

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Departamento de Biologia Geral
Programa de Pós-Graduação em Genética

Marla Mendes de Aquino

The Genetic Mosaic Beyond European Populations

Belo Horizonte

2022

Marla Mendes de Aquino

The Genetic Mosaic Beyond European Populations

Tese apresentada ao
Departamento Biologia Geral do
Instituto de Ciências Biológicas
da Universidade Federal de
Minas Gerais como requisito
parcial para a obtenção do título
de Doutora em Genética.

Orientador: Eduardo Martin
Tarazona Santos

Co-orientador: Victor Borda Pua

Belo Horizonte

2022

043

Aquino, Marla Mendes de.

The Genetic Mosaic Beyond European Populations [manuscrito] / Marla Mendes de Aquino. – 2022.

183 f. : il. ; 29,5 cm.

Orientador: Eduardo Martin Tarazona Santos. Co-orientador: Victor Borda Pua.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética.

1. Genética Populacional. 2. Variação Genética. I. Santos, Eduardo Martin Tarazona. II. Pua, Victor Octavio Borda. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 575



UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
 Programa de Pós-Graduação em Genética

ATA DE DEFESA DE DISSERTAÇÃO / TESE

ATA DA DEFESA DE TESE	155/2022
	entrada
Marla Mendes de Aquino	1º/2018
	CPF: 081.842.076-62

Às quatorze horas do dia **19 de abril de 2022**, reuniu-se a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: "**The Genetic Mosaic Beyond European Populations**", requisito para obtenção do grau de Doutora em **Genética**. Abrindo a sessão, o Presidente da Comissão, **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:

Prof./Pesq.	Instituição	CPF	Indicação
Eduardo Martin Tarazona Santos	UFMG	012.494.056-02	Aprovada
Maria Bernadete Lovato	UFMG	965.561.378-04	Aprovada
Michel Satya Naslavsky	Universidade de São Paulo	009.053.224-44	Aprovada
Maria Luiza Petzl Erler	Universidade Federal de Paraná	230.588.899-68	Aprovada
Victor Octavio Borda Pua	UFMG	700.751.756-06	Aprovada
Carlos Eduardo Guerra Amorim	California State University Northridge	998.707.461-87	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, 19 de abril de 2022.

Eduardo Martin Tarazona Santos

Maria Bernadete Lovato

Michel Satya Naslavsky

Maria Luiza Petzl Erler

Victor Octavio Borda Pua

Carlos Eduardo Guerra Amorim

Assinatura dos membros da banca examinadora:



Documento assinado eletronicamente por **Michel Satya Naslavsky, Usuário Externo**, em 19/04/2022, às 17:44, 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 **Victor Octavio Borda Pua, Usuário Externo**, em 19/04/2022, às 17:46, 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 19/04/2022, às 18:37, 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 **Carlos Eduardo Guerra Amorim, Usuário Externo**, em 19/04/2022, às 20:00, 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 **Maria Luiza Petzl Erler, Usuária Externa**, em 20/04/2022, às 15:25, 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 **Maria Bernadete Lovato, Professora do Magistério Superior**, em 21/04/2022, às 10:51, 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 **1391240** e o código CRC **B6D172E1**.



UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Genética

FOLHA DE APROVAÇÃO

"The Genetic Mosaic Beyond European Populations "

Marla Mendes de Aquino

Tese aprovada pela banca examinadora constituída pelos Professores:

Eduardo Martin Tarazona Santos
UFMG

Maria Bernadete Lovato
UFMG

Michel Satya Naslavsky
Universidade de São Paulo

Maria Luiza Petzl Eler
Universidade Federal de Paraná

Victor Octavio Borda Pua
UFMG

Carlos Eduardo Guerra Amorim
California State University Northridge

Belo Horizonte, 19 de abril de 2022.



Documento assinado eletronicamente por **Michel Satya Naslavsky, Usuário Externo**, em 19/04/2022, às 17:44, 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 **Victor Octavio Borda Pua, Usuário Externo**, em 19/04/2022, às 17:46, 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 19/04/2022, às 18:37, 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 **Carlos Eduardo Guerra Amorim, Usuário Externo**, em 19/04/2022, às 20:00, 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 **Maria Luiza Petzl Erler, Usuária Externa**, em 20/04/2022, às 15:24, 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 **Maria Bernadete Lovato, Professora do Magistério Superior**, em 21/04/2022, às 10:51, 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 **1391262** e o código CRC **09360836**.

Agradecimentos

Pela paciência digna de um monge, por cada discussão, conselho, e apoio incondicional, dedico não apenas o agradecimento dessa tese, mas todo conhecimento por mim adquirido nesses últimos 6 anos aos meus orientadores Eduardo Tarazona e Víctor Borda.

Agradeço também minha mãe solo, que não só me deu a vida, me criou com todo amor domundo, me permitiu ser tudo que sou hoje e construiu meu caráter, mas também enfrentou todos os obstáculos que apareceram para me dar o privilégio de poder seguir meus sonhos. Ao meu filho, que me ensinou que mesmo achando que não vai dar, eu vou fazer dar sim, por ele. Filho, mesmo todos achando que uma mulher precisa escolher entre ter vida profissional ou pessoal, você me mostrou todos os dias que mesmo sendo difícil, tudo é muito mais legal com você, e é justamente você que me inspira e me faz fazer o impossível. Você nunca foi um empecilho, mas sim a cereja do bolo da minha vida.

Ao meu marido, que ao meu lado virou um homem incrível, que nunca me ajudou, mas sempre carregou metade de todos os nossos problemas, como deve ser. Você me recarregou de amor todos os dias dessa caminhada, o que me ajudou a conseguir levar todo o doutorado de forma leve e agradável, sabendo que no fim do dia eu sempre teria você. Obrigada também por dividir seus pais comigo, eles formaram a base onde eu sempre corria nos momentos de dificuldade ao longo desse processo, obrigada sogros por serem muito mais que apenas sogros, vocês foram também meus pais e amigos nesses últimos anos.

Não poderia deixar também de agradecer a toda equipe do LDGH, em especial ao meu coorientador Victor, que me ensinou na prática todas as análises, que me incentivou e nunca deixou que meu tesão pela ciência se apagasse. Mesmo com um sério problema de comunicação, você foi além de um colega de trabalho mas um grande amigo que pretendo trabalhar em colaboração por muito tempo ainda. Ao Thiago que me fez rir todos os dias que conversamos, você foi o grande alívio cômico do meu doutorado, e também faz parte dos grandes amigos que conquistei nesse período. A Isabela e Carol por me acompanharem em todos os sofrimentos do artigo de nativos e em vários projetos que tive a honra de dividir com elas. A Camila que é a peça mais fundamental do laboratório, sem a qual os artigos não seriam publicados, e nem a paz reinaria, obrigada Camila por traduzir a gente pro Eduardo eo Eduardo pra gente.

A toda minha família por me treinar a habilidade comunicação, argumentação e de divulgação científica com todas as fake news enviadas no whatsapp. Amo vocês, mas lutarei até meus últimos dias para que entendam como a ciência é incrível e como a obediência cega não nos leva a nenhum progresso como sociedade. Mas dentre meus parentes que salvam, agradeço em particular a minhas primas Virginia e Elaine por compartilhar desse sentimento de vergonha e decepção. E às minhas tias Nivia, Soraia, Elisa, e Eliane por me incentivarem todos os dias a continuar estudando.

Agradeço em especial as Universidades Federais de excelência que me formaram, deixo registrado aqui meu desejo de que elas continuem assim, acessíveis para outras filhas de empregada doméstica que assim como eu deseje contribuir com o conhecimento científico do mundo.

Summary

Agradecimientos	3
List of Figures	7
Chapter 1	7
Chapter 2	7
Chapter 3	8
Chapter 4	10
List of Tables	10
Chapter 1	10
Chapter 2	10
Chapter 3	11
Chapter 4	12
List of Attachments	12
List of abbreviations	13
Resumo	14
Abstract	15
Introduction	16
Chapter 1 - Review: The history behind the mosaic of the Americas	19
Introduction	19
Methodology	19
Published article	21
Complementary Discussion	27
Chapter 2 - The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture	29
Introduction	29
Methodology	30
Section 1: Sampling, quality control and Datasets	30
1.3. Genotyping and Quality Control:	30
1.4. Merging datasets:	31
Section 2: Genetic relationships in Western South America	33
2.1. Methods	33
2.1.1. Population Structure using genotype based methods	33
Natives 1.9M dataset	34
Natives 500K dataset	34
2.2. Conclusions	35
Section 4: Dating the between-population homogenization of the arid Andes	37
4.1. Methods	37

4.1.1 Identical-by-descent segment analysis	37
RefinedIBD	38
Natives 1.9M dataset	38
Natives 500K dataset	39
4.1.2. IBDne	39
Natives 1.9M dataset	39
4.2. Conclusions	39
Section 5: Genetic Differentiation and Natural Selection in the Andes and Amazon	42
5.1. Introduction	42
5.2. Methods	43
Natural Selection Candidate SNPs	43
5.2.1. Population Branch Statistic	43
5.3. Conclusions	45
Published article	55
Chapter 3 - Identifying signatures of Natural Selection in Indian populations	65
Introduction	66
Materials and Methods	68
Datasets	68
Natural Selection Analysis	69
Methods Based on Population Differentiation	70
Methods Based on Linked Variation	71
Methods Based on Site Frequency Spectrum	72
Composite Methods	72
ASMC	73
Results	74
Discussion	77
Chapter 4 - Application of Polygenic Risk Scores (PRS) on admixed Brazilian Populations	105
Introduction	105
Materials and Methods	106
Sampling	106
Base Data	106
Target Data	108
Quality Control	109
Base Data	109
Target Data (EPIGEN-Brazil)	110
PRS calculation and p-value cutoff	110
Clumping and Thresholding	110
Statistical tests	111
Results	115
Discussion	117

Attachments	125
Collaborative Papers	125
1) The Iberian legacy into a young genetic xeroderma pigmentosum cluster in central Brazil. 126	
2) Origins, Admixture Dynamics, and Homogenization of the African Gene Pool in the Americas.	138
3) Human-SARS-CoV-2 interactome and human genetic diversity: TMPRSS2-rs2070788, associated with severe influenza, and its population genetics caveats in Native Americans	148
4) Tracing the distribution of European lactase persistence genotypes along the Americas.	152
5) A large Canadian cohort provides insights into the genetic architecture of human hair colour	167
Final remarks	179

List of Figures

Chapter 1

Figure 1. Infographic of the key events of the evolutionary history of the Americas.

Chapter 2

Figure 1. The genetic and geographic landscape for Western South American natives.

Figure 2. Evolution of IBD sharing between the Pacific Coast, Central Andes, Amazon Yunga and Amazon and its relationship with the archaeological chronology of the Andes.

Figure 3. Natural selection illustrated by a long-range haplotype plot and a Manhattan plot.

Figure S1. Geographical distribution for the 18 Peruvian Native populations sampled, plus the 65 sampled Native American populations and public data sets.

Figure S2. ADMIXTURE analysis for 18 Native American populations, as well as Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project.

Figure S3. Principal Component Analysis for 18 Native American Peruvian populations and Iberian individuals (IBS) from 1000 Genomes Project.

Figure S4. ADMIXTURE analysis for 18 Natives American Peruvian populations, Guatemala samples, Native Americans from Raghavan et al. 2015 and the Simons Project Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 500K Dataset).

Figure S5. Principal Component Analysis for 18 Native American Peruvian populations, Guatemala samples, Native Americans from Raghavan et al. 2015 and the Simons Project and Iberian (IBS) populations from 1000 Genomes Project.

Figure S20. Key historical events of Peruvian prehistory in four longitudinal regions: Peruvian Coast, Andes, Amazon Yunga and Amazon.

Figure S21. Heatmap representation of the shared Identical by descent (IBD) segments among Native Americans of the Natives 1.9M dataset.

Figure S22. IBDNe analysis to infer the dynamics of the effective population size (N_e) from 4 generations ago to the last 50 generations for the Andean populations.

Figure S23. Demographic model of the Andean, Amazonian and East Asian populations.

Figure S24. PBSn mean values Andean populations.

Figure S25. PBSn mean values Andean populations.

Figure S26. PBSn mean values for windows of 20 SNPs with 5 SNPs of overlapping in Andean populations.

Figure S27. PBSn mean values for windows of 20 SNPs with 5 SNPs of overlapping in Amazon populations.

Chapter 3

Figure 1. Schematic representation of approach to identify putative selective regions. We applied six different methods to identify outliers (top 1%, 0.5% and 0.01% results) and selected regions that were observed in all population groups and were outliers for at least two independent methods. Additionally, we performed analyses using a novel coalescence-based method implemented in the program ASMC

Figure 2. Overview of our results. A) Distribution of the number of regions identified for each chromosome, for all thresholds (1%, 0.5% and 0.1%); B) Distribution of the number of genes located within putative selective regions using 1%, 0.5%, and 0.1% thresholds for each chromosome; C) Distribution of the number of genes located within putative selective regions identified with two or three methods for each chromosome, for all thresholds (1%, 0.5% and 0.1%); D) Percentage of signals in the top 1% shared by pair of populations (WM Castes, WH Tribes, WH full sample, and India 1KG). The reference group is indicated in the Y-axis; for example, for the PBS method, 8.72% of the signals found in the WM Caste group are also found in the WM Tribe group, but just 7.78% of the signals identified in the WM Tribe group are found in the WM Caste group; E) Percentage of signals in the top 1% shared by different methods. The reference method is indicated in the Y-axis; for example, 33.5% of the signals identified using PBS are also observed with xpEHH.

Supplementary Figure 1. The geographical location of the samples analyzed in this study.

Supplementary Figure 2. Depicts a schematic representation of our approach. Additional details of the methods used in our analyses are provided in the main text.

Supplementary Figure 3. Admixture graphs showing our two approaches. A) Admixture graph including a preIndia group resulting from admixture from a European and an Asian source. B) Admixture graph including preTribe and preCaste groups as a result of admixture between a European and Asian source.

