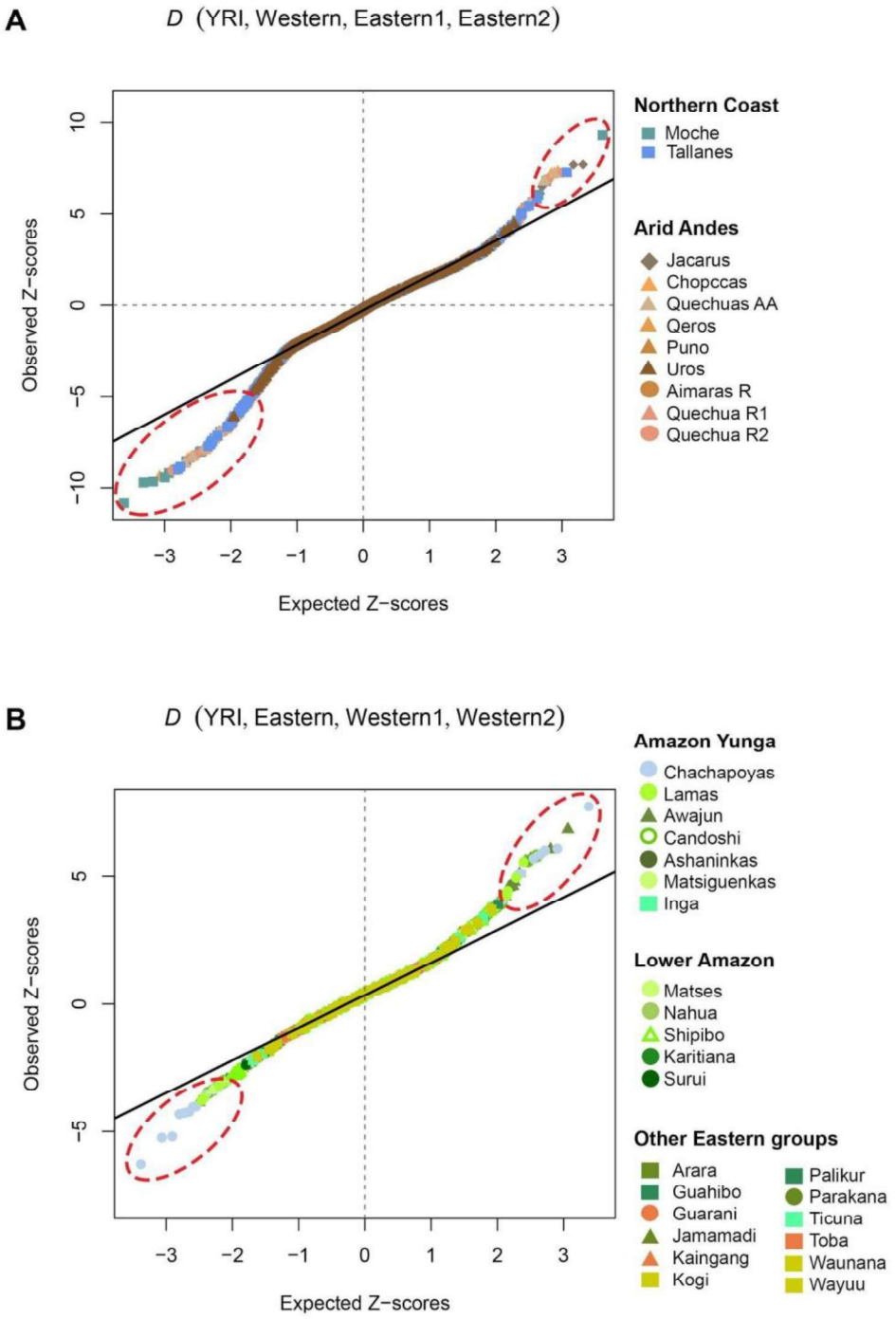
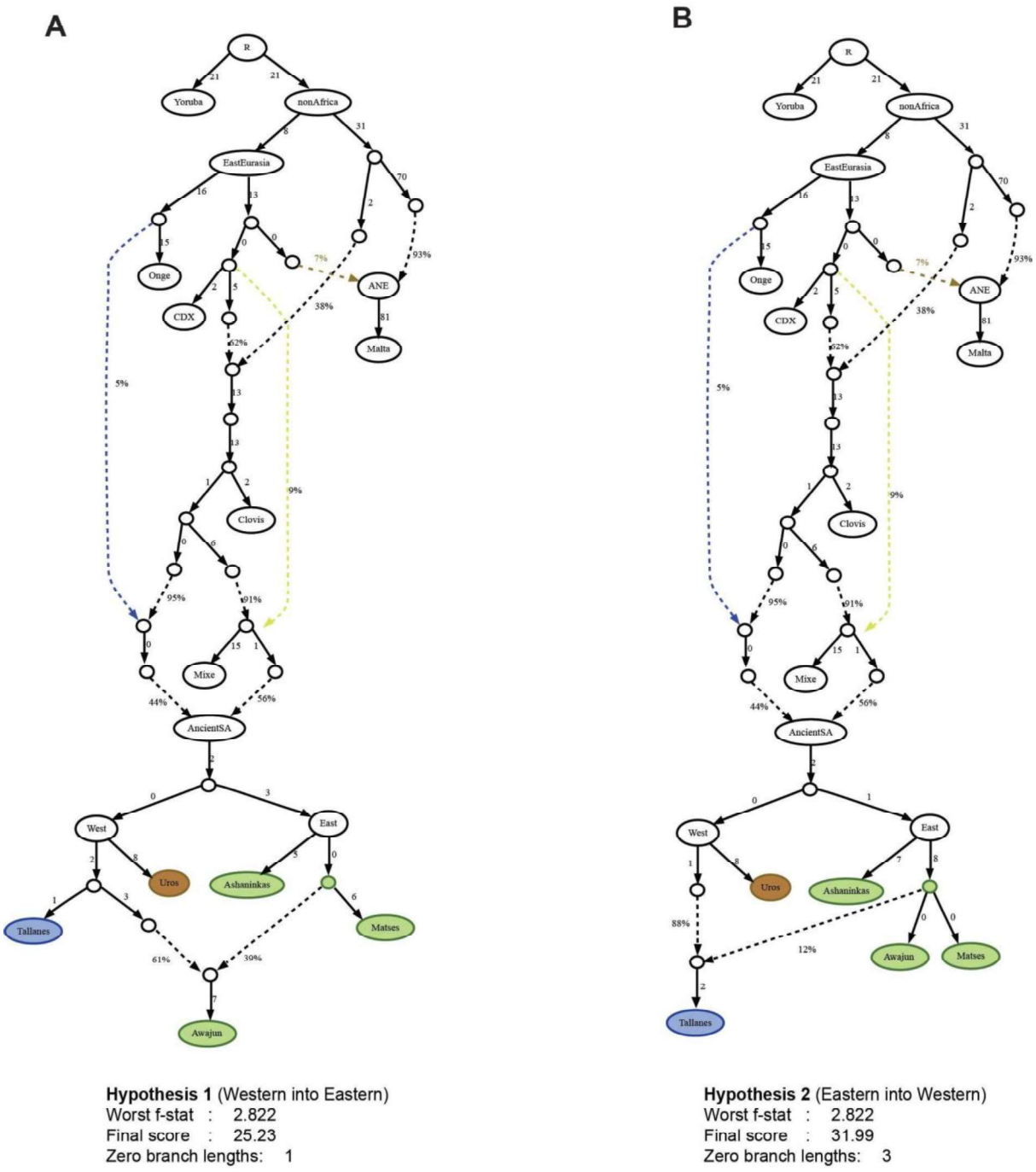


Figure S15. Quantile-quantile plot comparing Z-scores from D-statistics relating Western (Northern Coast and Arid Andes) and Eastern Andean slope populations (Amazon Yunga, Lower Amazon and other eastern groups) to those expected under a normal (green diagonal) distribution for the Dataset 230K. Red dashed circles show the Eastern populations with significant values of *D* statistics. **A)** We test the configuration (outgroups (Western (Eastern1, Eastern2))). We detected evidence of gene flow between Peruvian Coast and Eastern populations in the North Fertile Andes. **B)** We test the configuration (outgroups (Eastern (Western1, Western2))). We detected strong genetic affinity between Awajun, Candoshi, Lamas and Chachapoyas with Western populations.



Hypothesis 1 (Western into Eastern)
Worst f-stat : 2.822
Final score : 25.23
Zero branch lengths: 1

Hypothesis 2 (Eastern into Western)
Worst f-stat : 2.822
Final score : 31.99
Zero branch lengths: 3

Figure S16. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes. We explore the relationship of the Tallanes-Awajun and the direction of the gene flow. White balls in the intermediate nodes represent hypothetical ancestors for each divergence event. **A)** Admixture graph for testing the Hypothesis 1 (from Western to Eastern): the gene flow from the Northern Coast to Awajun. **B)** Admixture graph for testing the Hypothesis 2 (From Eastern to Western) the gene flow event into Tallanes. Hypothesis 1 is better supported considering its lower final score and number of zeroed branches in contrast to Hypothesis 2.