Supplementary Figure 4. ASMC, detailing in blue, regions found as possible signatures of natural selection in our study, the numbers in green indicate other studies where those regions were reported (1: Metsupalu et al. 2011, 2: Suo et al. 2012, 3: Karlsson et al. 2015, 4: Liu et al. 2017, 5: Perdomo-Sabogal et al, 2019). In red we show the genes present in regions with high enrichment but that was not found as an outlier in our study. A) chromosome 1, B) chromosome 2, C) chromosome 3, D) chromosome 4, E) chromosome 5, F) chromosome 6, G) chromosome 7, H) chromosome 8, I) chromosome 9, J) chromosome 10, K) chromosome 11, L) chromosome 12, M) chromosome 13, N) chromosome 14, O) chromosome 15, P) chromosome 16, Q) chromosome 17, R) chromosome 18, S) chromosome 19, T) chromosome 20, U) chromosome 21, V) chromosome 22.

Supplementary Figure 5. ASMC results for chromosome 2, showing a zoom in the region with the biggest enrichment of recent coalescence events (96.2Mb to 98.4Mb), on the top we describe all the genes located within this region.

Supplementary Figure 6. ASMC results for chromosome 16, showing at the top, the genes identified in the region between 29.46Mb and 32.07Mb. In blue we list the genes within the top 0.5% signals and in red the genes within the top 0.1% signals identified for at least one method. The red squares highlight genes that have also been identified in other studies, as detailed in supplementary table 1.

Supplementary Figure 7. ASMC results for chromosome 6, showing a zoom in the region from 29.6Mb to 32.8Mb, on the top we describe the genes located within this region.

Supplementary Figure 8. ASMC results for Chromosome 1 and 2, detailing in blue, regions found as possible signatures of natural selection in our study, the numbers in green indicate other studies where those regions were reported (1: Metsupalu et al. 2011, 2: Suo et al. 2012, 3: Karlsson et al. 2015, 4: Liu et al. 2017, 5: Perdomo-Sabogal et al, 2019). In red we show the genes present in regions with high enrichment but that was not found as an outlier in our study.

Chapter 4

Figure 1. Continental admixture of the EPIGEN Brazil populations, adapted from Kehdy et al., (2015): Brazilian regions, the studied populations, and their continental individual ancestry bar plots. N represents the numbers of EPIGEN individuals in the Original Dataset.

Figure 2. Barplot of prediction R-squared values vs p-value thresholds for each test. Colour gradient from blue for lowest R-squared value to red for highest R-squared value. The legend shows the R-squared values in the order of the p-value thresholds. A) GIANT x Bambuí. B) GIANT x Pelotas C) GIANT x Salvador D) PAGE x Bambuí E) PAGE x Pelotas F) PAGE x Salvador.

Figure 3: Flowchart of our PRS pipeline.

Figure 4. Scatterplot between observed Body-Mass Index (BMI) in the X-axis vs inferred PRS (Y-axis), for each test with the European ancestry proportion showed by the gradient colour of the points. A) GIANT x Bambuí, B) GIANT x Pelotas, C) GIANT x Salvador, D) Giant x All, E) PAGE x Bambuí, F) PAGE x Pelotas, G) PAGE x Salvador, H) PAGE x All.

Supplementary Figure 1. Histograms of the frequency distribution of PRS values for each test. In the X-axis are the PRS values and in the Y-axis are the frequency. A) GIANT x Bambuí. B) GIANT x Pelotas C) GIANT x Salvador D) PAGE x Bambuí E) PAGE x Pelotas F) PAGE x Salvador.

List of Tables

Chapter 1

Chapter 2

Dataset S1: Description of 19 studied Native American populations from Peruvian National Institute of Health and from Laboratory of Human Genetic Diversity.

Dataset S2: List of all samples included in the Native 500K dataset.

Dataset S3: List of all samples included in the Native 230K dataset.

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

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

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

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

Dataset S8: Highly Differentiated Variants Between Andean and Amazon Populations: Annotation from GWAS Catalog.

Dataset S9: Highly Differentiated Variants Between Andean and Amazon Populations: Annotation from PharmGKB.

Dataset S10: Highly Differentiated Variants Between Andean and Amazon Populations: Annotation from Sift and Polyphen.

Chapter 3

Supplementary Table 1: Details of the 435 genes results of our scanning for natural selection signatures. Showing also the shared signals with 1kpg high coverage data and the overlap with other studies.

Supplementary Table 2: Details of the 97 regions results of our scanning for natural selection signatures. Showing also the shared signals with 1kpg high coverage data and the overlap with other studies.

Because of the size and dynamics, those Supplementary tables 1 and 2 are in the following link:

<https://docs.google.com/spreadsheets/d/1J3xtV31hLrswFd2dOU8zgua9FJJCnIsk293lF6YVOvw/edit?usp=sharing>

Chapter 4

Table 1. Correlation results for each analyzed test. On the lines in red and green, we have the lowest values and highest values respectively.

Supplementary Table 1. Results of Pearson, Kendall and Spearman correlations compared to basedata from the GIANT Consortium and the PAGE Study. The tests were separated into three according to the percentage of European ancestry that increase from left to right. In Bambuí, each part has 298 individuals, in Pelotas, almost 1217 and 415 from Salvador.

Supplementary Table 2. Correlation results for mixed datas in the analyzed test.

Supplementary Table 3. Results of Pearson, Kendall and Spearman correlations compared to basedata from the GIANT Consortium and the PAGE Study. The tests were separated into three according to the percentage of European ancestry that increase from left to right. In Pelotas+Bambuí, each part has 1507 individuals, in Bambuí+Salvador almost 711 each, 1633 for the cohort from Salvador+Pelotas and 1920 for Salvador+Pelotas+Bambuí.

Supplementary Table 4. Results of Pearson, Kendall and Spearman correlations comparing the PRS calculated and different ancestries.

List of Attachments

Collaborative Papers

- 1) The Iberian legacy into a young genetic xeroderma pigmentosum cluster in central Brazil.
- 2) Origins, Admixture Dynamics, and Homogenization of the African Gene Pool in the Americas.
- 3) Human-SARS-CoV-2 interactome and human genetic diversity: Tmprss2-rs2070788, associated with severe influenza, and its population genetics caveats in Native Americans
- 4) Tracing the distribution of European lactase persistence genotypes along the Americas.
- 5) A large Canadian cohort provides insights into the genetic architecture of human hair colour

List of abbreviations

LDGH - Laboratory of Human Genetic Diversity
DNA - Deoxyribonucleic acid
IBD - Identity-by-Descent
PCA - Principal Analysis Component
GWAS - Genome-Wide Association Studies
CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
PGDP - Peruvian Genome Diversity Project
INS - Peruvian Institute of Health (Instituto Nacional de Salud)
UFMG - Universidade Federal de Minas Gerais
SPPC - South Pacific Peruvian Coast **U****Dataset** - Unrelated Dataset
LD - Linkage Disequilibrium
IBS - Iberian populations in Spain **LWK** - Luhya in Webuye, Kenya **CHS** - Han Chinese South
CDX - Chinese Dai in Xishuangbanna, China
CHB - Han Chinese in Beijing, China **CEU** - Northern Europeans from Utah **YRI** - Yoruba in Ibadan, Nigeria
NCI - National Cancer Institute
EM - Expectation-Maximization
NE - Effective Population Size
LH - Late Horizon
LIP - Late Intermediate Period
MH - Middle Horizon
EIP - Early Intermediate Period
EH - Early Horizon
IP - Initial Period
cM - Centimorgans
MAF - Minimum Allele Frequency **GIH** - Gujarati Indians in Houston, TX **ITU** - Indian Telugu in the UK
EAS - East Asian

Mb - Megabase

GRFs - gene regulatory factors

ZNF - zinc-finger genes

KRAB-ZNF - zinc-finger genes with a Krüppel-associated box

PBS - Population Branch Statistic

xpEHH - Cross Population Extended Haplotype Homozygosity

iHS - Integrated Haplotype Score

CLR - Composite likelihood ratio

GRoSS - Graph-aware Retrieval of Selective Sweeps **ASMC** -

Ascertained Sequentially Markovian Coalescent **PRS** - Polygenic

Risk Score

Resumo

A inclusão da diversidade humana nos estudos genéticos contribui não apenas para a melhor compreensão da nossa história, mas também para resolver questões biomédicas, como suscetibilidade a doenças, resposta a tratamentos médicos e elucidação de fenótipos complexos. Geneticistas sabem, há mais de uma década, que o foco em pessoas de ancestralidade europeia agrava as disparidades de saúde, e os avanços genéticos não são aplicados a quem mais precisa. Assim, esta tese apresenta trabalhos que contribuem para mudar esse cenário, buscando aumentar a visibilidade da riqueza genética que existe além das populações de ancestralidade europeia. Aqui, apresento resultados de pesquisas que incluem populações miscigenadas e não-miscigenadas com ancestralidades diversas, desde nativos sulamericanos, até africanos e asiático. O primeiro capítulo é uma mini revisão da história da humanidade no continente americano, contada pela genética de populações atuais e antigas, desde a chegada dos primeiros seres humanos até o processo de miscigenação, onde também apresentamos discussões complementares sobre as atualizações das descobertas feitas desde a publicação desse artigo em 2020. No capítulo 2, apresentamos um artigo publicado pelo meu grupo que resolve questões importantes da história genética andina e amazônica, por meio de métodos de genética de populações e seleção natural. No capítulo 3 me aprofundo nos métodos de varredura dessas assinaturas adaptativas e crio um pipeline de estudo de seleção natural multi-metodológico para diminuir ao máximo a chance de se obter falsos-positivos. O Capítulo 4 é uma adição importante para as populações geneticamente negligenciadas e seria uma parte complementar ao nosso capítulo 1, onde eu aplico os avanços com scores poligenicos de riscos as

populações miscigenadas, como a Brasileira. Explorar a diversidade da genética humana não é importante apenas para as populações negligenciadas, mas também é essencial para aumentar a capacidade de fazer novas descobertas. Assim, apresentamos nos anexos, cinco artigos exemplificando como o mosaico genômico de populações não europeias é uma rica fonte de informações de diferentes perspectivas, incluindo aspectos médicos, históricos e evolutivos. O trabalho aqui apresentado mostra de diversas formas como o esforço de geração de dados a partir de populações sub-representadas traz importantes contribuições para a ciência.

Abstract

The inclusion of human diversity in genetic studies contributes not only to a better understanding of our history but also to solve biomedical issues such as disease susceptibility, response to medical treatments and elucidation of complex traits. Geneticists have known for more than a decade that focusing on people of European ancestry exacerbates health disparities, and genetic advances can not be applied to those who need it most. Thus, this thesis presents works that contribute to change this scenario, seeking to increase the visibility of the genetic richness that exists beyond populations of European ancestry. Here, I present research findings that include admixed and non-admixed populations with diverse ancestries, from South American natives to Africans and Asians. The first chapter is a mini-review of the history of humanity on the American continent, told by the genetics of modern and ancient populations, from the arrival of the first humans to the process of admixture, where we also present complementary discussions about the updates of the discoveries made since the publication of this article in 2020. In chapter 2, we present an article published by my group that solves important questions in the Andean and Amazon genetic history, through methods of population genetics and natural selection. In chapter 3, I devote myself to the methods of scanning these adaptive signatures and create a multi-methodological pipeline for the study of natural selection to minimize the chance of obtaining false positives. Chapter 4 is an important addition to genetically neglected populations and would be a complementary part of our chapter 1, where we applied advances with polygenic risk scores to admixed populations such as the Brazilian. Exploring the diversity of human genetics is not only important for neglected populations, it is also essential to increase the capacity to make new discoveries. Thus, in the attachments, we present five articles exemplifying how the genomic mosaic of non-European populations is a rich source of information from different perspectives, including medical, historical and

evolutionary aspects. The work presented here shows in different ways how the effort to generate data from underrepresented populations makes important contributions to science.

Introduction

Our genetics influences is correlated with several factors in our life, some more obvious, such as whether you are more or less susceptible to a condition, and the colour of your eyes. But it also is correlated with some less obvious conditions, for example, it is your genetic ancestry that influences whether your genetics will be studied or ignored. A lot of criticism has been made about the differences in genetic studies between Europeans and non-Europeans, and the consequences of this are known, like the inequitable access to precision medicine for those with the highest burden of disease in poor countries (Gurdasani et al., 2019; Teh, 2019; Wojcik et al., 2019, Weissbrod et al., 2021).

The fact of the majority of genetic variants discovery efforts have been based on data from populations of European ancestry generates a bias in drug development, in clinical guidelines and exacerbates healthcare disparities, and disease prevalence. This bias can happen because critical variants may be missed if they have a low frequency or even didn't exist in European populations, besides other biological factors like differences in linkage disequilibrium between populations for example (Need and Goldstein 2009; Bustamante et al., 2011; Popejoy and Fullerton, 2016; Wojcik et al., 2019).

To avoid increased healthcare disparities, it is important that the scientific community adopts measures to ensure that populations of diverse ancestry are included in genomic studies and that countries of non-European ancestry give financial incentives for your own researchers to increase the knowledge about the genetic diversity and the genetic architecture of diseases of their peoples. During my Ph.D., I dedicated myself to increasing the visibility of the genetic richness that exists beyond populations of homogeneous European ancestry. Here, I present research results that include populations admixed and non-admixed with ancestries from native South America, Africans, South and East Asia.

In this sense, we need to have a strategic approach that uncovers from the basic genetic aspects to the methods that can directly be applied to the improvement of the quality of life of these people. So here, we focus on the study of population structure, the variants that can have a higher tendency to be transmitted to the next generation, and finally the genetic factors that can influence health policy management in these regions.

As an example of this approach, we present in chapter 1 a bibliographic mini-review that covers all the human history in the Americas, including the admixture and natural selection aspects that construct the genetic specificities of these people. This review was focused on the papers published between 2018 and 2020, and missed some important discussions, as the possible Polynesian contact with Natives in South America, and the medical specificities generated by the genetic admixed of the American people that can influence even in the calculation of polygenic risk score.

So here, in my thesis, I will try to complement these gaps, bringing first, our mini-review telling the history of humans in America. In chapter 2, I present a paper published by my group in 2020 that solve important questions of Andean and Amazonian genetic history, through methods of populations genetics (Principal Component Analysis, and ADMIXTURE), and natural selection (cross-population extended haplotype homozygosity - xpEHH, Population Branch Statistical - PBS). And in the following chapter 3, I present a work that is still in development, where I show a complete methodology to seek signatures of natural selection, bringing methods based on Population Differentiation, Linked Variation, Site Frequency Spectrum, Composite Methods, besides a recently developed method that uses admixture graphs to infer signatures of selection in specific branches of the graphs (Refoyo-Martínez et al., 2018), as well as a coalescence-based method (Palamara et al., 2018), all of them applied in the genetic neglected population of India. And in chapter 4 we present an effort to include admixed people in the recent polygenic risk score (PRS) revolution. This work is in the initial phase, in which we apply PRS to Brazilian target data and measure the influence of non-European ancestry on the accuracy of this calculation.

In addition to developing the main project on the populational history, adaptation, and medical factors of genetically neglected people, during my Ph.D. at the Laboratory of Human Genetic Diversity (LDGH), I had the opportunity to participate in different projects applying population genetics for historical and biomedical studies. So, I attached 5 articles that had my contribution:

i. "The Iberian legacy into a young genetic xeroderma pigmentosum cluster in central Brazil." That is an example of a study of genetically neglected people that takes into account aspects from the genetic history to the medical factors of the population. In this paper the genetic

knowledge about xeroderma pigmentosum was applied to a Brazilian population, making also an interplay with the historical migrations that form this population.

ii. "Origins, Admixture Dynamics, and Homogenization of the African Gene Pool in the Americas." Where we make a comparison of genetics and demography data to infer the patterns of the African migration to the Americas. Uncover an important part of the genetic history of those people.

iii. "Human-SARS-CoV-2 interactome and human genetic diversity: TMPRSS2-rs2070788, associated with severe influenza, and its population genetics caveats in Native Americans." In this paper, our group illustrates the importance of population genetics and of sequencing data in the design of genetic association studies in different human populations for human/SARS-CoV-2 interactome genes

iv. "Tracing the distribution of European lactase persistence genotypes along the Americas." Where we bring the famous genetic question about the evolution of lactase persistence in the different global populations.

v. "A large Canadian cohort provides insights into the genetic architecture of human hair colour." In this paper, my group conduct a population structure analysis in Canadian samples to improve the main goal of producing a GWAS meta-analysis of hair colour in a large cohort.

Chapter 1 - Review: The history behind the mosaic of the Americas

Introduction

The possibility to work with ancient DNA, besides all the new techniques to work at a genome level, led to a "boom" of genetic studies that aim to infer the human evolutionary history. In this context, the Laboratory of Human Genetic Diversity (LDGH) group was invited by Prof. Sarah Tishkoff and Prof. Joshua Akey, editors of Volume 62 of the journal *Current Opinion in Genetics & Development*, to write a mini-review that organized and summarized the most important discoveries published since 2018 about human evolution in the Americas, including also our point of view. This invitation reflects the efforts of our group to study the genetics of Latin American populations and their rich genetic, cultural and anthropological history.

We divide our review into three main questions: (i) How was the discovery and dispersion of the first humans by the Americas, (ii) How was the post-Columbian admixture, and (iii) Inferences about natural selection in the Americas human populations. As the first author of this paper, I actively participated in all sections, but I concentrated mostly on section one, and I had the collaborations of Ph.D. Victor Borba to write section II, and from Ph.D. Isabela Alvin in section III, with the orientation and editing of Prof. Eduardo Tarazona.

I was mostly in charge of Section I because inferences about the demographic history of Native American populations was an area where I devoted a large part of my master's and doctoral work. As I showed in chapter II of this thesis, I worked with the dynamic history of Andes and Amazon populations, using Identical-by-descent segments (IBD) to infer patterns of gene flow between and inside those groups. I was also able to collaborate with section II, because of my previous experience with genetic ancestry analysis, such as PCA, ADMIXTURE (Alexander et al., 2009) and admixture graphs. Besides my work with Population Branch Statistical, which helped me to understand better natural selection analysis and collaborate in section III.

Methodology

We organized the papers published between 2018 and 2020 about the conquest, dispersion and natural selection of humans in the Americas and summarized the most important information about each article.

To graphically summarize the milestones of the human evolutionary history in the Americas we prepared an image that shows all human steps in America since the arrival through the Beringia pathway, until the present days, which was used as our graphical abstract. This figure was chosen as the cover figure of this edition of Current Opinion in Genetics and Development (<https://www.sciencedirect.com/journal/current-opinion-in-genetics-and-development/vol/62/suppl/C>).



ELSEVIER



The history behind the mosaic of the Americas

Mendes¹, Isabela Alvim¹, Victor Borda² and Eduardo Tarazona-Santos¹

Focusing on literature published in 2018–2020, we review inferences about: (i) how ancient DNA is contributing to clarify the peopling of the Americas and the dispersal of its first inhabitants, (ii) how the interplay between environmental diversity and culture has influenced the genetic structure and adaptation of Andean and Amazon populations, (iii) how genetics has contributed to our understanding of the Pre-Columbian Tupi expansion in Eastern South America, (iv) the subcontinental origins and dynamics of Post-Columbian admixture in the Americas, and finally, (v) episodes of adaptive natural selection in the American continent, particularly in the high altitudes of the Andes.

Addresses

¹Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

²Laboratório de Bioinformática, LABINFO, Laboratório Nacional de Computação Científica (LNCC), Petrópolis, Rio de Janeiro, Brazil

Corresponding author:

Tarazona-Santos, Eduardo (edutars@icb.ufmg.br)

Current Opinion in Genetics and Development 2020, 62:72–77

This review comes from a themed issue on Genetics of human origin

Edited by Sarah Tishkoff and Joshua Akey

<https://doi.org/10.1016/j.gde.2020.06.007>

0959-437X/© 2020 Elsevier Ltd. All rights reserved.

Introduction

Focusing on post-2018 literature, we review studies on population genetics of the Americas on: (i) the demographic history of Native Americans, (ii) the genetics of post-Columbian admixture, and (iii) adaptive natural selection.

The peopling of the Americas

Archaeology, cranial and dental morphology, protein polymorphisms and DNA-uniparental markers have been traditional sources of knowledge for the peopling of the Americas [1–4]. Three recent advances were: development of model-based statistical methods that simultaneously consider genetic drift and gene flow; access to

genome-wide data; and more recently, studies on ancient DNA (aDNA).

The first milestone in the evolution of Native Americans is the split of their ancestors from East Asians, 36 KYA (Kilo-Years Ago), likely in Northeast Asia, with gene flow between the two differentiating groups persisting until ~25 KYA (likely still in Asia) [5]. This is consistent with results from Raghavan *et al.* [6]: the Asian population that was ancestral to modern Native Americans resulted from admixture between a population related to the Upper Paleolithic Mal'ta boy skeleton from south-central Siberia and an East Asian related population, ancestral to the Han from China.

A second milestone in the settlement of the Americas was the Beringian standstill [7,8], a period when the Ancestral Native American populations were isolated from Asian groups, which may have lasted between 4.6 KY [9] and 15 KY [10]. Moreno-Mayar *et al.* [5], studying aDNA from Upward Sun River dated around 11.5 KYA, inferred that an ancient Beringian population diverged from the ancestor of Native Americans 22–18 KYA, possibly in: (i) Northeast Asia/Siberia (i.e. before the Beringian standstill, which implies that the standstill population was structured); or (ii) East Beringia (during/after the Beringian standstill) [11]. The large number of archaeological sites dated 20 KYA or older found in northeast Asia compared to East Beringia [12,13] supports the first scenario.

Estimates of the effective population size for the founding population of the Americas is around a few hundred individuals [14]. Two possible southward routes for the first Americans were: (i) The Pacific Coast, where ice retreated ~16 KYA [11], supported by the oldest radiocarbon dates of the Cooper's Ferry site [15] and; (ii) the ice-free corridor between the Cordilleran and Laurentide ice sheets [13,16]. Demographic modeling using aDNA suggests a third milestone split of the Native American Ancestral group into two branches associated with Northern Native Americans (Ancestral B *in sensu* Scheib *et al.* [18]) and Central/Southern Native Americans (Ancestral A *in sensu* Scheib *et al.* [18]), dated 17–14.6 KYA [5,17,18]. This split likely occurred in the region between Eastern Beringia and the unglaciated North America [18,19].

The divergence between Central and South Amerindians still needs robust dates that consider back migration from northern South America to Central America. For now, we have the mtDNA-based estimates of 13–19 KYA for the

divergence between Peruvian and Panamanian natives, which is consistent with the oldest South American site of Monte Verde in Chile [20], and the estimates based on 1000 Genomes Project admixed individuals of ~ 12 – 13 KYA [21]. The dispersal across South America was rapid, occurring within a 1.5 KY span [22]. Possible routes of dispersion were the Pacific Coast [22–24] and the Atlantic Coast [25]. A method that evaluates the minimum number of contributing ancestral sources that better explain the genetic diversity in South Amerindians (qpWave, [26]) suggests four contributing populations [27]: three of them related to the Ancestral A (*in sensu* Scheib *et al.* [18]) population: (i) the Ancestral A; (ii) another Ancestral A, related to the Clovis Anzick-1 individual; (iii) another Ancestral A related to aDNA from Californian Channel Islands individuals (specifically to Andean populations) and finally (iv) a minor debatable contribution present in a few Brazilian native isolates, that Skoglund *et al.* [59] and Moreno-Mayar *et al.* [19] attribute to an Australasian-related source. However, studies of aDNA [27] and mitogenomes [22] did not replicate this biogeographic association.

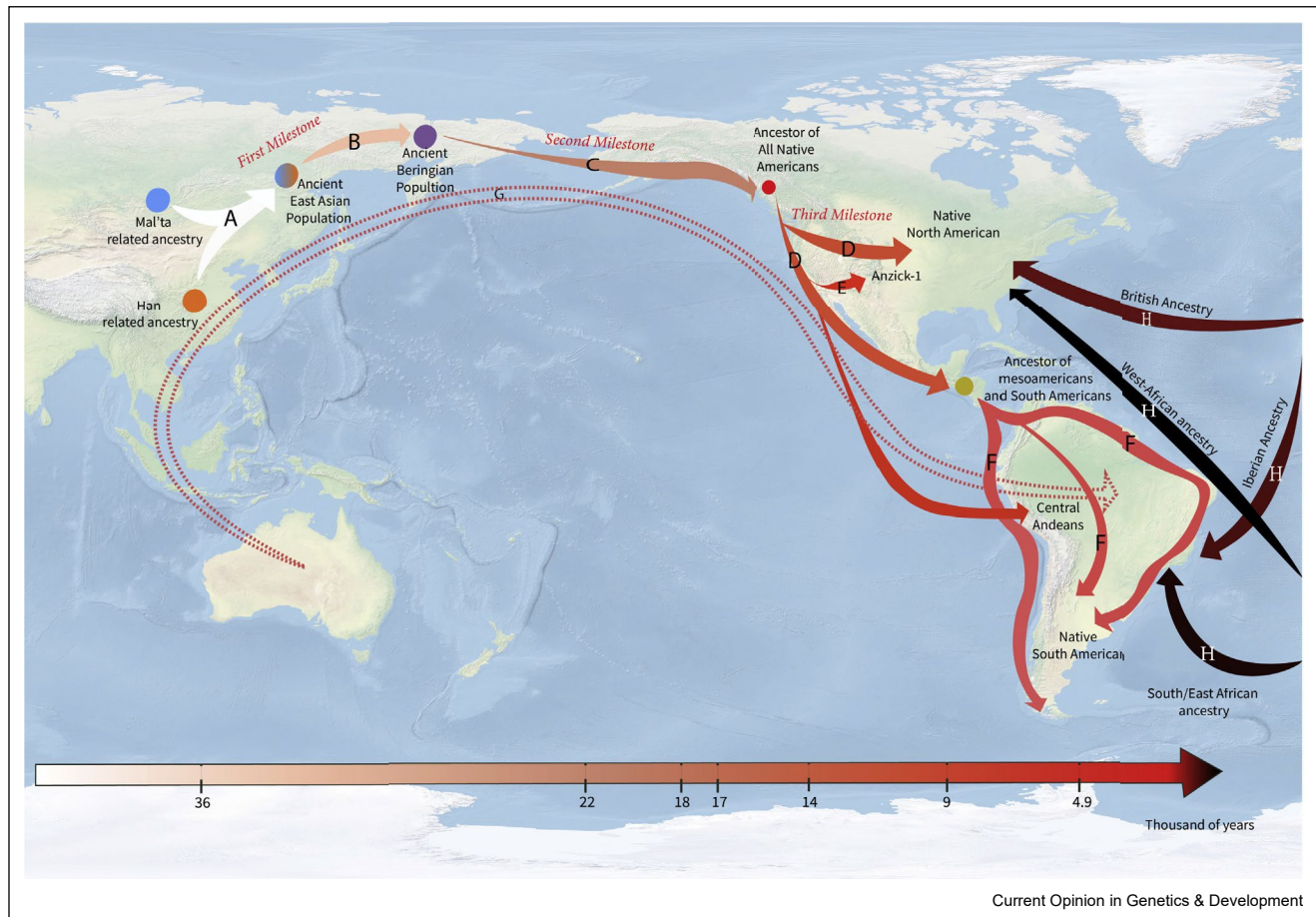
Western South America, being one of the worldwide cradles of civilization [28], has been a focus of population genetics studies [24,29,30,31,32]. The well known Inca Empire was the *tip of the iceberg* of a process that lasted for thousands of years. Notably, Western South America hosts a rich environmental diversity encompassing a desert coast, the Andean mountains and plateaus, as well as its adjacent Amazonian tropical forest. Populations from these biogeographic regions (Coast, Andean and Amazonian populations) split around 8–12 KYA [24,29]. Barbieri *et al.* [31] have reported episodes of gene flow between Amazonian populations, which suggest that this region is not necessarily characterized by highly isolated groups, as previously thought. Borda *et al.* [32] have revealed how the interplay between environmental diversity and culture influenced the genetic structure and adaptation of Andeans and Amazonians. Borda *et al.* [32] show that the between-population homogenization of the central-southern Andes and its differentiation with respect to Amazonian populations of similar latitudes observed by Tarazona-Santos *et al.* [33] do not extend northward. The east-west gene flow between the north coast of Peru, Andes, and Amazonia was concomitant with cultural and socioeconomic interactions suggested by archeology. This geographic pattern of genetic diversity mimics the environmental and cultural differentiation between the fertile north Andes, where altitudes are lower; and the arid south, where the Andes are higher and act as a barrier to gene flow. Also, the genetic homogenization between the populations of the arid Andes is not only due to migration during the Inca Empire or subsequent periods, but started at least as early as the expansion of the important pre-Inca Wari Empire (600–1000 years ago) [32].