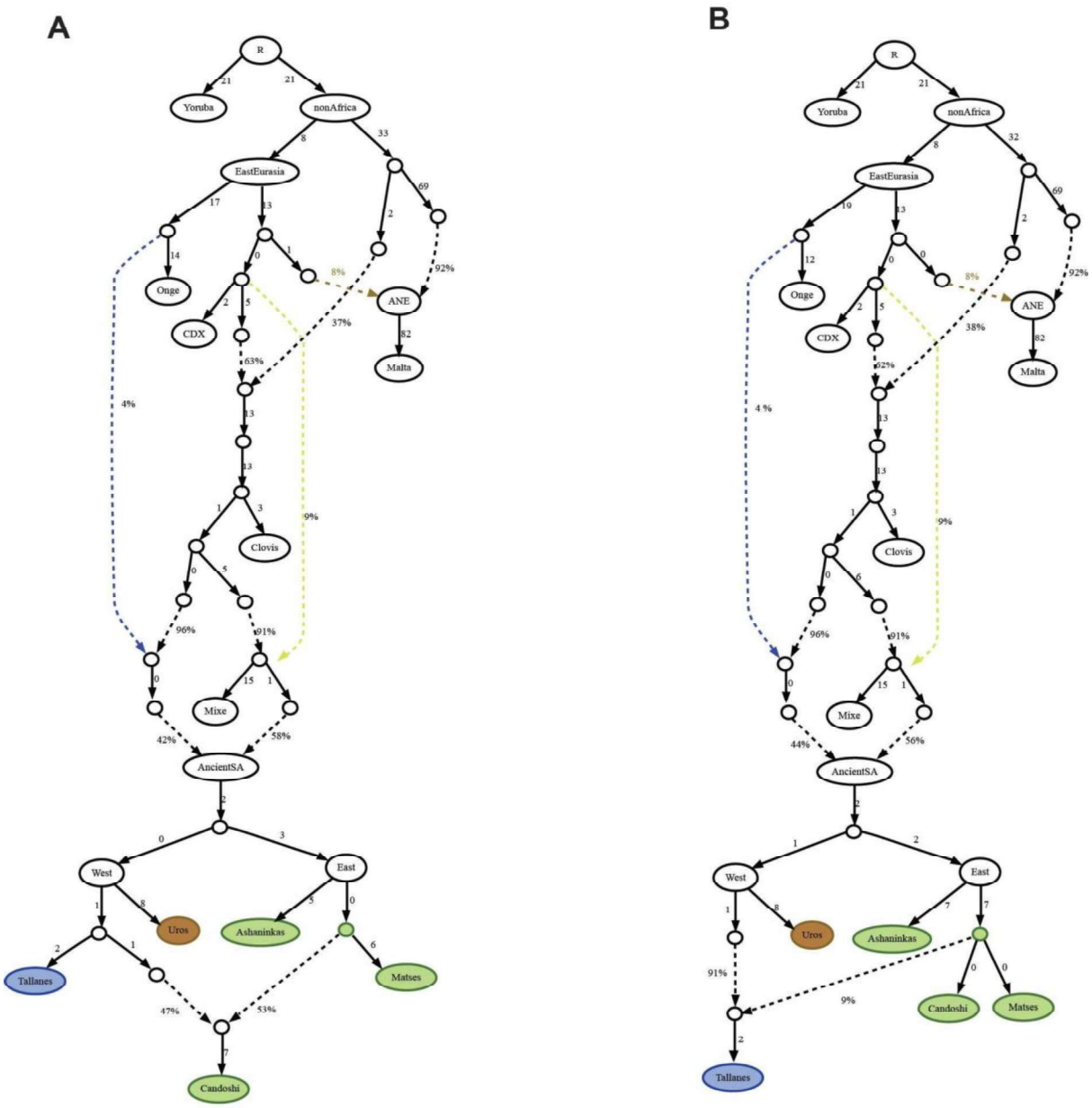
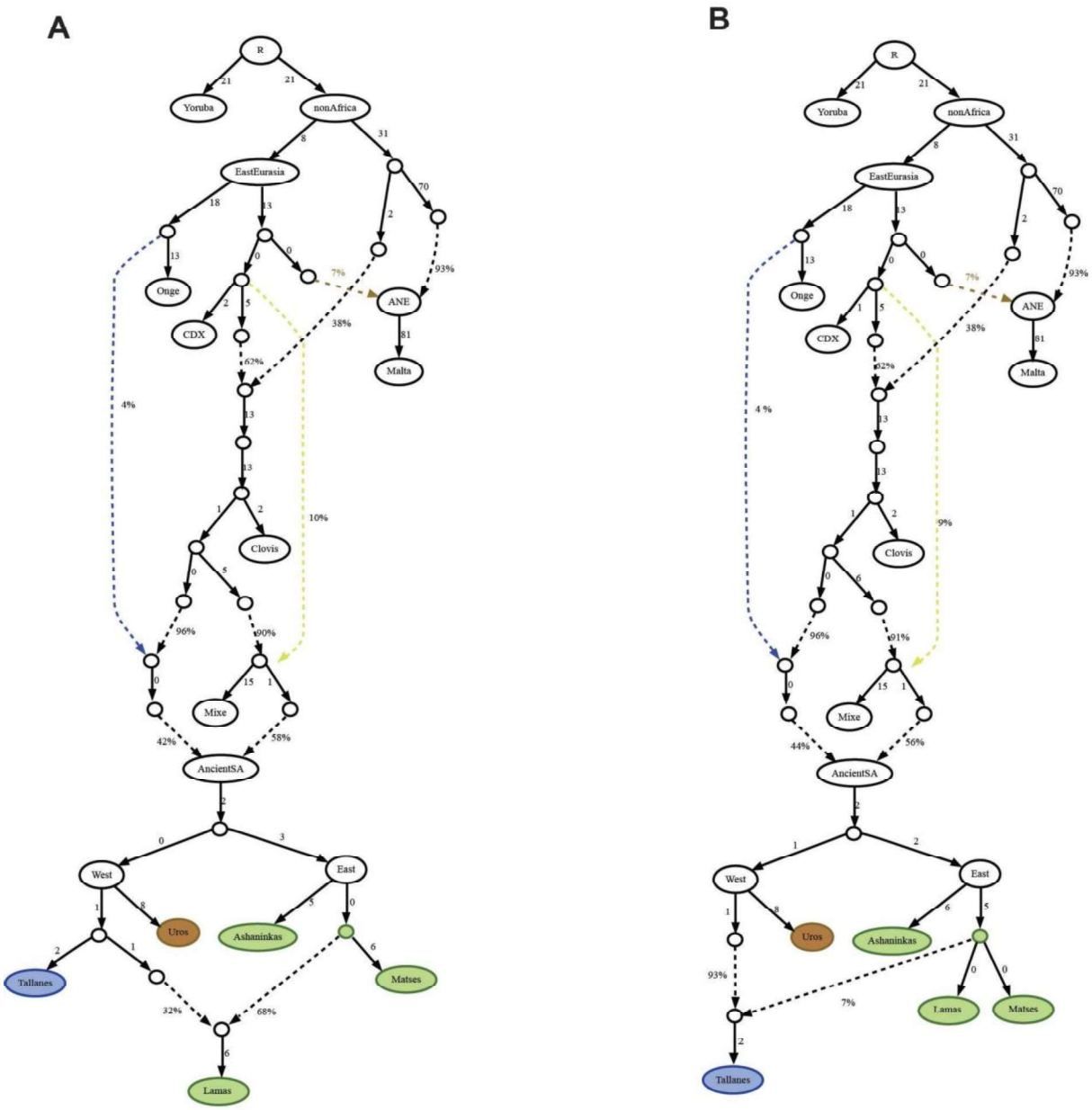


Figure S17. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes. We explore the relationship of the Tallanes-Candoshi and the direction of the gene flow. White balls in the intermediate nodes represent hypothetical ancestors for each divergence event. **A)** Admixture graph for testing the Hypothesis 1 (from Western to Eastern): the gene flow from the Northern Coast into Candoshi. **B)** Admixture graph for testing the Hypothesis 2 (From Eastern to Western) the gene flow event from Candoshi into Tallanes. Hypothesis 1 is better supported considering its number of zeroed branches in contrast to Hypothesis 2.



Hypothesis 1 (Western into Eastern)
Worst f-stat : 2.822
Final score : 28.684
Zero branch lengths: 0

Hypothesis 2 (Eastern into Western)
Worst f-stat : 2.823
Final score : 25.863
Zero branch lengths: 2

Figure S18. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes. We explore the relationship of the Tallanes-Lamas and the direction of the gene flow. White balls in the intermediate nodes represent hypothetical ancestors for each divergence event. **A)** Admixture graph for testing the Hypothesis 1 (from Western to Eastern): the gene flow from the Northern Coast into Lamas. **B)** Admixture graph for testing the Hypothesis 2 (From Eastern to Western) the gene flow event from Lamas into Tallanes. Hypothesis 1 is better supported considering that it does not include any zeroed branches in contrast to Hypothesis 2.

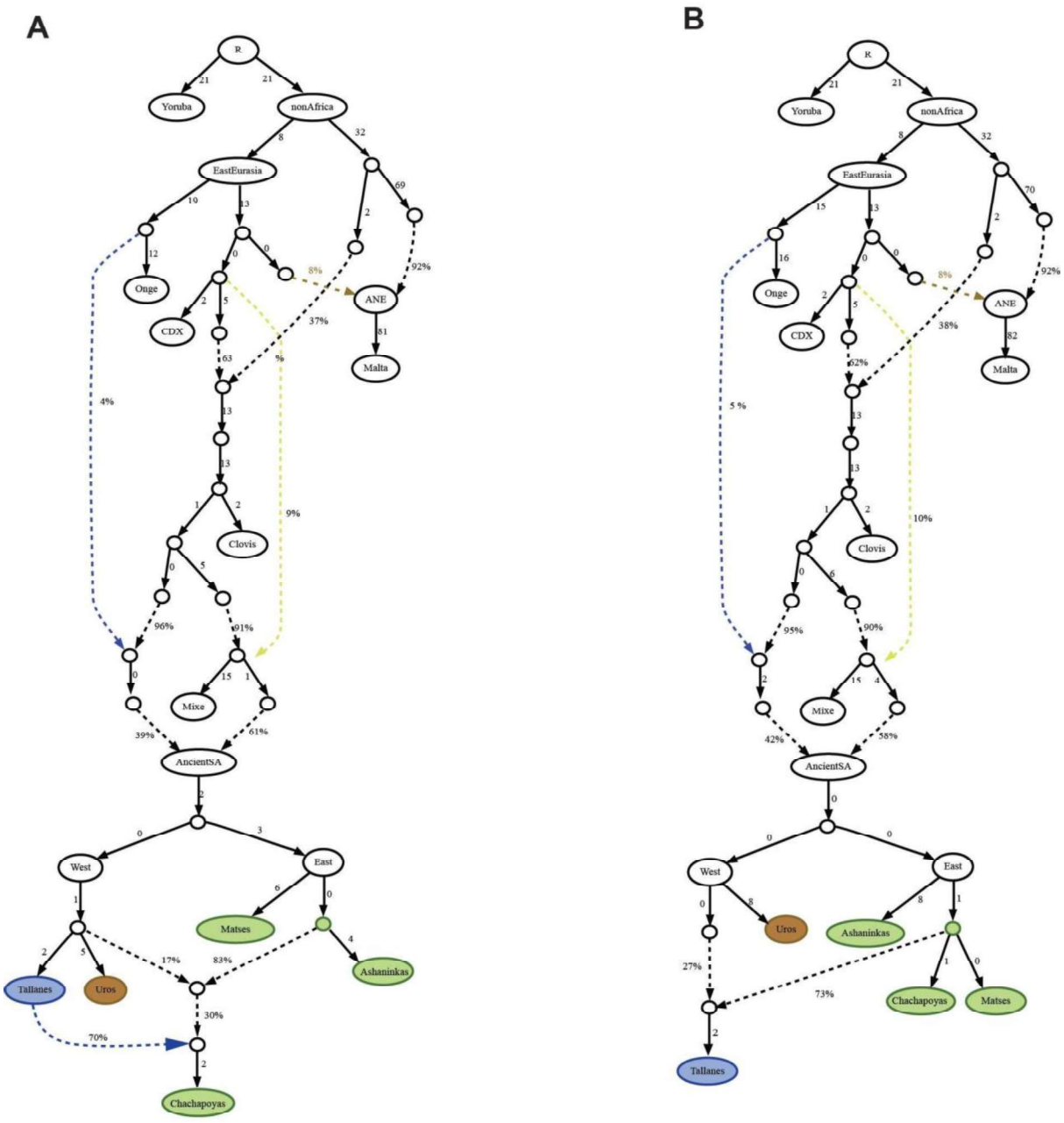


Figure S19. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes. We explore the relationship of the Tallanes-Chachapoyas and the direction of the gene flow. White balls in the intermediate nodes represent hypothetical ancestors for each divergence event. **A)** Admixture graph for testing the Hypothesis 1 (from Western to Eastern): the gene flow from the Northern Coast into Chachapoyas. **B)** Admixture graph for testing the Hypothesis 2 (From Eastern to Western) the gene flow event from Chachapoyas into Tallanes. Hypothesis 1 is better supported considering its significant f statistic, lower final score and just one zeroed branch lengths.