The Tupi were one of the most numerous ethnic groups living in the XVth Century on the Brazilian coast, but in the XVIIIth Century they were almost extinct. Castro e Silva *et al.* [34] studied one of the few remnant coastal Tupi populations and tested two alternative hypotheses about the North-to-South Tupi Expansion (by-litoral versus by-inland). Genomic data supports the first hypothesis: a Pre-Columbian migration from Amazon to the northeast coast, giving rise to Tupi' coastal populations, and a single migration southward that originated the Guaraní' people from Brazil and Paraguay. Castro e Silva *et al.* [34] dedicated their article to Francisco M Salzano, who co-authored the paper, and sadly passed away in 2018 being 91 years old, and was one of the most influential and appreciated scholars studying the human biology of the American continent populations (Figure 1).

The mosaic of the Americas

Because Latin American ancestry results from Post-Columbian admixtures between Europeans, Africans, and Native Americans, their genomes are like a mosaic of fragments (i.e. *tracts*) deriving from those ancestries [35]. The shift from inferences of admixture proportion to inferences of the admixture dynamics in population genetics is like the shift from the era of photographs to that of filmmaking. One family of methods to infer admixture dynamics relies on the distribution of the *tract* lengths (i.e. contiguous DNA blocks inherited from a parental population) [36–38]. Kehdy *et al.* [38] revealed that the low Native American ancestry of admixed Brazilians, characterized by short *tracts*, was almost entirely introduced immediately after the first arrival of Europeans into the Americas, which is consistent with the decimation of Native Americans in Brazil. Noteworthy, methods to infer admixture dynamics do not infer when the immigrants arrived (a demographic event), but date intensification of biological admixture. For instance, Harris *et al.* [29] inferred that current Peruvian *mestizos* (predominantly Native American) living in cities founded 400–500 years ago by Spaniards harbor the signature of biological admixture with Spaniards occurring only ~ 200 years ago. This is because individuals with predominant European or Native American ancestries may have coexisted without admixing for generations, or because Native American ancestors of current *mestizos* may predominantly have arrived in the cities (where admixture occurred) only recently from rural populations where they were isolated. Similar dates of ~ 250 years ago have been inferred for intensification of admixture in Mexico, Colombia, Brazil, Chile, and Peru, using the *chromopainter-based* method based on haplotypes [39], that relies on the pattern of linkage disequilibrium generated by admixture. Also using the *chromopainter-based* methods to analyze the African Diaspora, Gouveia *et al.* [40] captured a continental trend: in most of the Americas, intercontinental admixture intensification occurred between 1750 and 1850, which correlates with the peak

Figure 1



Infographic of the key events of the evolutionary history of the Americas.

The color of the bottom horizontal arrow represents the temporal scale from past to present. Authors that present results and discuss each event are evidenced in the gray square. The dashed arrow is related to a controversial event. A) Moreno-Mayar *et al.* [5^{*},19]; Waters *et al.* [11]. B) Raghavan *et al.* [17]; Moreno-Mayar *et al.* [5^{*},19]. C) Potter *et al.* [13]; Moreno-Mayar *et al.* [5^{*},19]; Waters [11]; Pinotti *et al.* [9]. D) Potter *et al.* [13]; Waters [11]; Davis *et al.* [15]; Scheib *et al.* [18]; Moreno-Mayar [5^{*},19]. E) Posth *et al.* [27]. F) Gravel *et al.* [21]; Moreno-Mayar *et al.* [5^{*},19]; Harris *et al.* [29^{*}]. G) Skoglund *et al.* [59]; Moreno-Mayar [5^{*},19]. H) Gouveia *et al.* [40^{**}]; Baharian *et al.* [60]; Lindo *et al.* [24].

of the slave trade from Africa. Furthermore, Gouveia *et al.* [40^{**}] performed a systematic comparison of population history inferred from genomic data to historical demographic records from SlaveVoyage database (<https://www.slavevoyages.org>). This kind of comparison of genetics and demography data may be interpreted as a tribute to Luca Cavalli-Sforza, who passed away in 2018. Indeed, systematic integration of genetic and demographic data has solid roots in human populations genetics, partly in the work by Cavalli-Sforza more than sixty years ago in the Parma Valley [41]. However, this kind of comparison has become rare in the era of human population genomics.

Recent studies are detecting sources of admixture at a subcontinental geographic resolution. In Latin America Chacón-Duque *et al.* [39] differentiated Spanish

contribution to Spanish-speaking populations from Portuguese contribution to Brazil and detected South/East Mediterranean ancestry across Latin America that likely reflects the clandestine colonial migration of Christian converts of Jews origin (Conversos). The roots of the African Diaspora is also being mapped [38,40^{**},42–44]: (i) West-Central African ancestry is predominant in the Americas, (ii) Western African ancestry and South/Eastern African Bantu-associated ancestries show a longitudinal pattern, with the former more common in northern latitudes of the Americas and the latter ancestry more common in southern South America. An interesting result by Gouveia *et al.* [40^{**}] is that while African intra-population diversity was not lost during the African Diaspora, there was a between-population homogenization of the African gene pool in the Americas. With respect to Native

American ancestry, studies in different countries show that in Mestizo populations, Native American admixture predominantly originates from nearby indigenous groups [29,39,45].

Inferences about natural selection

The most commonly used methods to infer natural selection are still based on allele frequencies or their spectra or on long-range haplotypes/linkage disequilibrium. Some gene sets are more often reported in Native American populations: (i) the immune system appears very frequently [24,32,46–51], including signals of convergent selection in tropical forests from Amazon and Africa (CCL28) [52]; (ii) adaptation of lipid metabolism [50,53], for example, SCP2 in Amazon [52], KCNH1 in Alaska [51]; and (iii) adaptations to extreme environments such as high soil concentration of Arsenic in the Atacama Desert, where variants in the gene AS3MT were selected for efficiency in arsenic metabolism [54] and with cold (HS3ST4) in Alaska [51].

Adaptation to high altitude by Andean populations is a classic topic of physiological and anthropological studies in South America, revealing hematological and respiratory adaptations to hypoxia [55,56]. Genome-wide scans for natural selection in Andean populations have identified genes related to the hypoxia-inducible factors pathway (EGLN1,ET-1), oxidative stress (FAM213A) and cardiovascular function and development (DST,NOS2, VEGFB,TBX5,HAND2-AS1) [24,32,56].

In admixed populations of the Americas, genomic regions or gene sets for which the contribution of European, African or Native American ancestry is beyond what would be expected given their genome-wide proportions, are signatures of natural selection by adaptive introgression. Using this concept, Norris *et al.* [47] identified signals for immune system pathways such as T cell receptors signaling, antigen processing and presentation, and cytokine-receptors interaction, shared between four populations from Peru, Colombia, Puerto Rico, and Mexico. This gene-set approach is interesting, but it poses statistical challenges related to significance tests for gene-sets that we still need to better understand to avoid false positives.

Prospects

Because current studies are mostly based on very few individuals for each site, which may bias the results, we still need aDNA studies based on more individuals. Another approach is that used by Mas-Sandoval *et al.* [57], who explored an interesting reconstruction of Native American haplotypes from admixed individuals, which is important in places where indigenous populations no longer exist. Methodologies to study admixture dynamics would benefit if they include complexities such as ancestry-dependent assortative mating, including

ancestry-related sex bias, which may result in interesting questions, methods, and conclusions [58]. Signatures of polygenic natural selection remain to be explored in Native Americans, as well as functional validation of natural selection claims using both candidate genes and the developing arsenal of functional genomics.

Conflict of interest statement

Nothing declared.

Acknowledgements

To write this review we were inspired by previous work and ideas of present and past members of the *Laboratório de Diversidade Genética Humana*. We thank Vinicius Furlan, Carolina Silva-Carvalho, Hana'isa Sant'Anna, Thiago P Leal, Fabricio Santos, Maria Cátira Bortolini, Nelson Fagundes, Sandro Bonatto and Tabita Hunemeier for suggestions. The authors would like to recognize the collaboration between Indigenous peoples with scientists that make all these studies possible. MM was supported by a Mitacs Globalink Research Award (FR37903) and by *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) (88887.474324/2020-00). IA (88882.349066/2019-01), and VB (88882.195664/2018-01) also were supported by *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) and ET-S by *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) from Brazil.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Greenberg JH, Turner CG, Zegura SL, Campbell L, Fox JA, Laughlin WS, Weiss KM, Woolford E: The settlement of the Americas: a comparison of the linguistic, dental, and genetic evidence [and Comments and Reply]. *Curr Anthropol* 1986, 27:477-497.
 2. Pena SDJ, Santos FR, Bianchi NO, Bravi CM, Carnese FR, Rothhammer F, Gerelsaikhan T, Munkhtuja B, Oyunsuren T: A major founder Y-chromosome haplotype in Amerindians. *Nat Genet* 1995, 11:15-16.
 3. Dillehay Tom: *The Settlement of the Americas: a New Prehistory*. Basic Books; 2001.
 4. Gonza'lez-Jose' R, Bortolini MC, Santos FR, Bonatto SL: The peopling of America: craniofacial shape variation on a continental scale and its interpretation from an interdisciplinary view. *Am J Phys Anthropol* 2008, 137:175-187.
 5. Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspina AS, Sikora M *et al.*: Terminal Pleistocene Alaskangenome reveals first founding population of Native Americans. *Nature* 2018, 553:203-207.
- This paper and that of Posth *et al.* [27], published together, were pivotal to reveal the contribution of aDNA to our understanding of the settlement and early demographic history of the American continent. It used demographic models that consider both drift and gene flow and important samples from Late Pleistocene found in Alaska to infer the genetic composition of the founding population in the Americas and date milestone population splits such as that involving the Beringia Standstill.
6. Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, Korneliusen TS, Grønnow B, Appelt M, Gulløv HC, Friesen TM *et al.*: The genetic prehistory of the new world Arctic. *Science*(80-) 2014, 345.
 7. Szathmari EJ: mtDNA and the peopling of the Americas. *Am J Hum Genet* 1993, 53:793-799.
 8. Bonatto SL, Salzano FM: Diversity and age of the four major mtDNA haplogroups, and their implications for the peopling of the new world. *Am J Hum Genet* 1997, 61:1413-1423.
 9. Pinotti T, Bergström A, Geppert M, Bawn M, Ohasi D, Shi W, Lacerda DR, Solli A, Norstedt J, Reed K *et al.*: Y chromosome