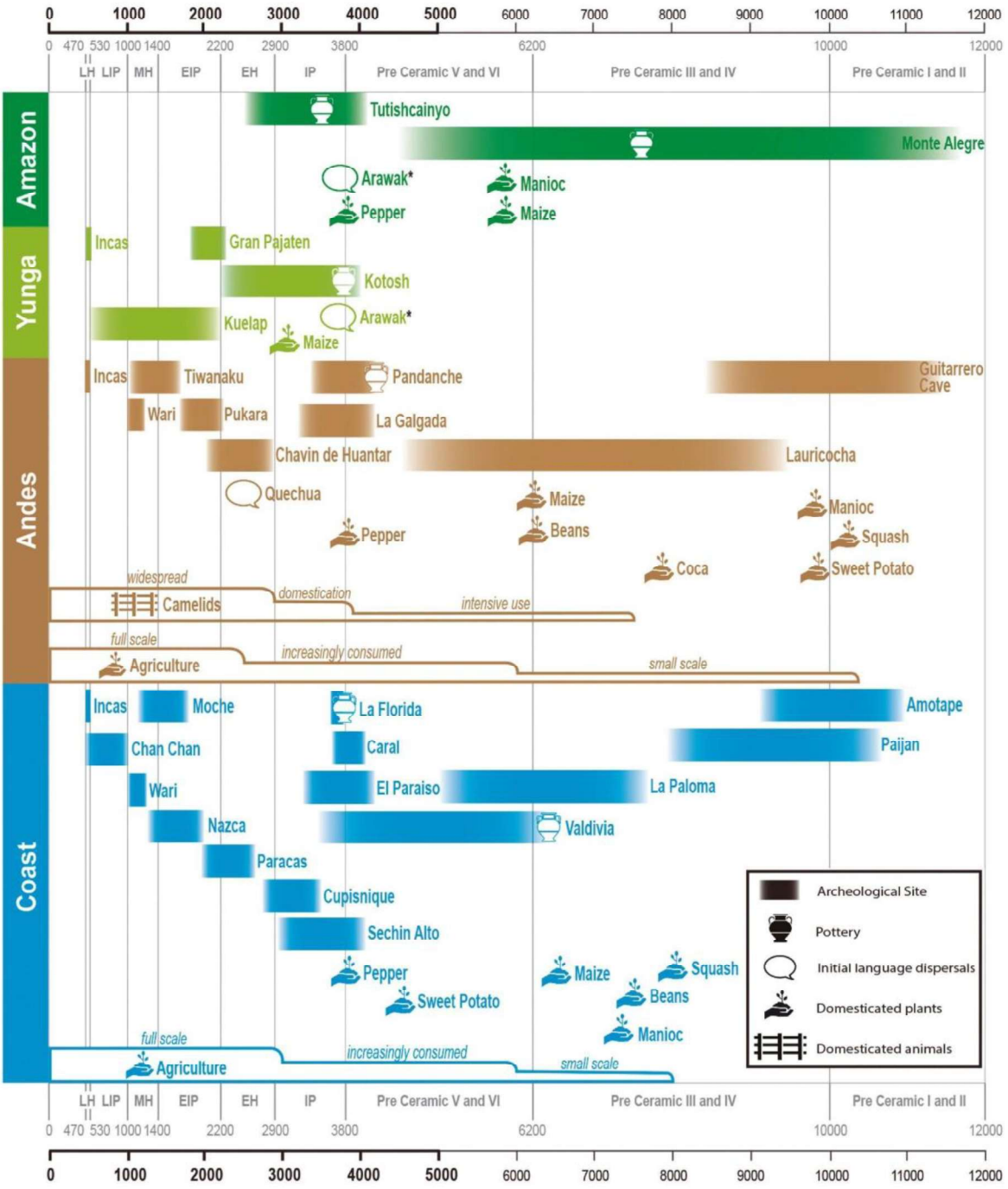


Figure S20. Key historical events of Peruvian prehistory in four longitudinal regions: Peruvian Coast, Andes, Amazon Yunga and Amazonia. Pottery and cultivars symbols represent the earliest archaeological record for the region. To account for time uncertainties, This figure showed the events in the chronology plot without clearly defined chronological borders. Timeline on the top and bottom is represented in Years before present. LH: Late Horizon, LIP: Late Intermediate Period, MH: Middle Horizon, EIP: Early Intermediate Period, EH: Early Horizon, IP: Initial Period. *Controversial geographic region of Arawak origin. Each step in Agriculture and Camelids representations shows an increase in their relative importance. Adapted from ref. 92, which is licensed under CC BY 4.0.

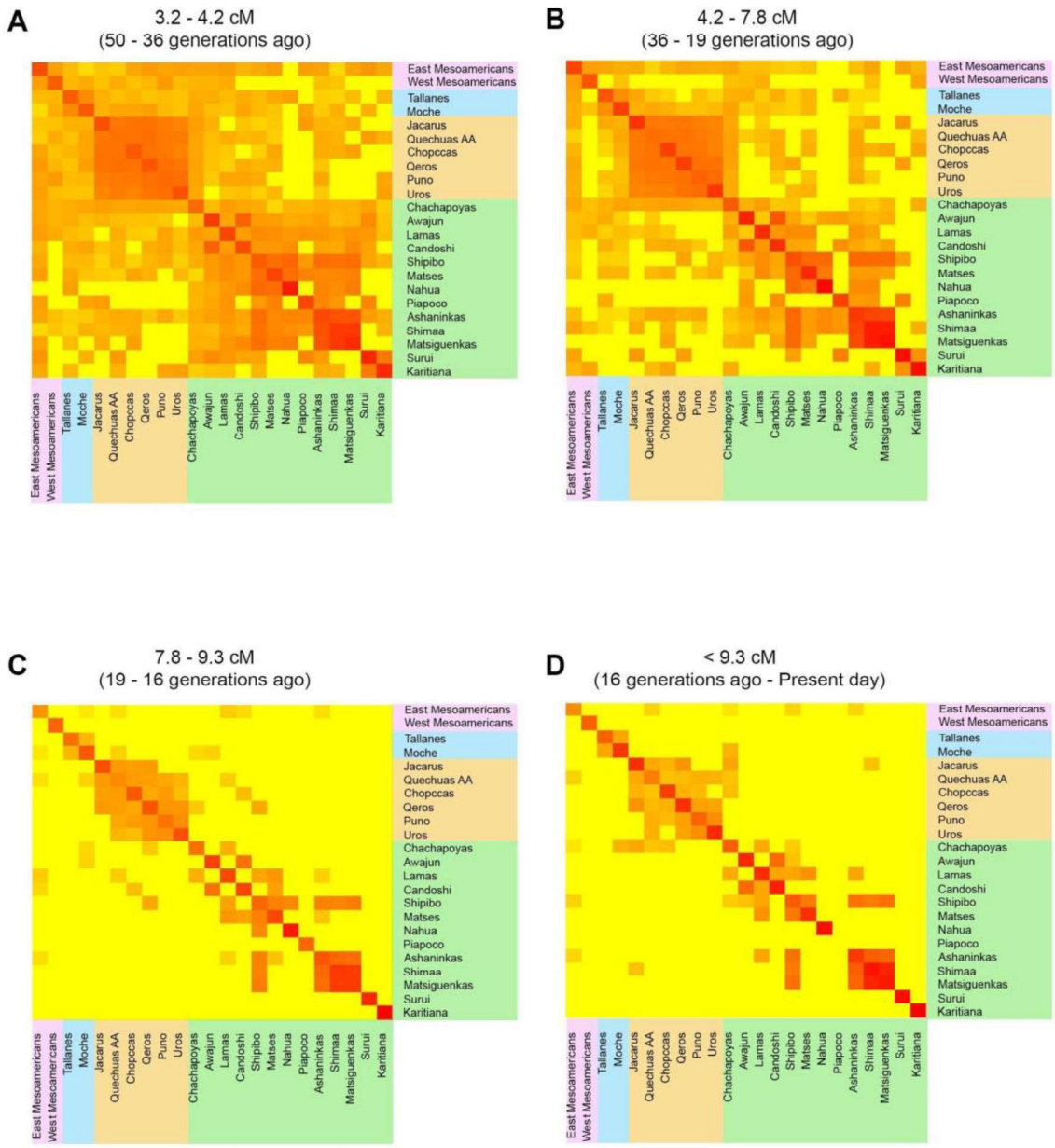


Figure S21. Heatmap representation of the shared Identical by descent (IBD) segments among Native Americans of the Natives 1.9M dataset. Each heatmap represents an interval of segments size and is correlated with time generation for the most common recent ancestor. A) An interval from 3.2 to 4.2 cM correlated with 50 to 36 generations ago. B) The second interval from 4.2 to 7.8 cM correlated with 36 to 19 generations ago. C) The third interval from 7.8 to 9.3 cM correlated with 19 to 16 generations ago. D) And the last interval for all segments longer than 9.3 cM correlated with 16 generations ago to the present day.