- sequences reveal a short Beringian standstill, rapid expansion, and early population structure of Native American founders. *Curr Biol* 2019, 29:149-157.e3.
10. Graf KE, Buvit I: Human dispersal from Siberia to Beringia assessing a Beringian standstill in light of the archaeological evidence. *Curr Anthropol* 2017, 58:S583-S603.
 11. Waters MR: Late Pleistocene exploration and settlement of the Americas by modern humans. *Science (80-)* 2019, 365.
 12. Buvit I, Izuho M, Terry K, Konstantinov MV, Konstantinov AV: Radiocarbon dates, microblades and Late Pleistocene human migrations in the Transbaikal, Russia and the Paleo-Sakhalin-Hokkaido-Kuril Peninsula. *Quat Int* 2016, 425:100-119.
 13. Potter BA, Baichtal JF, Beaudoin AB, Fehren-Schmitz L, Haynes CV, Holliday VT, Holmes CE, Ives JW, Kelly RL, Llamas B *et al.*: Current evidence allows multiple models for the peopling of the Americas. *Sci Adv* 2018, 4:1-9.
 14. Fagundes NJR, Tagliani-Ribeiro A, Rubicz R, Tarskaia L, Crawford MH, Salzano FM, Bonatto SL: How strong was the bottleneck associated to the peopling of the Americas? New insights from multilocus sequence data. *Genet Mol Biol* 2018, 41:206-214.
 15. Davis LG, Madsen DB, Becerra-Valdivia L, Higham T, Sisson DA, Skinner SM, Stueber D, Nyers AJ, Keen-Zebert A, Neudorf C *et al.*: Late Upper Paleolithic occupation at Cooper's Ferry, Idaho, USA, ~16,000 years ago. *Science (80-)* 2019, 365:891-000897.
 16. Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A, Kashani BH, Ritchie KH, Scozzari R, Kong QP *et al.*: Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. *Curr Biol* 2009, 19:1-8.
 17. Raghavan M, Steinrucken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Avila-Arcos MC, Malaspina A-S *et al.*: Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science (80-)* 2015, 349 aab3884-aab3884.
 18. Scheib CL, Li H, Desai T, Link V, Kendall C, Dewar G, Griffith PW, Morseburg A, Johnson JR, Potter A *et al.*: Ancient human parallel lineages within North America contributed to a coastal expansion. *Science (80-)* 2018, 360:1024-1027.
 19. Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T *et al.*: Early human dispersals within the Americas. *Science* 2018, 362.
 20. Fuselli S, Tarazona-Santos E, Dupanloup I, Soto A, Luiselli D, Pettener D: Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. *Mol Biol Evol* 2003, 20:1682-1691.
 21. Gravel S, Zakharia F, Moreno-Estrada A, Byrnes JK, Muzzio M, Rodriguez-Flores JL, Kenny EE, Gignoux CR, Maples BK, Guiblet W *et al.*: Reconstructing Native American migrations from whole-genome and whole-exome data. *PLoS Genet* 2013, 9.
 22. Brandini S, Bergamaschi P, Fernando Cerna M, Gandini F, Bastaroli F, Bertolini E, Cereda C, Ferretti L, Gómez-Carballa A, Battaglia V *et al.*: The Paleo-Indian entry into South America according to mitogenomes. *Mol Biol Evol* 2018, 35:299-311.
 23. Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C *et al.*: Genetic variation and population structure in Native Americans. *PLoS Genet* 2007, 3:2049-2067.
 24. Lindo J, Achilli A, Perego UA, Archer D, Valdiosera C, Petzelt B, Mitchell J, Worl R, Dixon EJ, Fifield TE *et al.*: Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity. *Proc Natl Acad Sci U S A* 2017, 114:4093-0004098.
 25. Gómez-Carballa A, Pardo-Seco J, Brandini S, Achilli A, Perego UA, Coble MD, Diegoli TM, Alvarez-Iglesias V, Martínez-Torres F, Olivieri A *et al.*: The peopling of South America and the trans-Andean gene flow of the first settlers. *Genome Res* 2018, 28:767-779.
 26. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N *et al.*: Reconstructing Native American population history. *Nature* 2012, 488:370-374.
 27. Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, Rohland N, Narasimhan V, Adamski N, Bertolini E *et al.*: Reconstructing the deep population history of central and South America. *Cell* 2018, 0:1-13.
- This paper and those of Moreno-Mayar *et al.* [5, 19], published together, were pivotal to reveal the contribution of aDNA to our understanding of the settlement and early demographic history of the American continent. It used demographic models that consider both drift and gene flow and important samples from Late Pleistocene as well as genetic exchange between South and North America that impacted the genetic composition of the Indigenous populations in South America. We also highlight the relationship between ancient individuals from the California Channel Islands and Central Andes populations, and the shared ancestry between Clovis-associated Anzick-1 skeleton with Chilean, Brazilian, and Belizean individuals.
28. Solis RS, Haas J, Creamer W: Dating Caral, a preceramic site in the Supe Valley on the central coast of Peru. *Science (80-)* 2001, 292:723-726.
 29. Harris DN, Song W, Shetty AC, Levano KS, Ca'ceres O, Padilla C, Borda V, Tarazona D, Trujillo O, Sanchez C *et al.*: Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proc Natl Acad Sci U S A* 2018, 115:E6526-E6535.
- The largest sequence-based study of Native-Americans individuals (150 Peruvian whole-genome sequences), including admixed individuals and studying the admixture dynamics of Peruvian populations. They make inferences on the demographic history of these populations since they first split 12 KYA into the present, covering both Pre-Columbian and Post-Columbian times, and focusing on the dynamics between the three Peruvian geographic regions: Andean, Amazon and Pacific Coast.
30. Gnecci-Ruscione GA, Sarno S, De Fanti S, Gianvincenzo L, Giuliani C, Boattini A, Bortolini E, Di Corcia T, Sanchez Mellado C, Davila Francia TJ *et al.*: Dissecting the pre-Columbian genetic ancestry of Native Americans along the Andes-Amazonia divide. *Mol Biol Evol* 2019, 36:1254-1269.
 31. Barbieri C, Barquera R, Arias L, Sandoval JR, Acosta O, Zurita C, Aguilar-Campos A, Tito-Álvarez AM, Serrano-Osuna R, Gray RD *et al.*: The current genomic landscape of Western South America: Andes, Amazonia, and Pacific Coast. *Mol Biol Evol* 2019, 36:2698-2713.
 32. Borda V, Alvim I, Aquino MM, Silva C, Soares-Souza GB, Leal TP, Scliar MO, Zamudio R, Zolani C, Padilla C *et al.*: The genetic structure and adaptation of Andean highlanders and Amazonian dwellers is influenced by the interplay between geography and culture. *bioRxiv* 2020 <http://dx.doi.org/10.1101/2020.01.30.916270>.
- The authors present new data and describe in detail the genetic structure of Andes and Amazon populations, interpreting the results in terms of the environmental diversity and cultural developments in Western South America, one of the cradles of civilization. The study capitalizes inferences on the genetic structure of populations to design genomewide scans of natural selection in the Andes and the Amazonia.
33. Tarazona-Santos E, Carvalho-Silva DR, Pettener D, Luiselli D, De Stefano GF, Labarga CM, Rickards O, Tyler-Smith C, Pena SD, Santos FR: Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. *Am J Hum Genet* 2001, 68:1485-1496.
 34. Castro e Silva MA, Nunes K, Lemes RB, Mas-Sandoval A, Guerra Amorim CE, Krieger JE, Mill JG, Salzano FM, Bortolini MC, Pereira da C *et al.*: Genomic insight into the origins and dispersal of the Brazilian coastal natives. *Proc Natl Acad Sci U S A* 2020, 117:2372-2377.
 35. Soares-Souza G, Borda V, Kehdy F, Tarazona-Santos E: Admixture, genetics and complex diseases in Latin Americans and US Hispanics. *Curr Genet Med Rep* 2018, 6:208-223.
 36. Gravel S: Population genetics models of local ancestry. *Genetics* 2012, 191:607-619.
 37. Liang M, Nielsen R: The lengths of admixture tracts. *Genetics* 2014, 197:953-967.
 38. Kehdy FSG, Gouveia MH, Machado M, Magalhães WCS, Horimoto AR, Horta BL, Moreira RG, Leal TP, Scliar MO, Soares-

- Souza GB *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, 112:8696-8701.
39. Chaco'n-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acun'a-Alonzo V, Barquera R, Quinto-Sa'nchez M, Go'mez-Valde's J, Everardo Mart'nez P, Villamil-Ram'rez H *et al.*: Latin Americans show wide-spread converso ancestry and imprint of local Native ancestry on physical appearance. *Nat Commun* 2018, 9.
 40. Gouveia Mateus H *et al.*: Origins, admixture dynamics and homogenization of the African gene pool in the Americas. *Mol Biol Evol* 2020, 37:1647-1656.
- A continental analysis that infers the subcontinental origin of the African gene pool of the Americas. They compare genomic and historical demographic data related to the African diaspora into the Americas.
41. Cavalli-Sforza Luigi Luca *et al.*: *Consanguinity, Inbreeding, and Genetic Drift in Italy*. Princeton University Press; 2013.
 42. Fortes-Lima C, Bybjerg-Grauholm J, Marin-Padr3n LC, Gomez-Cabezas EJ, Bækvad-Hansen M, Hansen CS, Le P, Hougaard DM, Verdu P, Mors O *et al.*: Exploring Cuba's population structure and demographic history using genome-wide data. *Sci Rep* 2018, 8:1-13.
 43. Fortes-Lima C, Mtetwa E, Schlebusch C: Unraveling African diversity from a cross-disciplinary perspective. *Evol Anthropol* 2019, 28:288-292.
 44. Ongaro L, Scliar MO, Flores R, Raveane A, Marnetto D, Sarno S, Gnecci-Ruscione GA, Alarco'n-Riquelme ME, Patin E, Wangkumhang P *et al.*: The genomic impact of European colonization of the Americas. *Curr Biol* 2019, 29:3974-3986.e4.
 45. Moreno-Estrada A, Gignoux CR, Fern'andez-Lo'pez JC, Zakharia F, Sikora M, Contreras AV, Acun'a-Alonzo V, Sandoval K, Eng C, Romero-Hidalgo S *et al.*: The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science (80-)* 2014, 344:1280-1285.
 46. Hsueh WC, Bennett PH, Esparza-Romero J, Urquidez-Romero R, Valencia ME, Ravussin E, Williams RC, Knowler WC, Baier LJ, Schulz LO *et al.*: Analysis of type 2 diabetes and obesity genetic variants in Mexican Pima Indians: marked allelic differentiation among Amerindians at HLA. *Ann Hum Genet* 2018, 82:287-299.
 47. Norris ET, Wang L, Conley AB, Rishishwar L, Marin'o-Ram'rez L, Valderrama-Aguirre A, Jordan IK: Genetic ancestry, admixture and health determinants in Latin America. *BMC Genomics* 2018, 19.
 48. Oliveira MLG de, Veiga-Castelli LC, Marcorin L, Debortoli G, Pereira ALE, Fracasso NC de A, Silva G do V, Souza AS, Massaro JD, Sim3es AL *et al.*: Extended HLA-G genetic diversity and ancestry composition in a Brazilian admixed population sample: implications for HLA-G transcriptional control and for case-control association studies. *Hum Immunol* 2018, 79:790-799.
 51. Calonga-Soli's V, Malheiros D, Beltrame MH, De Brito Vargas L, Dourado RM, Issler HC, Wassem R, Petzl-Erler ML, Augusto DG: Unveiling the diversity of Immunoglobulin Heavy Constant Gamma (IGHG) gene segments in Brazilian populations reveals 28 novel alleles and evidence of gene conversion and natural selection. *Front Immunol* 2019, 10.
 50. 3vila-Arcos MC, McManus KF, Sandoval K, Rodriguez-Rodr'guez JE, Villa-Islas V, Martin AR, Luisi P, Pen'aloza-Espinosa RI, Eng C, Huntsman S *et al.*: Population history and gene divergence in Native Mexicans inferred from 76 human exomes. *Mol Biol Evol* 2020, 37:994-1006 <http://dx.doi.org/10.1093/molbev/msz282>.
 - Reynolds AW, Mata-Mi'guez J, Miro'-Herrans A, Briggs-Cloud M, Sylestine A, Barajas-Olmos F, Garcia-Ortiz H, Rzhetskaya M, Orozco L, Raff JA *et al.*: Comparing signals of natural selection between three indigenous North American populations. *Proc Natl Acad Sci U S A* 2019, 116:9312-9317.
 52. Amorim CEG, Daub JT, Salzano FM, Foll M, Excoffier L: Detection of convergent genome-wide signals of adaptation to tropical forests in humans. *PLoS One* 2015, 10:1-19.
 53. Mychaleckyj JC, Havt A, Nayak U, Pinkerton R, Farber E, Concannon P, Lima AA, Guerrant RL: Genome-wide analysis in Brazilians reveals highly differentiated native American genome regions. *Mol Biol Evol* 2017, 34:559-574.
 54. Vicun'a L, Fernandez MI, Vial C, Valdebenito P, Chaparro E, Espinoza K, Ziegler A, Bustamante A, Eyheramendy S: Adaptation to extreme environments in an admixed human population from the Atacama desert. *Genome Biol Evol* 2019, 11:2468-2479.
 55. Tarazona-Santos E, Lavine M, Pastor S, Fiori G, Pettener D: Hematological and pulmonary responses to high altitude in Quechuas: a multivariate approach. *Am J Phys Anthropol* 2000, 111:165-176.
 56. Julian CG, Moore LG: Human genetic adaptation to high altitude: evidence from the Andes. *Genes (Basel)* 2019, 10:150.
 57. Mas-Sandoval A, Arauna LR, Gouveia MH, Barreto ML, Horta BL, Lima-Costa MF, Pereira AC, Salzano FM, Hu'nemeier T, Tarazona-Santos E *et al.*: Reconstructed lost Native American populations from Eastern Brazil are shaped by differential J3/Tupi ancestry. *Genome Biol Evol* 2019, 11:2593-2604.
 58. Zaidi AA, Makova KD: Investigating mitonuclear interactions in human admixed populations. *Nat Ecol Evol* 2019, 3:213-222.
 59. Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hu'nemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D: Genetic evidence for two founding populations of the Americas. *Nature* 2015 <http://dx.doi.org/10.1038/nature14895>.
 60. Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, Bustamante CD, Kenny EE, Williams SM, Aldrich MC *et al.*: The great migration and African-American genomic diversity. *PLoS Genet* 2016, 12:1-27.

Complementary Discussion

In this mini-review, we provide a contextualization of each milestone in the human evolutionary history in the Americas, which was just possible with careful research about the most recent methods and techniques developed in this area, and the importance of each one in the construct of the knowledge that we have today. We highlight the papers since 2018 that were crucial to better understand this chapter of human history. But since our paper was published, several discoveries were made about the history of Humans in the Americas. Among them, we highlight the discussion of the possible pre-Columbian contact between Natives in South America and Polynesians.

Known for your great capacity for exploring, the Polynesian people reached regions from New Zealand to Easter Island (the closest to South America). But a recent study of ~800 Polynesian modern samples shows evidence that maybe these people had a single contact with a Native American group most closely related to the indigenous inhabitants of present-day Colombia. Other intriguing evidence of this contact is the sweet potato domesticated in Polynesia for hundreds of years before Europeans arrived but that was also found in abundance in the Andes (Roullier, et al 2013).

Another important aspect to be added to this work is the application of genetic medical techniques, as polygenic risk score, to the admixed American people, as the Brazilian. The accuracy of an inferred polygenic risk score is typically highest in the population from which summary statistics were derived because of factors as Linkage disequilibrium and allele frequency. And how the majority of genome-wide association studies (GWASs) are performed in Europeans, we face a significant bias when we try to apply it in American populations (Martin et al., 2017). Here, we dedicate chapter 4 to the application of a polygenic risk score in a Brazilian population, in which we show preliminary results of the correlation between the loss in the accuracy of this inference and the percentage of European ancestry.

The literature review is a crucial step in all scientific works. It gives an overview of the study subject allowing the contextualization and orientation of the research. The paper presented in this chapter was very helpful to the conclusion of our manuscript (chapter 2) and especially for the decisions regarding the next steps to be taken. Besides that, in this mini-review, we highlight that the most used methods to infer natural selection are based on allelic

frequencies and their spectra, or to detect more subtle signals from genes in common biological networks. This inspired us to start development of a complete pipeline to search for signatures of natural selection in populations with very little information about that, taking into account different methods based on length of haplotypes, population differences, high frequency of derived alleles, besides specific Softwares as GRoSS (Refoyo-Martínez et al., 2019) and ASMC (Palamara et al., 2018). This multi-methodological strategy decreases the chance of catching false positive signals, which is a huge problem of the studies of natural selection that use exclusive genetic data. The development of this pipeline was one of my main projects in my sandwich period in the University of Toronto - Canada supported by a Mitacs Globalink Research Award (FR37903) and the Brazilian government program CAPES-print (88887.474324/2020-00), the preliminary results of this project are presented in Chapter 3 of this thesis.

Chapter 2 - The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture

Introduction

The South American region has great historical importance, as the last continental region in the world to be populated by humans and cradle of great civilizations with abundant archaeological records. Despite that, this region has few studies that seek to understand the current genetic composition of native populations, which are still surrounded by uncertainties. Among these, there are mainly questions about the migratory population dynamics that gave rise to the first South Americans more than 14,000 years ago, and about how the dispersions after these events fostered the rich genetic and cultural diversity of the continent (Mendes et al, 2020).