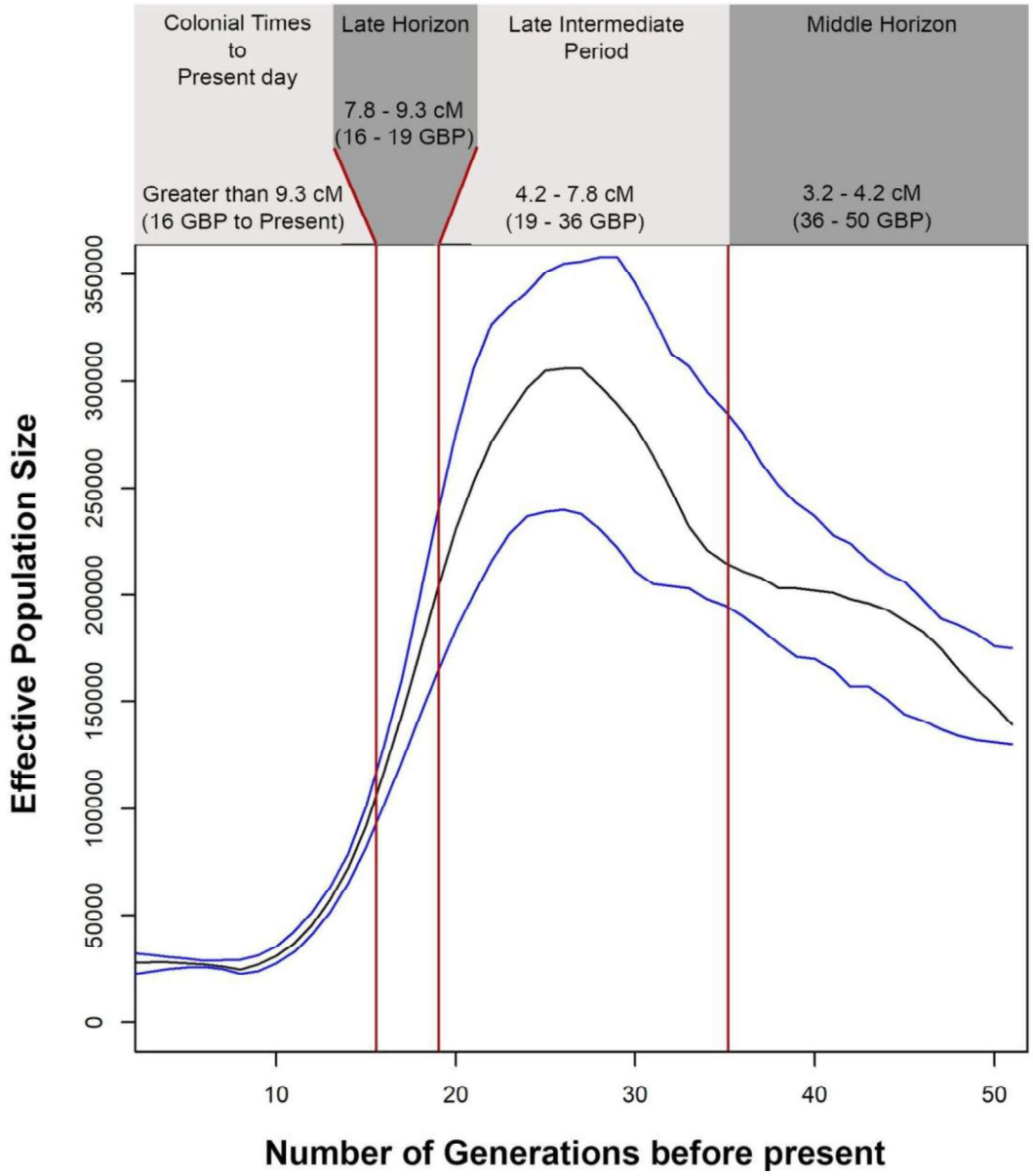


Figure S22. IBDNe analysis to infer the dynamic of the effective population size (Ne) from 4 generations ago to the last 50 generations for the Andean populations (Quechuas_AA, Aimaras_P, Chopccas, Qeros and Uros) as a whole. We used the Natives 1.9M dataset. The x axis represents the number of generations from the present to the past. The y axis represents the estimated value of the Ne. Blocks separated by red lines in the graph correspond to the intervals of the IBD heatmaps (Fig 2). GBP: Generations before present.

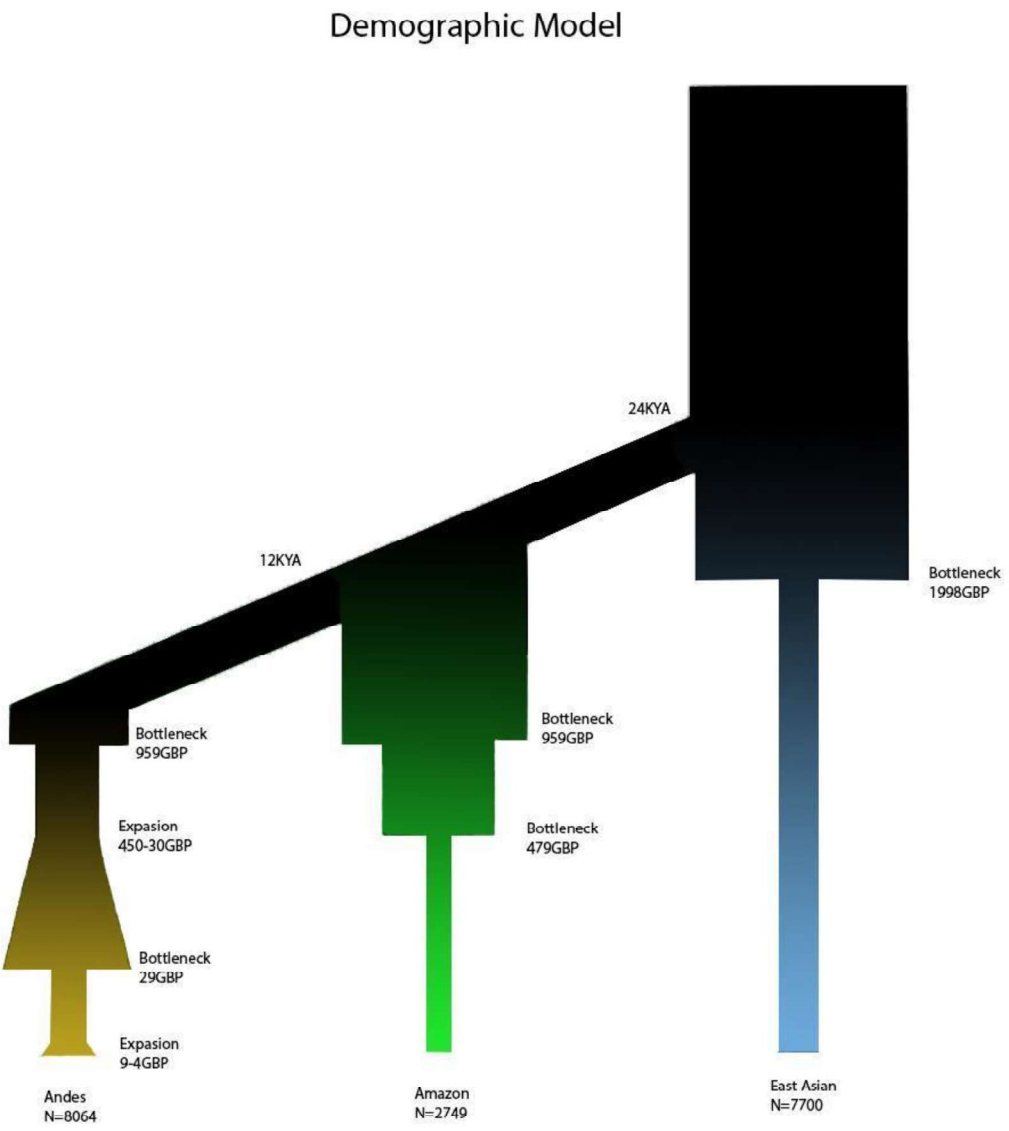


Figure S23. Demographic model of the Andean, Amazonian and East Asian populations. This model was used for the simulations made to calculate the p-value of the obtained PBSn values.

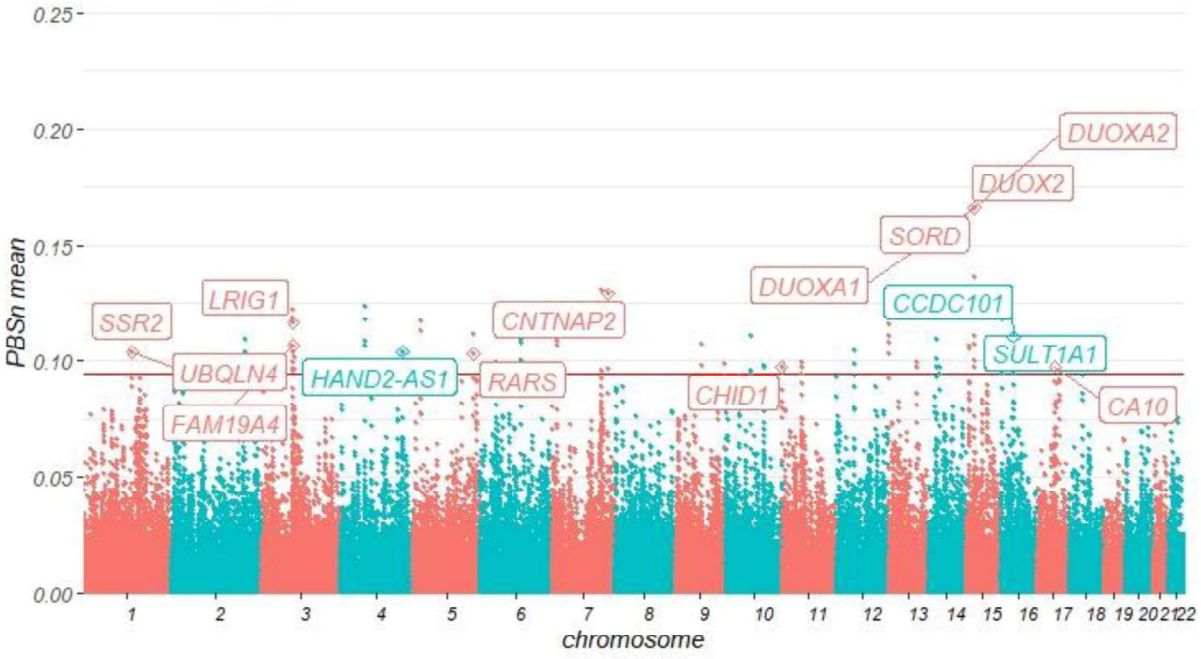


Figure S24. PBSn mean values Andean populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of PBSn mean (red line).

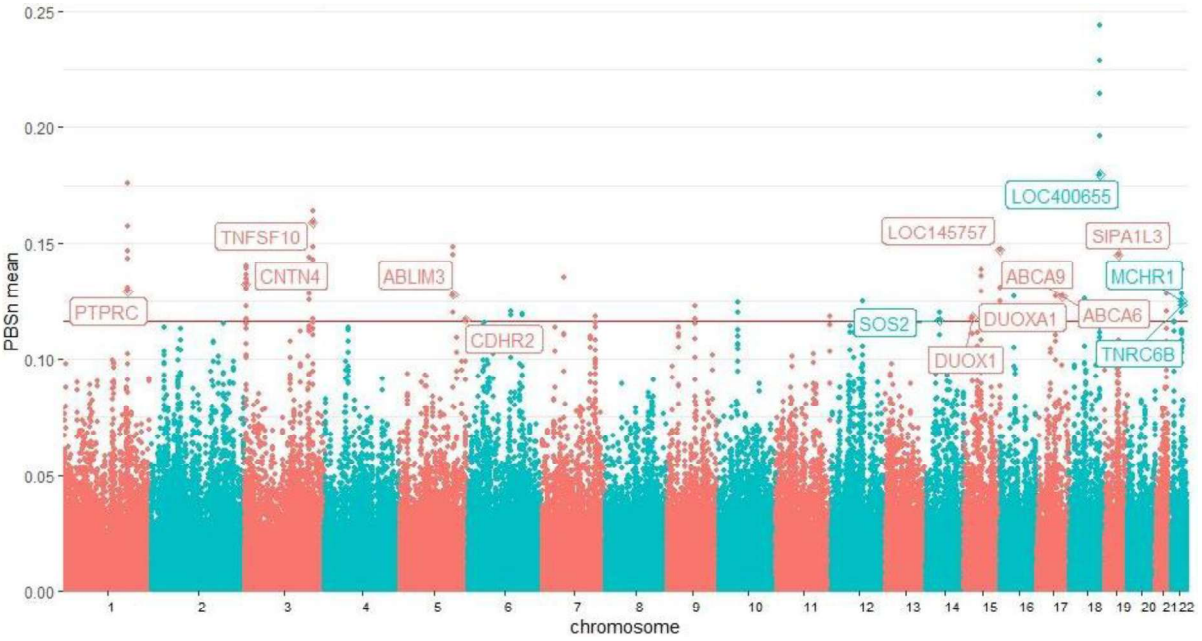


Figure S25. PBSn mean values Amazon populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of PBSn mean (red line).

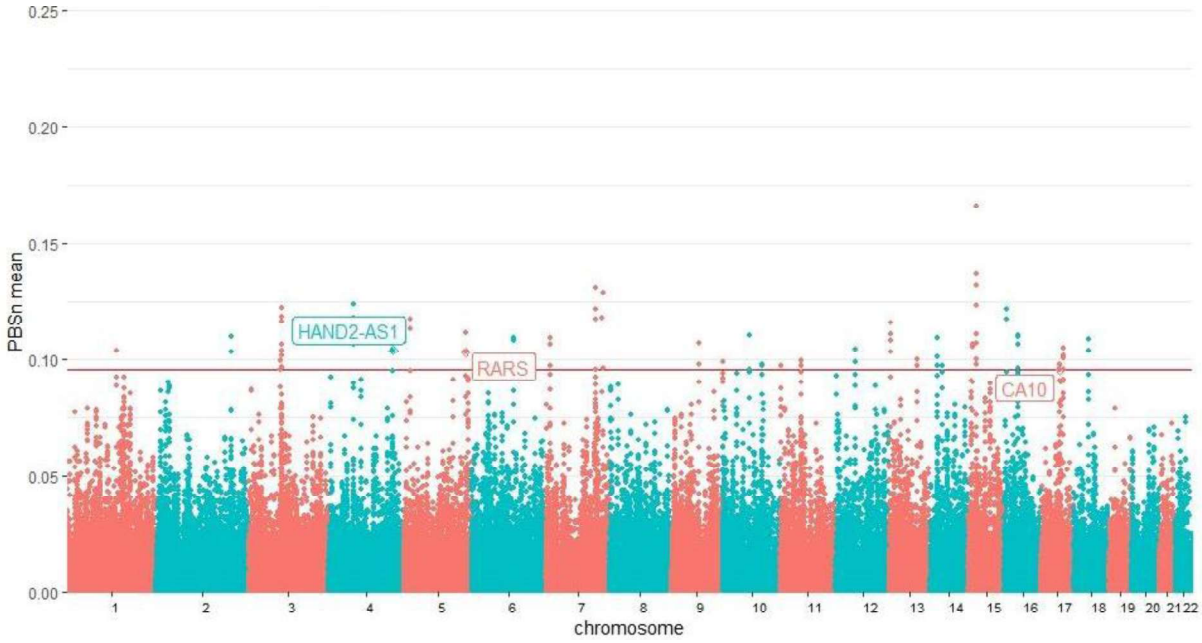


Figure S26. PBSn mean values for windows of 20 SNPs with five SNPs of overleap in Andean populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of windows PBSn mean (red line) that also present high values for xpEHH are labeled.

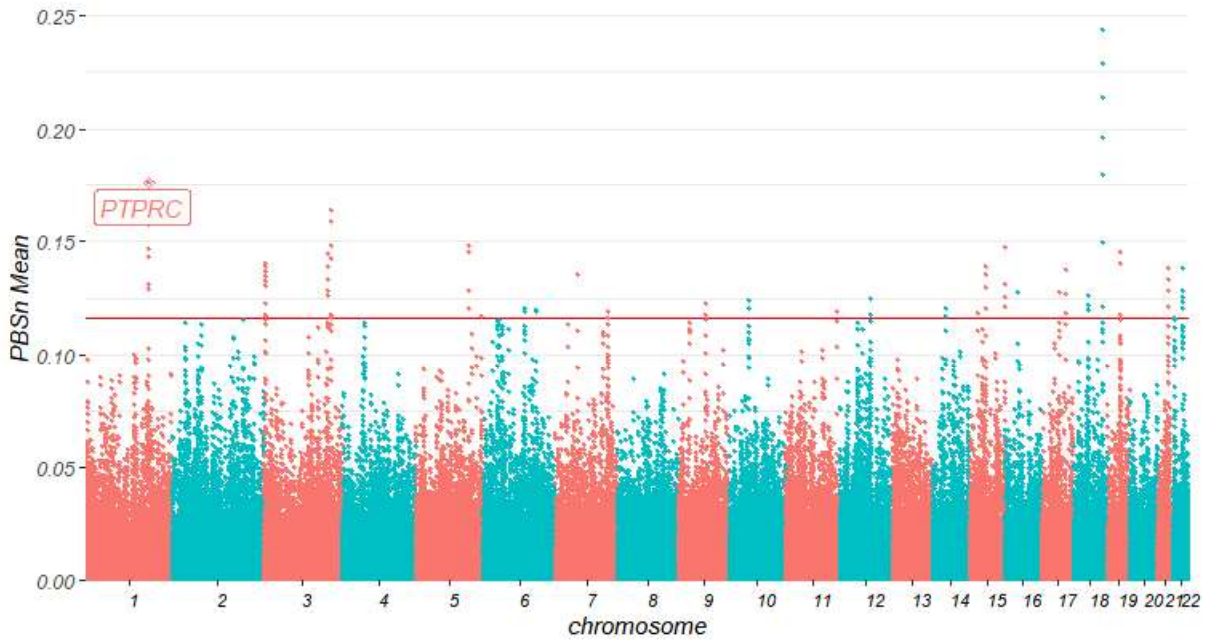


Figure S27. PBSn mean values for windows of 20 SNPs with 5 SNPs of overleap in Amazon populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of windows PBSn mean (red line) that also present high values for xpEHH are labeled.

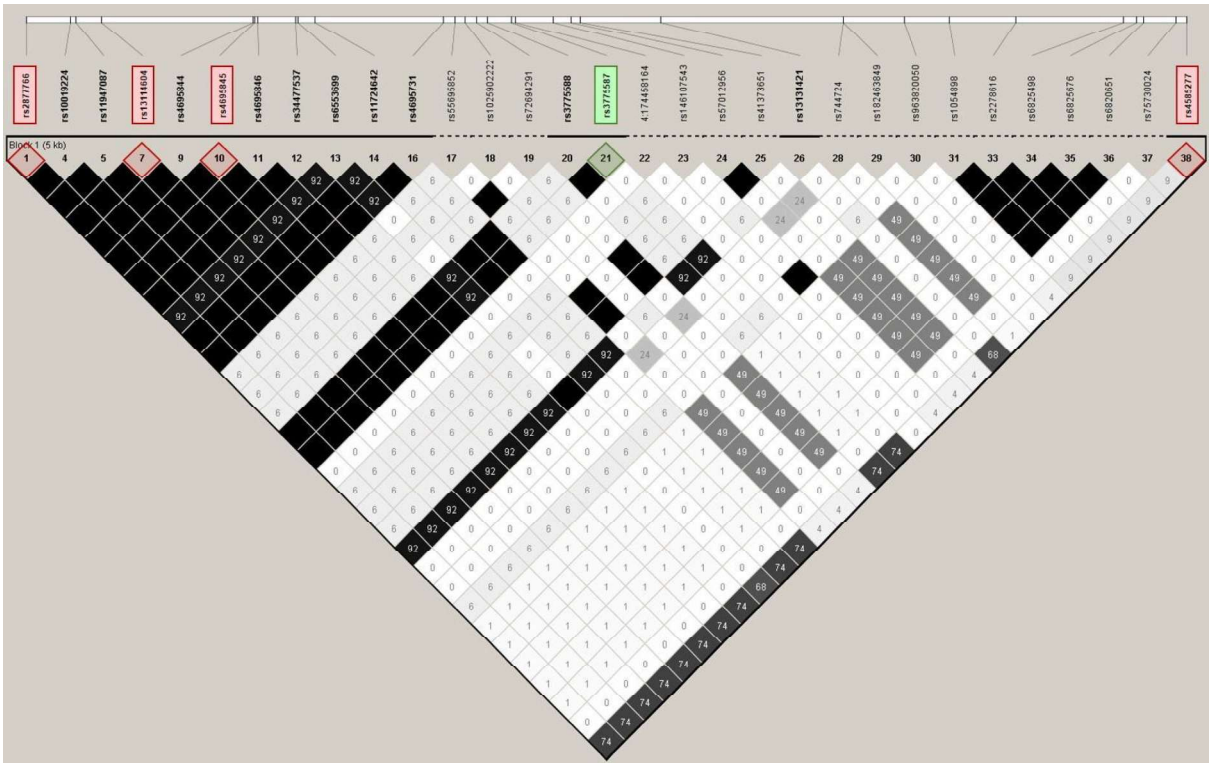


Figure S28. Linkage Disequilibrium between rs3775587 and the SNPs found to be under selection in the gene HAND2-AS1. SNPs with signals in PBS and xpEHH analysis are in red and SNP rs3775587, mapped within the putative enhancer GH04J173536 is in green.