Two important examples of South American regions that have gone through different cultural, linguistic and ecological processes are the Amazon and Andean areas. The native populations living in the east of the Andean mountains were characterized as groups with high rates of interpopulation differentiation and low levels of gene flow. However, there is linguistic and archaeological evidence suggesting that the Amazonian groups were not that isolated, but maintained a certain level of cultural and economic interaction both among themselves and with the Andean region. Another fundamental difference between these groups (Amazon and Andes) is that while the Andean populations experienced processes of establishment of empires and civilizations that allowed connections within this region that resulted in cultural expansions, the Amazon populations went through processes of greater genetic isolation from each other, being the populations of this region smaller, in terms of effective population size, and more isolated. The divergence between the Andes and the Amazon was inferred to occur approximately 12,000 years ago. Since then, natural selection due to the Andean and Amazonian environment may have shaped the genetic composition of both groups, enabling, among other factors, the adaptation to such different environments and the different patterns of genetic composition found in these groups.

As a part of the efforts of the LDGH group to fill the gap in genomic studies in neglected populations, the main project developed by me during my Ph.D. aimed to analyze the genetic

diversity of natives in South America with three focuses: 1) population structure and ancestry components of the West of South America. 2) to infer and analyze segments identical-by-descent (IBD), and apply this approach to historical questions in the native population, 3) Identify variants that have suffered selective pressure since the settlement of South America, specifically in relation to the occupation of two environments, the Andes and the Amazon, and interpreting those results in a biomedical context.

Methodology

As one of the first authors of the paper “The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture”, I actively participated in all sections, reading and critically discussing all topics, besides directly performing some of the analyzes. So I will include below an adaptation of our supplementary material including only the description of our sampling/datasets and the analysis carried out directly by me, but you can found the entire Supplementary material at the link: <https://www.pnas.org/doi/10.1073/pnas.2013773117#supplementary-materials>

Section 1: Sampling, quality control and Datasets

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 (Φ_{ij}) 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 $\Phi_{ij} \geq 0.25$ to define relatedness or not. A pair of individuals with Φ_{ij} 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:

- 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).
- 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.
- 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?

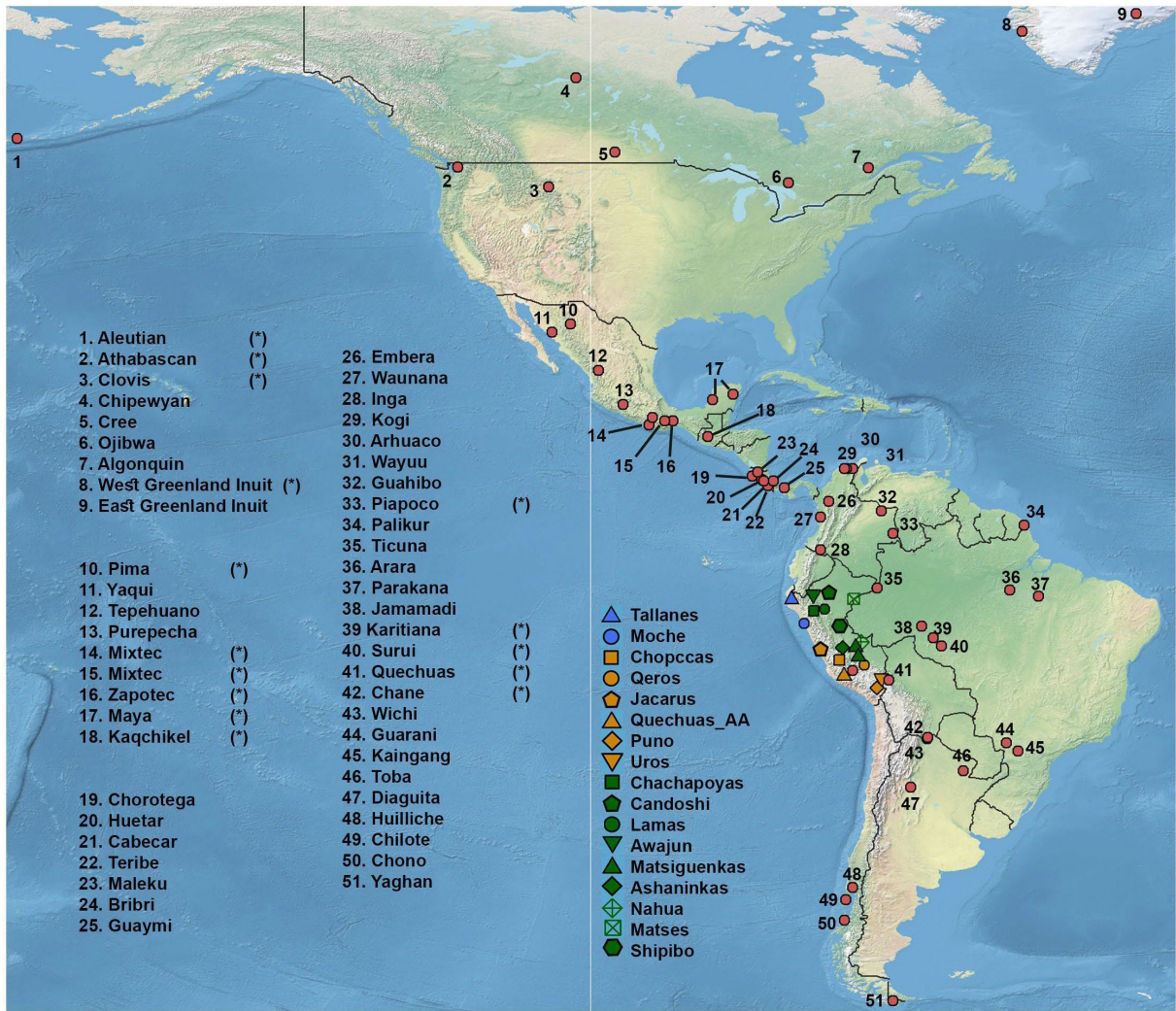


Figure S1. Geographical distribution for the 18 Peruvian Native populations sampled, plus the 65 sampled Native American populations and public data sets (Mallick *et al.* 2016, Raghavan *et al.* 2015, Reich *et al.* 2012). All samples except Clovis and Athabascan were included in a data set of ~ 230,000SNPs.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

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?

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 (Matsigenkas). As in ADMIXTURE, both Matsigenkas groups, Shima (Matsigenkas 1) and Sepahua (Matsigenkas 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 SPPC 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 (Jaquarus, Quechuas_AA and Chopccas) showed ~10% of SPPC related ancestry. Matsigenkas individuals were observed as a highly differentiated population since it has aspecific 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 Matsigenkas 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.

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.

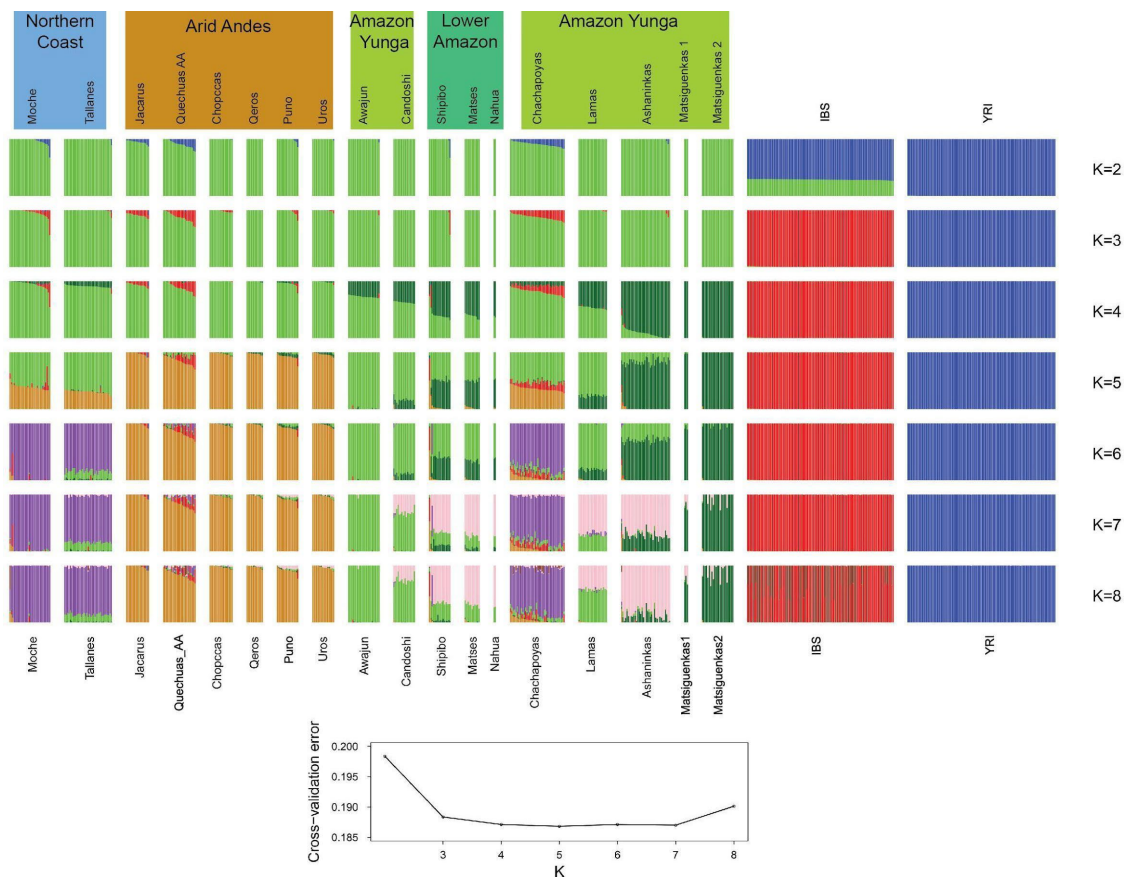


Figure S2. ADMIXTURE analysis for 18 Native American populations, as well as Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 1.9M Dataset). Figure shows results

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 validation error corresponds to K=5 in which we observed four Native American, one European and one African cluster.

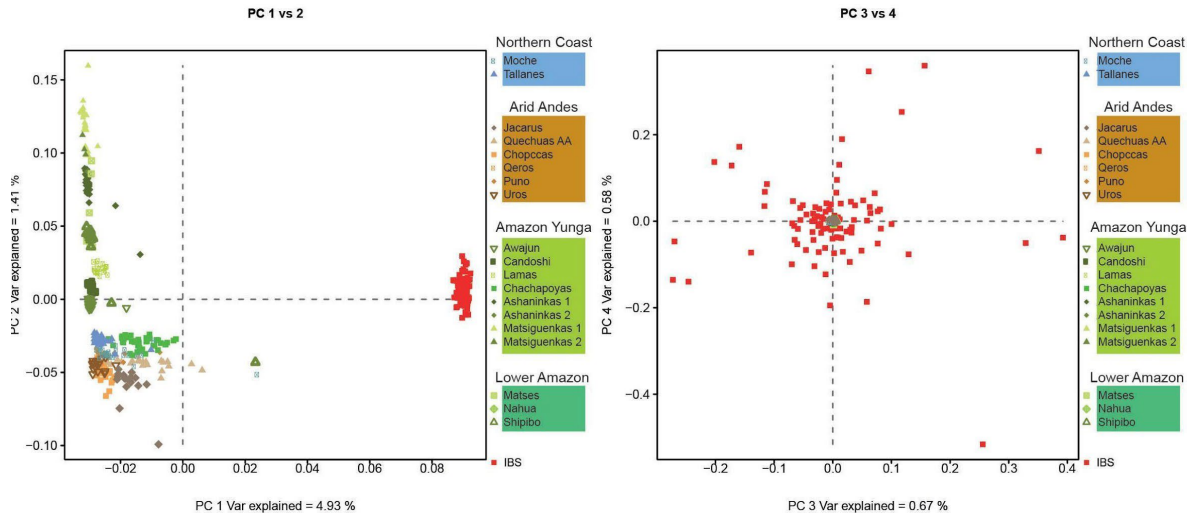


Figure S3. Principal Component Analysis for 18 Native American Peruvian populations and Iberian individuals (IBS) from 1000 Genomes Project (Natives 1.9M Dataset). Shades of blue are related to Coast populations. Orange-brown colors are related to Andean populations and green colors are related to Amazon.

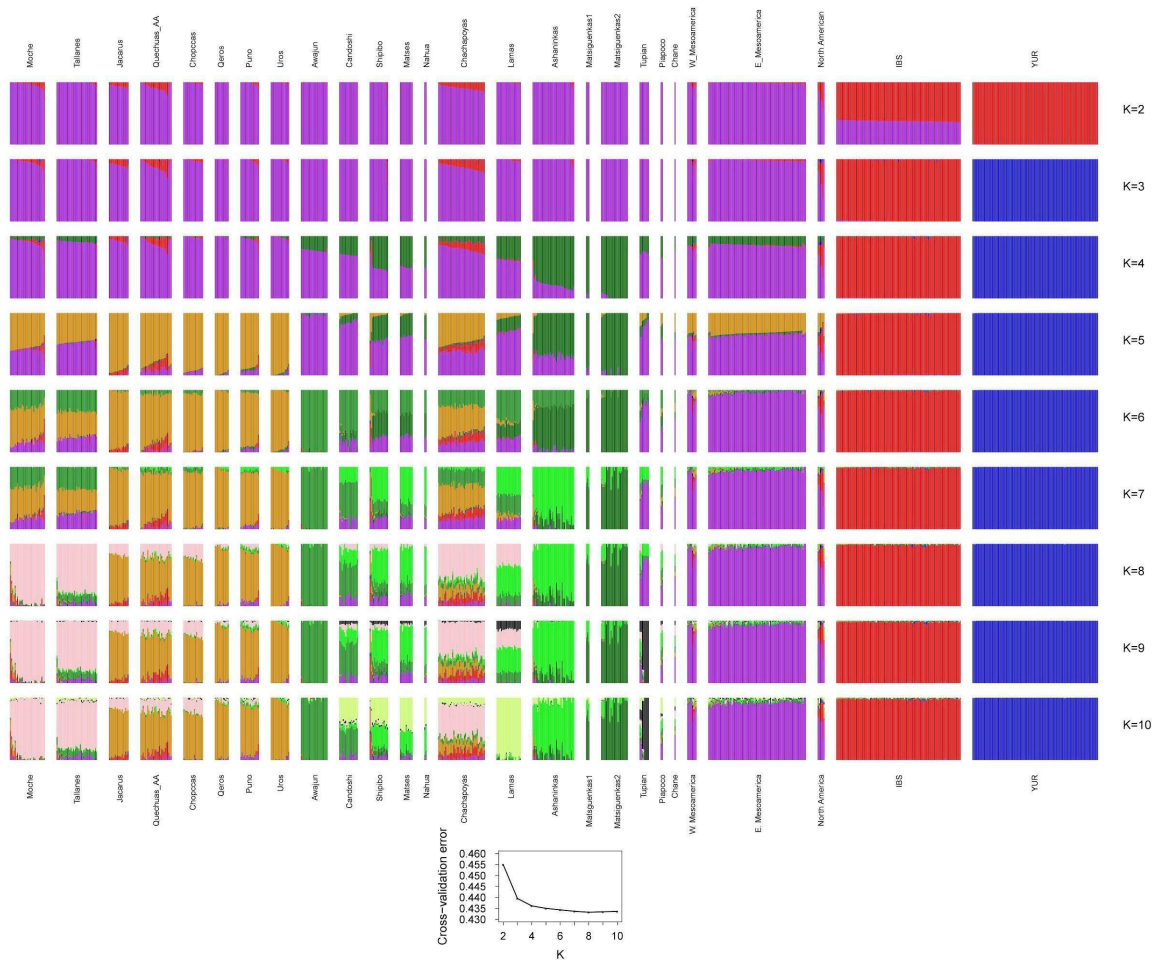


Figure S4. ADMIXTURE analysis for 18 Native American Peruvian populations, Guatemala samples, Native Americans from Raghavan *et al.* 2015 and the Simons Project (Mallick *et al.* 2016) Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 500K Dataset). Figure shows results for 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.