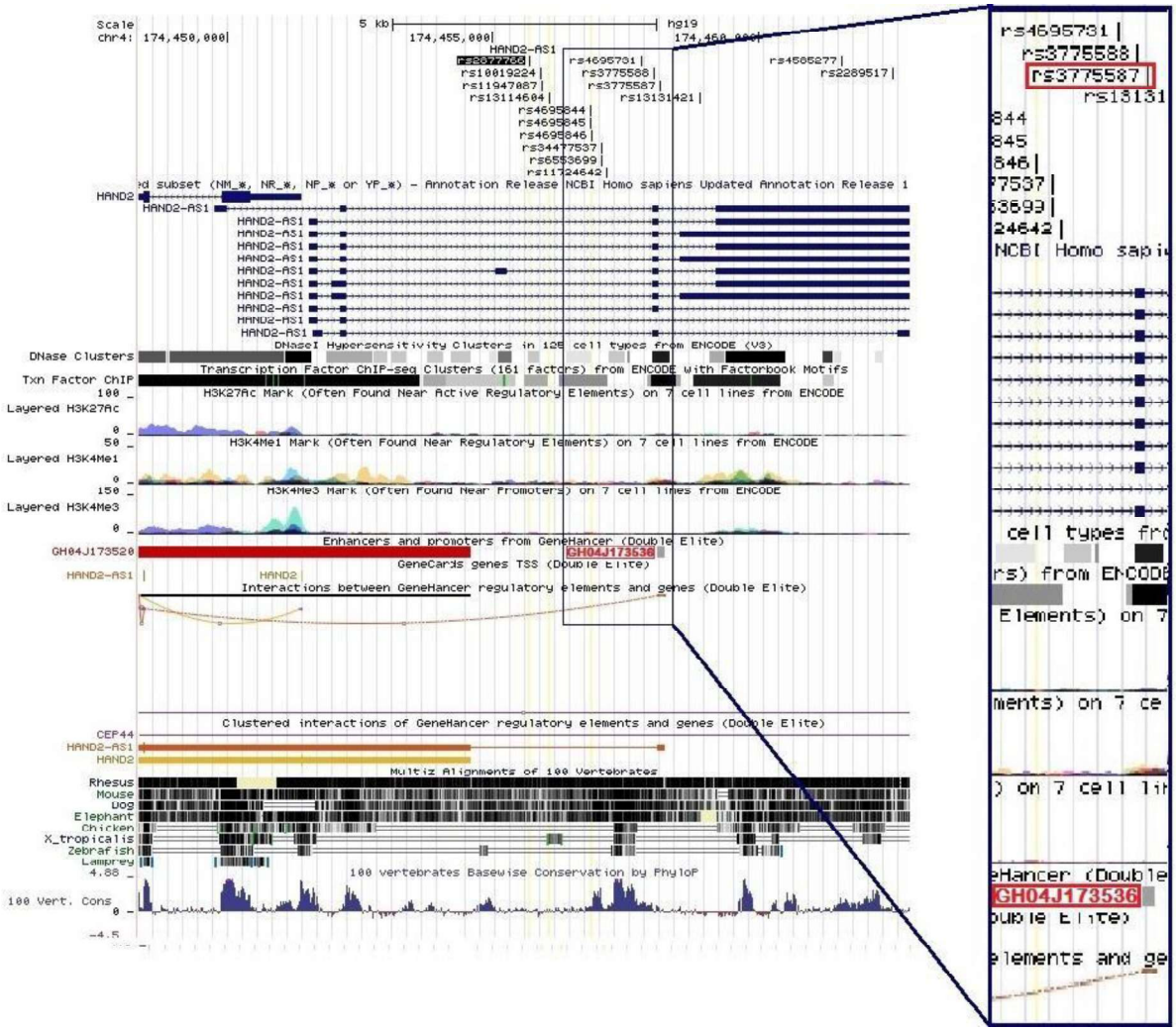


Figure S29. UCSC Genome Browser view of HAND2-AS1 locus with the SNPs located in regions found to be under selection, DNase I hypersensitivity clusters, Transcription Factor ChIP-seq binding sites, and the histone modifications H3K27ac (Often Found Near Active Regulatory Elements), H3K4me1 (Often Found Near Regulatory Elements) and H3K4me3 (Often Found Near Promoters) on cell lines from the ENCODE Project, and GeneHancer (see Supplementary Methods) and vertebrate conservation data. According to GeneHancer, rs2877766 and other SNPs lie within an ~2.5Kb intronic region of HAND2-AS1, which is located between a promoter/enhancer (GH04J173520) and an enhancer (GH04J173536).

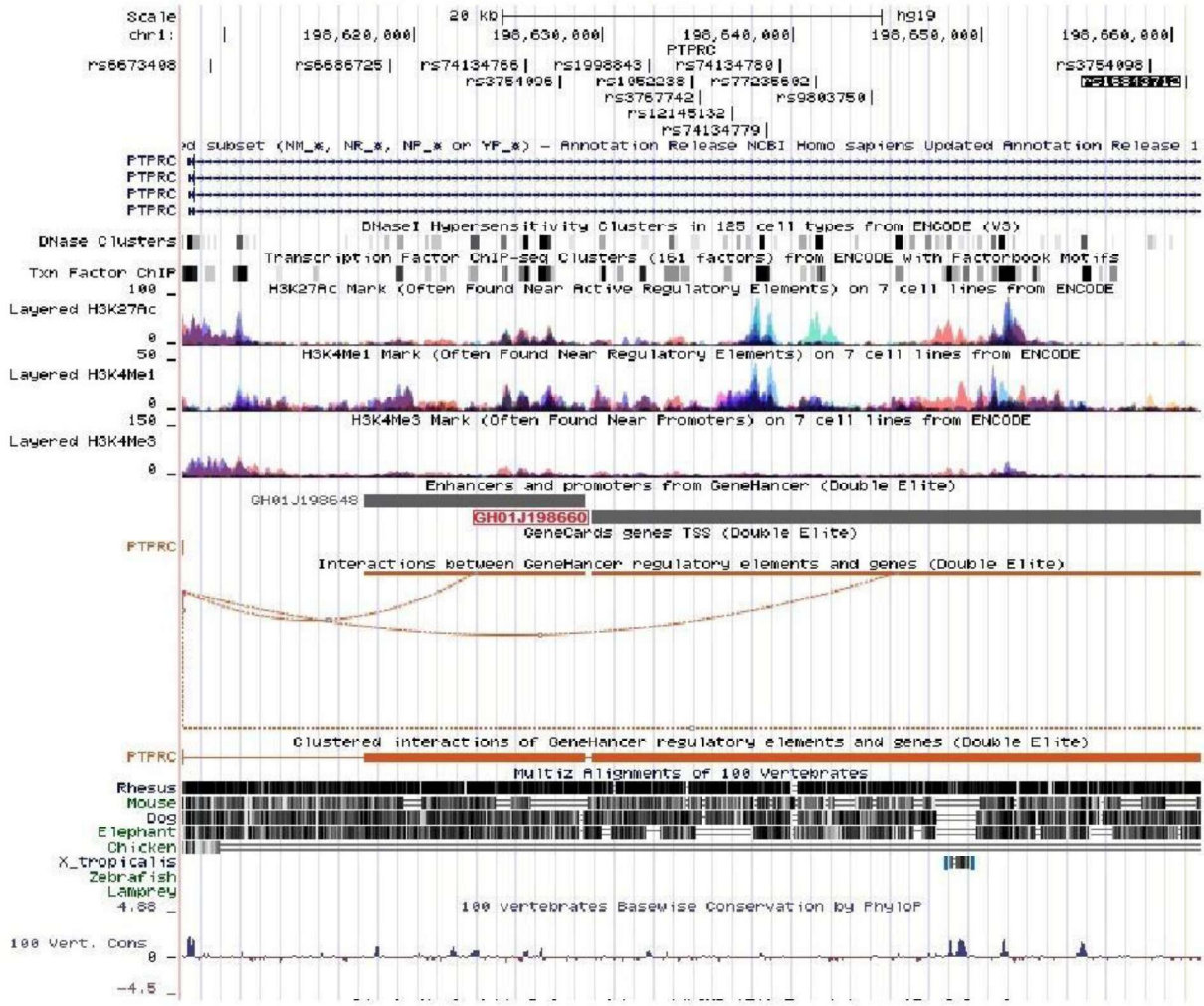


Figure S30. UCSC Genome Browser view of PTPRC locus with the SNPs located in regions found to be under selection, DNase I hypersensitivity clusters, Transcription Factor ChIP-seq binding sites, and the histone modifications H3K27ac (Often Found Near Active Regulatory Elements), H3K4me1 (Often Found Near Regulatory Elements) and H3K4me3 (Often Found Near Promoters) on cell lines from the ENCODE Project, and GeneHancer (see Supplementary Methods) and vertebrate conservation data. According to GeneHancer, rs16843712 and other SNPs lie within an intronic enhancer (GH01J198660) of PTPRC.

Legends for Datasets S1 to S3

Dataset S1: Description of 19 studied Native American populations from Peruvian National Institute of Health and from Laboratory of Human Genetic Diversity. Ashaninka population was sampled twice independently, for this reason, we merge these samples in a unique Ashaninka group and a total of 18 studied populations.
(Dataset_S1_Samples Dataset 1.9M.xlsx)

Dataset S2: List of all samples included in the Native 500K dataset.
(Dataset_S2_Samples Dataset500K.xlsx)

Dataset S3: List of all samples included in the Native 230K dataset.
(Dataset_S3_Samples Dataset 230K.xlsx)

Dataset S4: SNPs under selection in Andean populations according to Population Branch Statistic (PBS) test. (Dataset_S4_PBS_results_AND.xlsx)

Dataset S5: SNPs under selection in Amazon populations according to Population Branch Statistic (PBS) test. (Dataset_S5_PBS_results_AMZ.xlsx)

Dataset S6: SNPs under selection in Andean populations according to Population Branch Statistic (PBS) and Cross-Population Extended Haplotype Homozygosity (XP-EHH) tests
(Dataset_S6_0.5percXPEHH_AND.xlsx)

Dataset S7: SNPs under selection in Amazon populations according to Population Branch Statistic (PBS) and Cross-Population Extended Haplotype Homozygosity (XP-EHH) tests

(Dataset_S7_0.5percXPEHH_AMZ.xlsx)

Dataset S8: Highly Differentiated Variants Between Andean and Amazon Populations: Annotation from GWAs Catalog. CHR: chromosome, FST: Level of genetic differentiation between groups, A1: alternative allele, AMZ: Amazon populations, AND: Andean populations, PEL: Peruvians from Lima, EAS: East asian populations, EUR: European populations, WAFR: West African populations. (Dataset_S8_GWAs Hits.xlsx)

Dataset S9: Highly Differentiated Variants Between Andean and Amazon Populations: Annotation from PharmGKB. CHR: chromosome, FST: Level of genetic differentiation between groups, A1: alternative allele, AMZ: Amazon populations, AND: Andean populations, PEL: Peruvians from Lima, EAS: East asian populations, EUR: European populations, WAFR: West African populations. (Dataset_S9_PharmGKB.xlsx)

Dataset S10: Highly Differentiated Variants Between Andean and Amazon Populations: Annotation from Sift and Polyphen. CHR: chromosome, Wild.AA: Wild Aminoacid, Mutant.AA: Mutant Aminoacid, FST: Level of genetic differentiation between groups, A1: alternative allele, AMZ: Amazon populations, AND: Andean populations, PEL: Peruvians from Lima, EAS: East asian populations, EUR: European populations, WAFR: West African populations. (Dataset_S10_Missense.xlsx)

SI References

1. W. B. Church, A. von Hagen, “Chachapoyas: Cultural Development at an Andean Cloud Forest Crossroads” in *The Handbook of South American Archaeology*, H. Silverman, W. H. Isbell, Eds. (Springer New York, 2008), pp. 903–926.