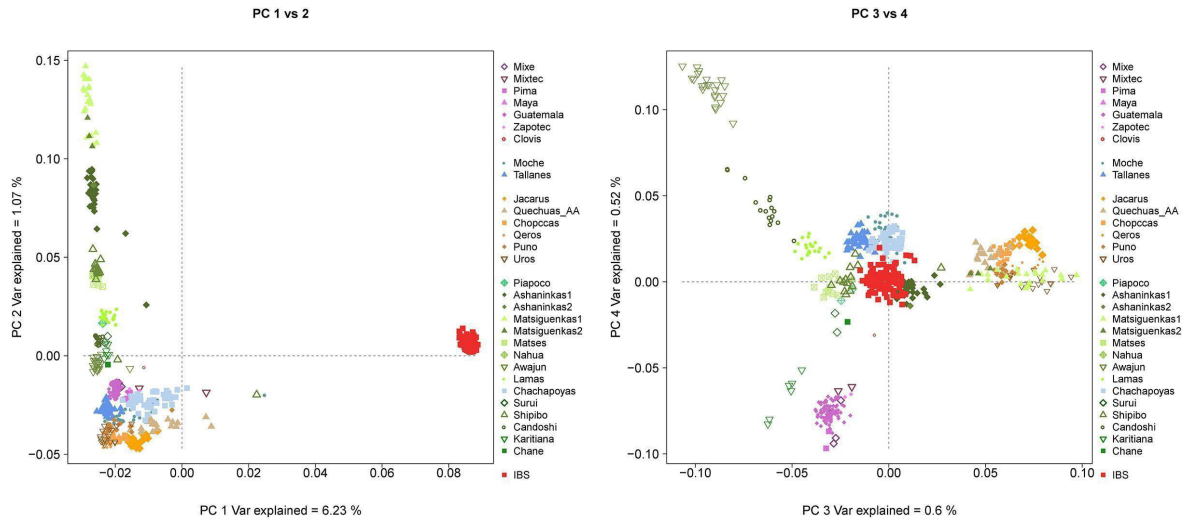


Figure S5. Principal Component Analysis for 18 Native American Peruvian populations, Guatemala samples, Native Americans from Raghavan *et al.* 2015 and the Simons Project (Mallick *et al.* 2016) and Iberian (IBS) populations from 1000 Genomes Project (Natives 500K Dataset). Shades of blue are related 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.

Section 4: Dating the between-population homogenization of the aridAndes

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):

$$E \approx 3/2L$$

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 presentday) |

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:

$$abIBD = \frac{\sum_{i,j} l_{ij}}{N_{pairs}}$$

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: $N_a * N_b$ (for different populations), and $N_a(N_a-1)/2$ (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 (N_e) with software **IBDne** (41). This algorithm infers the pattern of N_e 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 N_e , 2) it assumes a closed population, 3) it assumes a homogeneous 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 (N_e) 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 N_e 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 (N_e), 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.

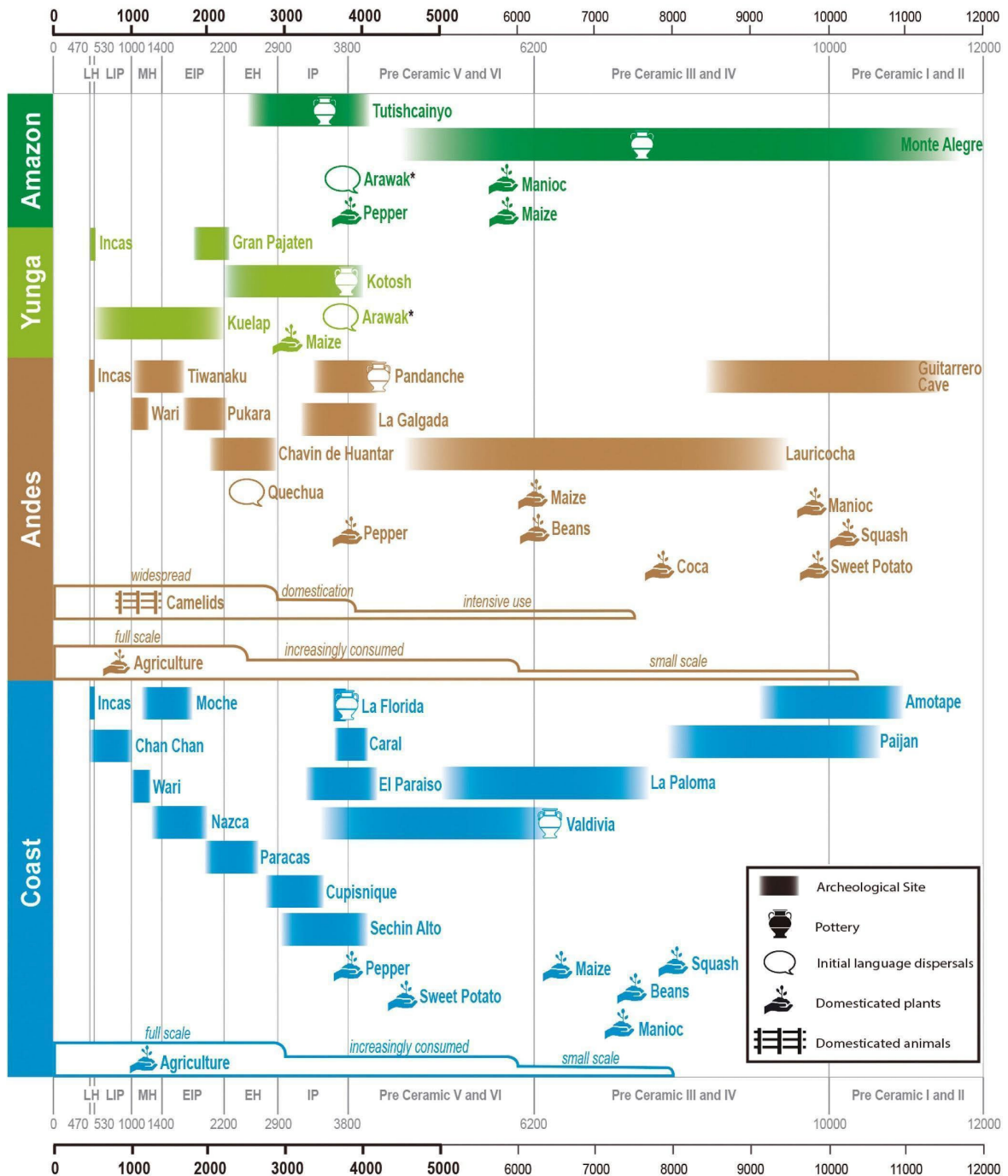


Figure S20. Key historical events of Peruvian prehistory in four longitudinal regions: Peruvian Coast, Andes, Amazon Yunga and Amazonia. Pottery and cultivars symbols represent the earliest archaeological record for the region. To account for time uncertainties, This figure showed the events in the chronology 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](https://creativecommons.org/licenses/by/4.0/).

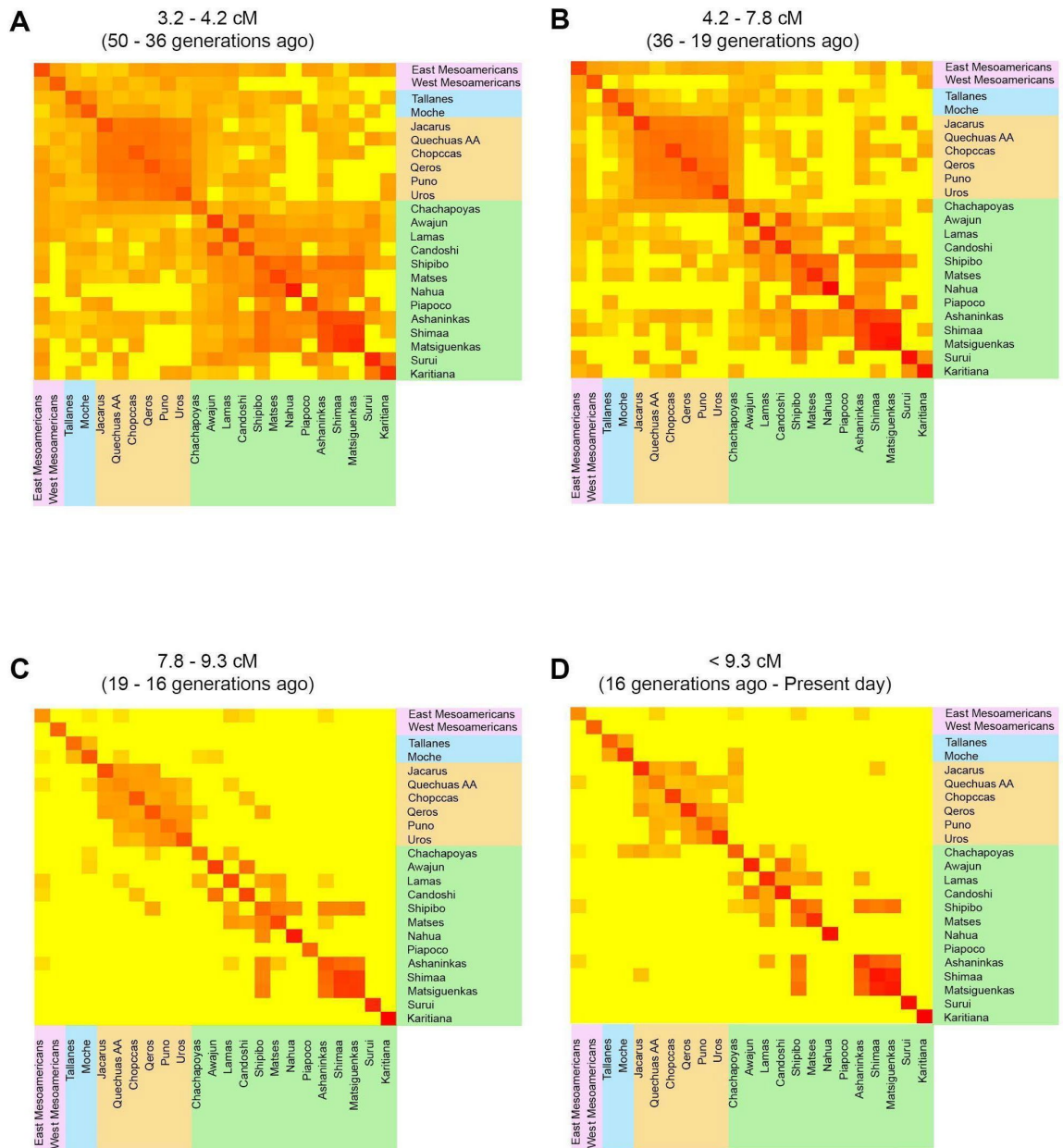


Figure S21. Heatmap representation of the shared Identical by descent (IBD) segments among Native Americans of the Natives 1.9M dataset. Each heatmap represents an interval of segments size and is correlated with time generation for the most common recent ancestor. A) An interval from 3.2 to 4.2 cM correlated 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.