2. I. Schellerup, Wayko-Lamas: a Quechua community in the Selva Alta of North Peru under change. *Geografisk Tidsskrift*, 199–208 (1999).

3. G. Seitz, *Cultural Discontinuity: The New Social Face of the Awajun* (Amakella Publishing, 2017).

4. J. M. Guallart, *La tierra de los cinco ríos* (Pontificia Universidad Católica del Perú, Instituto Riva Agüero, 1997).

5. L. Campbell, “Language isolates and their history” in *Language Isolates*, (Routledge, 2017), pp. 1–18.

6. S. Purcell, *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

7. W. C. S. Magalhães, *et al.*, EPIGEN-Brazil Initiative resources: a Latin American imputation panel and the Scientific Workflow. *Genome Res.* **28**, 1090–1095 (2018).

8. A. L. Price, N. A. Zaitlen, D. Reich, N. Patterson, New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).

9. F. S. G. Kehdy, *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.* **112**, 8696–8701 (2015).

10. 1000 Genomes Project Consortium, *et al.*, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

11. J. Z. Li, *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).

12. D. Reich, *et al.*, Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).

13. S. Mallick, *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

14. M. Raghavan, *et al.*, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).

15. R. E. Green, *et al.*, A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).

16. N. Patterson, *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).

17. B. K. Maples, S. Gravel, E. E. Kenny, C. D. Bustamante, RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).

18. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in

unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

19. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

20. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).

21. O. Delaneau, J. Marchini, J.-F. Zagury, A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).

22. G. Hellenthal, *et al.*, A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).

23. S. Leslie, *et al.*, The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).

24. L. van Dorp, *et al.*, Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. *PLoS Genet.* **11**, e1005397 (2015).

25. J.-C. Chacón-Duque, *et al.*, Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat. Commun.* **9**, 5388 (2018).

26. G. A. Gnecchi-Ruscone, *et al.*, Dissecting the Pre-Columbian Genomic Ancestry of Native Americans along the Andes–Amazonia Divide. *Mol. Biol. Evol.* **36**, 1254–1269 (2019).

27. D. W. Lathrap, The antiquity and importance of long-distance trade relationships in the moist tropics of pre-Columbian South America. *World Archaeol.* **5**, 170–186 (1973).

28. H. Silverman, W. Isbell, *Handbook of South American Archaeology* (Springer Science & Business Media, 2008).

29. C. Quintana, R. T. Pennington, C. U. Ulloa, H. Balslev, Biogeographic Barriers in the Andes: Is the Amotape—Huancabamba Zone a Dispersal Barrier for Dry Forest Plants? *Ann. Mo. Bot. Gard.* **102**, 542–550 (2017).

30. J. Guffroy, “Cultural Boundaries and Crossings: Ecuador and Peru” in *The Handbook of South American Archaeology*, H. Silverman, W. H. Isbell, Eds. (Springer New York, 2008), pp. 889–902.

31. J. R. Sandoval, *et al.*, The Genetic History of Peruvian Quechua-Lamistas and Chankas: Uniparental DNA Patterns among Autochthonous Amazonian and Andean Populations. *Ann. Hum. Genet.* **80**, 88–101 (2016).

32. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).

33. A. Bergström, *et al.*, A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science* **357**, 1160–1163 (2017).

34. M. Raghavan, *et al.*, Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014).

35. S. R. Browning, B. L. Browning, High-resolution detection of identity by descent in unrelated individuals. *Am. J. Hum. Genet.* **86**, 526–539 (2010).

36. D. Speed, D. J. Balding, Relatedness in the post-genomic era: is it still useful? *Nat. Rev. Genet.* **16**, 33–44 (2015).

37. E. A. Thompson, Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* **194**, 301–326 (2013).

38. S. Baharian, *et al.*, The Great Migration and African-American Genomic Diversity. *PLoS Genet.* **12**, e1006059 (2016).

39. V. Pankratov, *et al.*, East Eurasian ancestry in the middle of Europe: genetic footprints of Steppe nomads in the genomes of Belarusian Lipka Tatars. *Sci. Rep.* **6**, 30197 (2016).

40. B. L. Browning, S. R. Browning, Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).

41. S. R. Browning, B. L. Browning, Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *Am. J. Hum. Genet.* **97**, 404–418 (2015).

42. A. Gusev, *et al.*, Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009).

43. J. Haas, S. Pozorski, T. Pozorski, *The Origins and Development of the Andean State* (Cambridge University Press, 1987).

44. C. Stanish, The Origin of State Societies in South America. *Annu. Rev. Anthropol.* **30**, 41–64 (2001).

45. S. C. Stearns, R. M. Nesse, D. R. Govindaraju, P. T. Ellison, Evolutionary perspectives on health and medicine. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1691–1695 (2010).

46. E. Vasseur, L. Quintana-Murci, The impact of natural selection on health and disease: uses of the population genetics approach in humans. *Evol. Appl.* **6**, 596–607 (2013).

47. R. Lewin, *Human Evolution: An Illustrated Introduction* (John Wiley & Sons, 2009).

48. S. Fan, M. E. B. Hansen, Y. Lo, S. A. Tishkoff, Going global by adapting local: A review of recent human adaptation. *Science* **354**, 54–59 (2016).

49. F. M. Salzano, The role of natural selection in human evolution - insights from Latin America. *Genet. Mol. Biol.* **39**, 302–311 (2016).

50. X. Yi, *et al.*, Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).

51. D. L. Hartl, A. G. Clark, A. G. Clark, *Principles of population genetics* (Sinauer associates Sunderland, MA, 1997).

52. J. Goudet, Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Resour.* **5**, 184–186 (2005).

53. A. Benazzo, A. Panziera, G. Bertorelle, 4P: fast computing of population genetics statistics from large DNA polymorphism panels. *Ecol. Evol.* **5**, 172–175 (2015).

54. C. C. Cockerham, B. S. Weir, Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics* **40**, 157–164 (1984).

55. L. L. Cavalli-Sforza, Human diversity in *Proc. 12th Int. Congr. Genet*, (1969), pp. 405–416.

56. J. E. Crawford, *et al.*, Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans. *Am. J. Hum. Genet.* **101**, 752–767 (2017).

57. 1000 Genomes Project Consortium, *et al.*, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

58. I. Shlyakhter, P. C. Sabeti, S. F. Schaffner, Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics* **30**, 3427–3429 (2014).

59. S. W. Buskirk, R. E. Peace, G. I. Lang, Hitchhiking and epistasis give rise to cohort dynamics in adapting populations. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8330–8335 (2017).

60. P. C. Sabeti, *et al.*, Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).

61. P. C. Sabeti, *et al.*, Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

62. Z. A. Szpiech, R. D. Hernandez, selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Molecular Biology and Evolution* **31**, 2824–2827 (2014).

63. G. Soares-Souza, “Novas Abordagens para Integração de Bancos de Dados e Desenvolvimento de Ferramentas Bioinformáticas para Estudos de Genética de Populações,” Universidade Federal de Minas Gerais. (2014).

64. S. T. Sherry, *et al.*, dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

65. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–7 (2005).

66. B. Jassal, *et al.*, The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).

67. B. Braschi, *et al.*, Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res.* **47**, D786–D792 (2019).

68. A. Buniello, *et al.*, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

69. I. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7**, Unit7.20 (2013).

70. Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, A. P. Chan, Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).

71. R. Vaser, S. Adusumalli, S. N. Leng, M. Sikic, P. C. Ng, SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).

72. F. Hsu, *et al.*, The UCSC Known Genes. *Bioinformatics* **22**, 1036–1046 (2006).

73. M. Ashburner, *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

74. The Gene Ontology Consortium, The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

75. M. Whirl-Carrillo, *et al.*, Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92**, 414–417 (2012).

76. V. C. Jacovas, *et al.*, Selection scan reveals three new loci related to high altitude adaptation in Native Andeans. *Sci. Rep.* **8**, 12733 (2018).

77. D. N. Harris, *et al.*, Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proc. Natl. Acad. Sci. U. S. A.*, 201720798 (2018).

78. J. C. Barrett, B. Fry, J. Maller, M. J. Daly, Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).

79. A. Carré, *et al.*, When an Intramolecular Disulfide Bridge Governs the Interaction of DUOX2 with Its Partner DUOXA2. *Antioxid. Redox Signal.* **23**, 724–733 (2015).

80. W. J. Kent, *et al.*, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

81. T. J. P. Hubbard, *et al.*, Ensembl 2007. *Nucleic Acids Res.* **35**, D610–7 (2007).

82. Y. Ahmad, *et al.*, The proteome of Hypobaric Induced Hypoxic Lung: Insights from Temporal Proteomic Profiling for Biomarker Discovery. *Sci. Rep.* **5**, 10681 (2015).

83. Y. Shen, *et al.*, Ischemic preconditioning inhibits over-expression of arginyl-tRNA synthetase gene Rars in ischemia-injured neurons. *J. Huazhong Univ. Sci. Technolog. Med. Sci.* **36**, 554–557 (2016).

84. X. Cheng, H. Jiang, Long non-coding RNA HAND2-AS1 downregulation predicts poor survival of patients with end-stage dilated cardiomyopathy. *J. Int. Med. Res.* **47**, 3690–3698 (2019).

85. Y.-Z. Fu, *et al.*, Human Cytomegalovirus Tegument Protein UL82 Inhibits STING-Mediated Signaling to Evade Antiviral Immunity. *Cell Host Microbe* **21**, 231–243 (2017).

86. D. Xie, *et al.*, Exploring the associations of host genes for viral infection revealed by genome-wide RNAi and virus-host protein interactions. *Mol. Biosyst.* **11**, 2511–2519 (2015).

87. A. van der Vliet, K. Danyal, D. E. Heppner, Dual oxidase: a novel therapeutic target in allergic disease. *Br. J. Pharmacol.* **175**, 1401–1418 (2018).

88. S. Meer, Y. Perner, E. D. McAlpine, P. Willem, Extraoral plasmablastic lymphomas in a high human immunodeficiency virus endemic area. *Histopathology* (2019) https:/doi.org/10.1111/his.13964.

89. A. S. Motani, *et al.*, Evaluation of AMG 076, a potent and selective MCHR1 antagonist, in rodent and primate obesity models. *Pharmacol Res Perspect* **1**, e00003 (2013).

90. E. M. van Leeuwen, *et al.*, Genome of The Netherlands population-specific imputations

identify an ABCA6 variant associated with cholesterol levels. *Nat. Commun.* **6**, 6065 (2015).

91. A. Piehler, W. E. Kaminski, J. J. Wenzel, T. Langmann, G. Schmitz, Molecular structure of a novel cholesterol-responsive A subclass ABC transporter, ABCA9. *Biochem. Biophys. Res. Commun.* **295**, 408–416 (2002).

92. Scliar, M.O., Gouveia, M.H., Benazzo, A. *et al.* Bayesian inferences suggest that Amazon Yunga Natives diverged from Andeans less than 5000 ybp: implications for South American prehistory. *BMC Evol Biol* **14,** 174 (2014).

Anexo 2.2 Material Suplementar do artigo: "The need to diversify genomic studies: insights from Andean highlanders and Amazonians"

STAR Methods

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to the Lead Contact, Eduardo Tarazona Santos (eduardotarazona@ufmg.br; edutars@gmail.com).

Materials availability

There are no materials related with the present Comment.

Data and code availability

Genotype data have been deposited in the European Genome-phenome Archive (EGA), https://www.ebi.ac.uk/ega/home (accession nos. EGAD00010001958, EGAD00010001990, EGAD00010001991, EGAD00010001992), as described in the Key Resources Table. They are available upon request if access is granted. To request access, contact Eduardo Tarazona Santos (eduardotarazona@ufmg.br; edutars@gmail.com).

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

LDGH Data: The data from the Laboratory of Human Genetic Diversity were collected by the Universidad Peruana Cayetano Heredia from two populations in the Andes (one Quechua and one Aymara-speaking), and two in the Amazon Yunga populations (Ashaninkas and Matsiguenkas), considering only individuals >18 years old (adults). The Quechua-speaking population was sampled in a large area comprising two continuous regions, Ayacucho and Apurimac, for this reason, these individuals were grouped as Quechuas_AA. In the case of the Aymara-speaking population, individuals were collected near the Titicaca lake shore in the Puno region. This sampling was conducted under approval of the Institutional Reviews Boards from the Universidad Peruana Cayetano Heredia, Asociación Benéfica PRISMA, Universidade Federal de Minas Gerais, and Johns Hopkins University.

INS Data: The present work is part of the Peruvian Genome Diversity Project (PGDP) of the Peruvian Institute of Health (Instituto Nacional de Salud - INS). The protocol for this study was approved by The Research and Ethics Committee (OI003-11 and OI-087-13) of the INS. Fifteen Native American populations (INS data) were sampled as part of the PGDP, including populations from the South Pacific Peruvian Coast (SPPC), Andean and Amazon regions. In the SPPC, samples were collected from Moche and Tallanes populations, and in the Andes from Qeros,

Chopccas, Jaqarus, and Uros. In the Amazon, samples were collected from 6 populations of the Amazon Yunga (Chachapoyas, Lamas, Awajun, Candoshi, Ashaninkas, and Matsiguenkas), a transition region between the Andean mountains and the Lower Amazon, and two populations from the lower Amazon (Shipibo-Conibo, Matses, and Nahuas). All native participants were adults (> 18 years old), and required to have their four grandparents born in the selected native ancestral population.

METHOD DETAILS

Genotyping and Quality Control: A total of 289 individuals from INS samples were genotyped using the Illumina Human Omni array 2.5M at the INS. The total number of genotyped SNPs was 2,391,739. The 127 samples from LDGH were genotyped by the Illumina facility (California) using the HumanOmni2.5–8v1 array.

Quality control was performed for both datasets using the PLINK 1.7 software (5) and in-house scripts (8), removing individuals and SNPs with >10% of missing data and loci with 100% of heterozygosity, no chromosomal information and A/T-C/G genotypes. After QC, the datasets were merged resulting in à total of 2,077,858 SNPs for 418 individuals from 19 populations. Both groups, INS and LDGH datasets, include independent samples of the Ashaninka population from the same region, for this reason, we merged these individuals into a unique Ashaninka group leading to a total of 18 populations (Table S2).

To filter related individuals, we removed SNPs in high linkage disequilibrium (LD) in each population, as it affects the inferences of relatedness, with PLINK 1.7 using the flag --indep-pairwise with the following parameters: size of window: 200 SNPs; step between continuous windows: 25 SNPs, LD threshold (r^2): 0.1.

Merging datasets: We merged the LDGH and INS datasets with data from individuals of European (IBS), African (YRI), and East Asian (CDX) ancestries from the 1000 Genomes project (2) using the software PLINK 1.7 (5). This dataset has 1,927,769 SNPs and 673 individuals.

Population Structure using genotype-based methods: The genetic clustering using genotype-based methods presented in Figure 1 was originally published by Borda et al. (1).

Before performing population structure analyses that require unrelated individuals, we used autosomal SNPs to estimate the kinship coefficients (Φij) for each pair of individuals for each population using the option –genome in PLINK 1.7. We considered a threshold Φij 0.25 to define relatedness (first-degree relatives: parent-offspring pair and full sibling). We further used a network approach (9) to identify which individuals should be removed, preserving a maximum number of unrelated individuals. After applying the kinship filter, we kept 358 individuals.

The ADMIXTURE (7) was run in 50 replicates in unsupervised mode for different values of K, starting from 2 ancestral clusters until cross-validation error started to increase. The results for K=5 (corresponding to the lowest cross-validation error) are represented as pie charts for each population in Figure 1 and mean values are in supplementary table S1.

The ADMIXTURE analysis (7) assumes independence among SNPs, for this reason, linked SNPs were removed using PLINK 1.7 with the option indep-pairwise 200 25 0.4 for each dataset. The final number of SNPs for this analysis was 625,736. These analyses were presented in Borda et al. (2020), where the complete results and the detailed methodology are described.

Variants annotation with PharmGKB: The Clinical Annotations Data from PharmGKB was downloaded from https://www.pharmgkb.org/downloads in June 2023. We calculated the frequencies of 58 SNPs with clinical annotations levels 1 or 2 in our dataset with the software PLINK 1.7 (supplementary table S2). The 95% Confidence Interval of the frequencies was calculated in R environment by performing binomial random sampling 10000 times, conditioning on the observed allele frequency and sample size.

F_{ST} calculation: F_{ST} was estimated with the 4P software, which is based on the Weir and Cockerham (1984) algorithm (4). F_{ST} estimates between groups of population were made considering two groupings: (i) all Andean (Jacarus, Chopccas, Quechuas, Qeros, Puno, Uros) vs all Amazonian populations (Chachapoyas, Awajun, Candoshi, Lamas, Matses, Shipibo, Ashaninkas, Matsiguenkas); and (ii) all Andean vs adjacent Amazonian populations (Shipibo, Shimaa, Matsiguenkas, Ashaninkas) (supplementary table S2).

Genotype Frequencies and Hardy-Weinberg expected values calculation: Genotype frequencies and counts were calculated with software PLINK 1.7 for each population. The expected counts of alleles and genotypes were calculated with Hardy-Weinberg equation for SNPs rs2231142 and rs9923231. We calculated the chi-square statistic by hand and considered its significance ($p<0.05$) threshold of 3.84, for one degree of freedom (supplementary table S3).

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
dataset_raw_NativesAmericans_LDGH_august2020	Borda et al. (1)	EGAD00010001958
NativeAmericans_InstitutoNacionaldeSalud_hg38_autosomic_2group	Borda et al.(1)	EGAD00010001990
NativeAmericans_InstitutoNacionaldeSalud_hg38_autosomic_3group	Borda et al. (1)	EGAC00001001729
1000 Genomes Project	The 1000 Genomes Project Consortium (2)	https://www.internationalgenome.org/
Variant and Clinical Annotations Data	Whirl-Carrillo et al. (3)	https://www.pharmgkb.org/downloads
Software and algorithms		
4P	Benazzo et al. (4)	https://github.com/anbena/4p
Plink	Purcell et al. (5)	http://pngu.mgh.harvard.edu/purcell/plink/
R	R Core Team (6)	https://www.r-project.org/
ADMIXTURE	Alexander et al. (7)	https://dalexander.github.io/admixture/

Key resources table

References:

1. Borda, V., Alvim, I., Mendes, M., Silva-Carvalho, C., Soares-Souza, G.B., Leal, T.P., Furlan, V., Scliar, M.O., Zamudio, R., Zolini, C., et al. (2020). 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. *117*, 32557–32565.
2. 1000 Genomes Project Consortium, et al., A global reference for human genetic variation. Nature 526, 68–74 (2015).
3. Whirl-Carrillo, M., et al., Pharmacogenomics knowledge for personalized medicine. Clin. Pharmacol. Ther. 92, 414–417 (2012).
4. Benazzo, A., Panziera, A., Bertorelle, G. 4P: fast computing of population genetics 55 statistics from large DNA polymorphism panels. Ecol. Evol. 5, 172–175 (2015).
5. Purcell, S. et al., PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
6. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
7. Alexander, D.H., Novembre, J., Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655-64. (2009).
8. Magalhães, W. C. S. et al., EPIGEN-Brazil Initiative resources: a Latin American imputation panel and the Scientific Workflow. Genome Res. 28, 1090–1095 (2018).

9. Leal, T.P. et al. NAToRA, a relatedness-pruning method to minimize the loss of dataset size in genetic and omics analyses. Comput Struct Biotechnol J. 2022 Apr 9;20:1821-1828.

Supplemental Tables:

Supplementary table S1, Populational ancestry corresponding to Figure 1A populations. (Supplemental_Table_S1.docx)

Supplementary Table S2, Allele frequencies and Fst values for the 19 pharmacogenetic clinically relevant SNVs analyzed in Andean and Amazonian populations, as well as in continental populations (Supplemental_Table_S2.docx)

Supplementary Table S3, Population genetics statistics for the pharmacogenetic variants ABCG1 - rs2231142 and VKORC1 - rs9923231 in the studied populations. (Suplemental_Table_S3.docx)