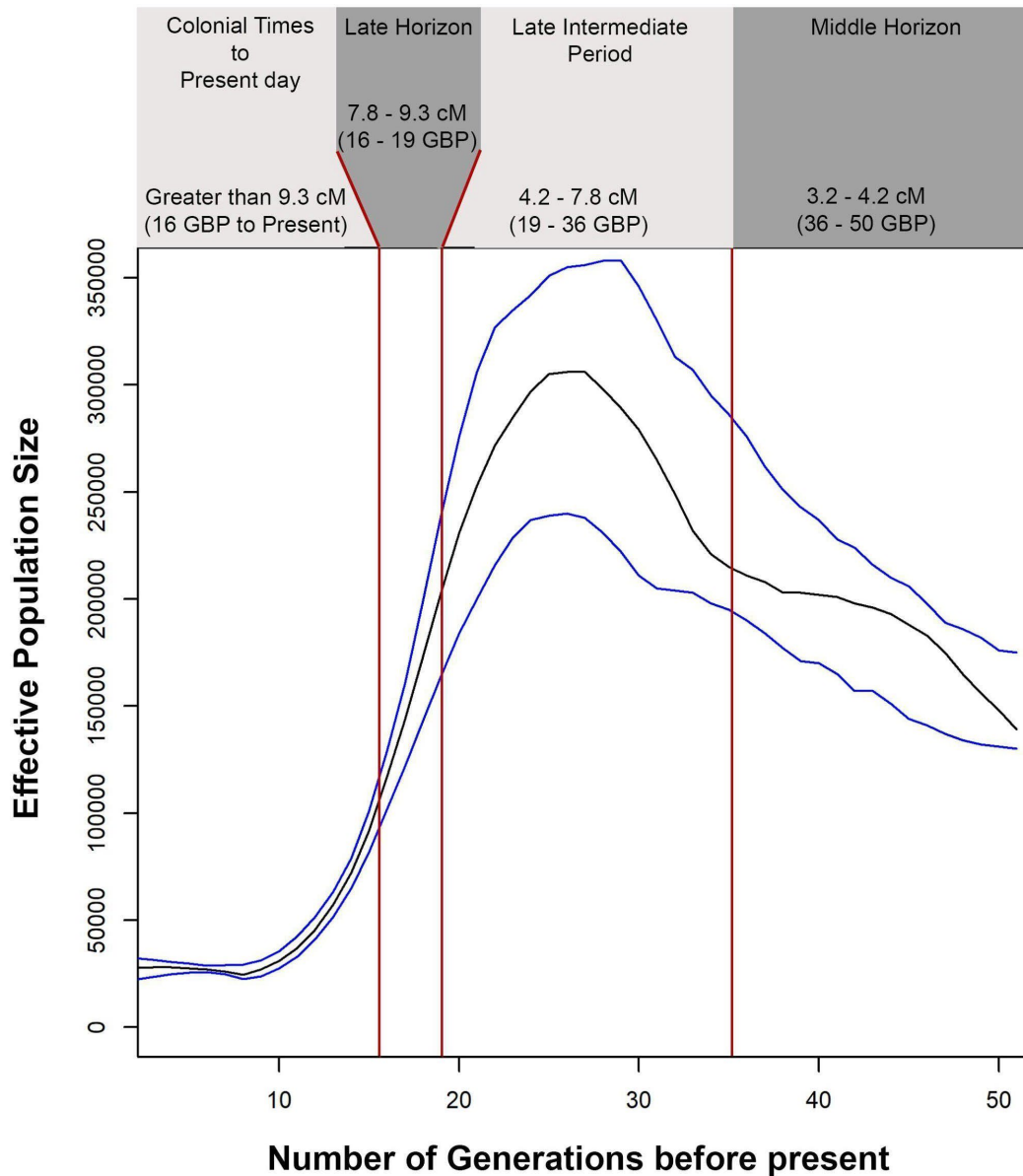


Figure S22. IBDNe analysis to infer the dynamic of the effective population size (N_e) from 4 generations ago to the last 50 generations for the Andean populations (Quechuas_AA, Aymaras_P, Chopccas, Qeros and Uros) as a whole. We used the Natives 1.9M dataset. The x axis represents the number of generations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

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, Matsigenkas (including Matsigenkas 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):

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

Where:

PBS_n : 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 0.000000045# 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 Chopecas (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).

5.3. 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: *MCHRI*: $PBSn=0.26$ 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).

Demographic Model

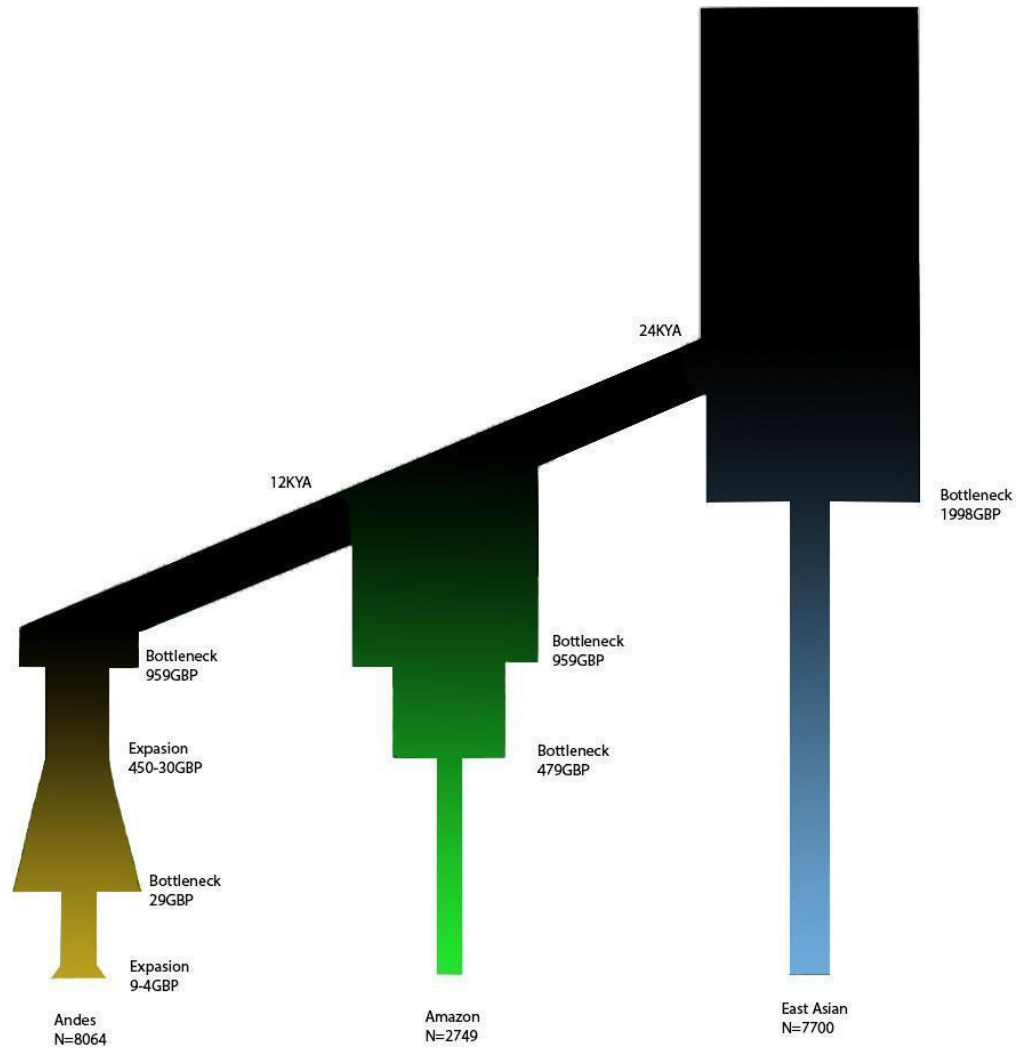


Figure S23. Demographic model of the Andean, Amazonian and East Asian populations. This model was used for the simulations made to calculate the p-value of the obtained PBSn values.

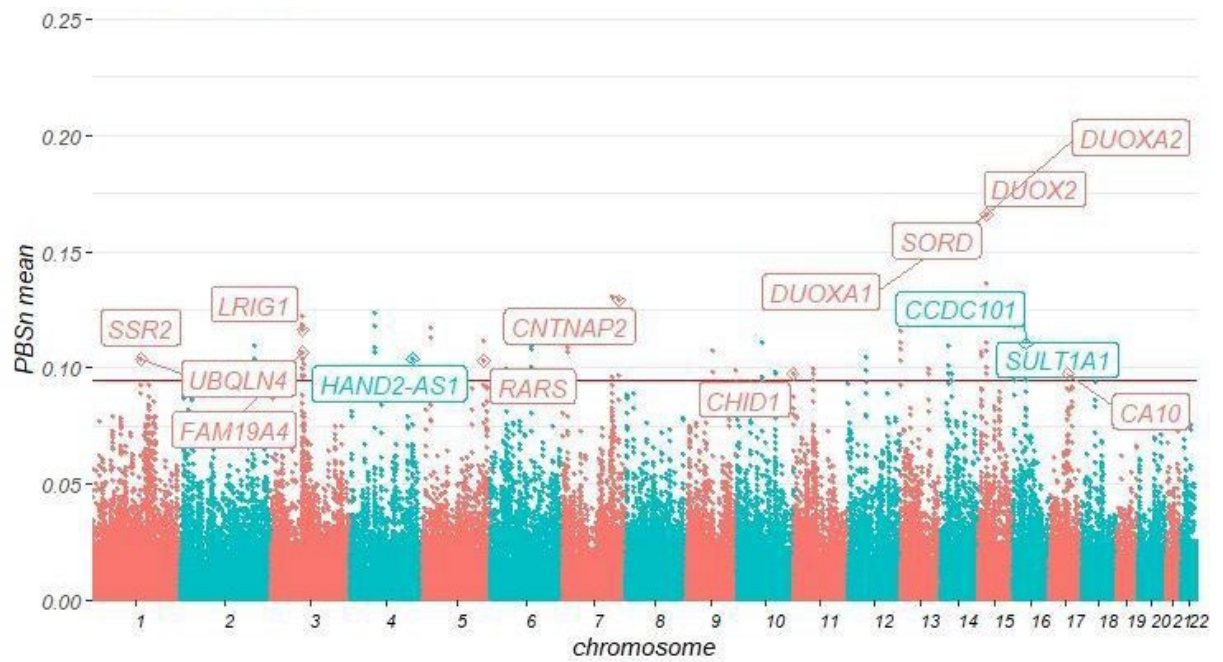


Figure S24. PBSn mean values Andean populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of PBSn mean (red line).

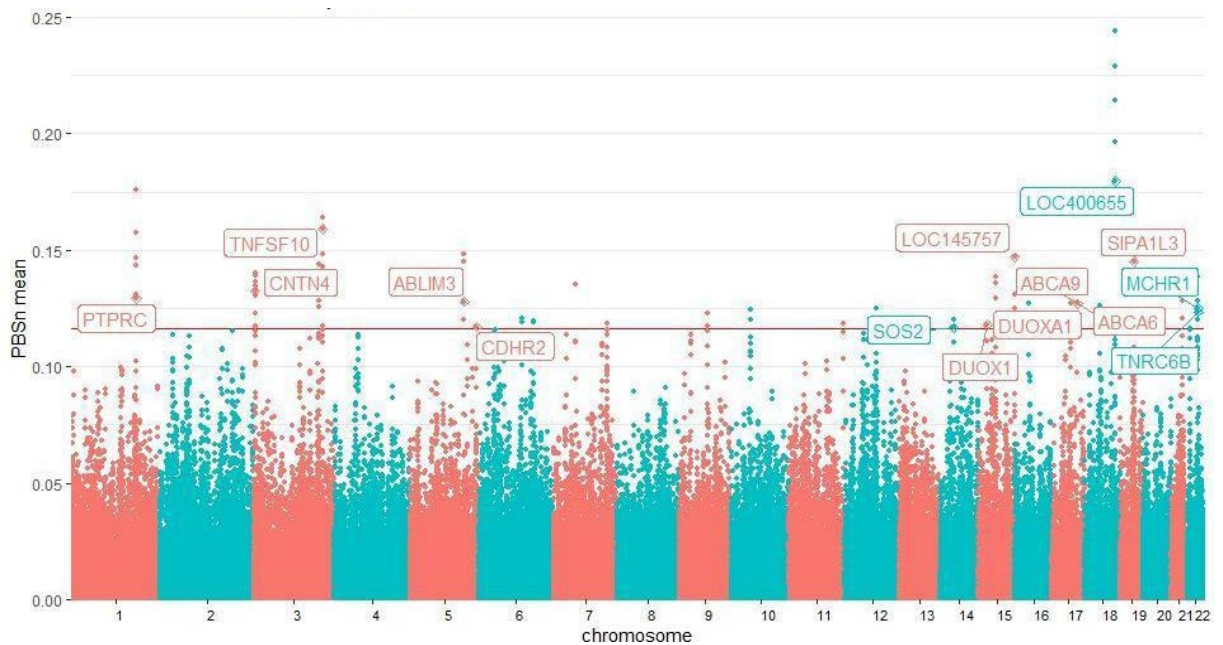


Figure S25. PBSn mean values Amazon populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of PBSn mean (red line).

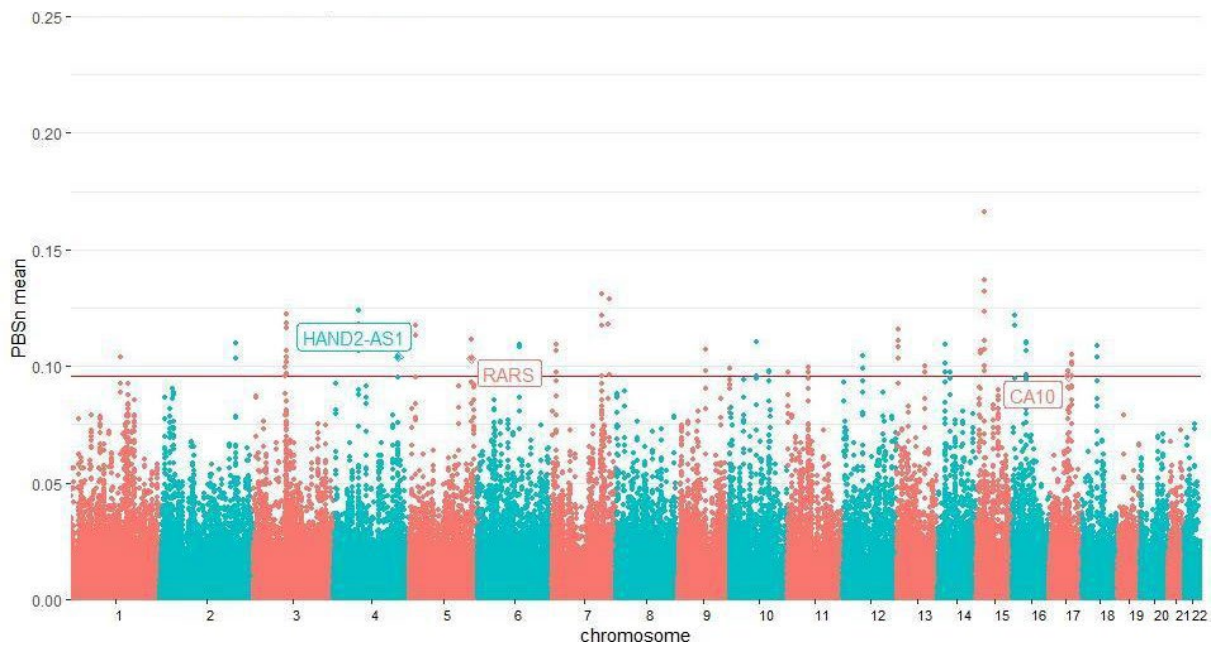


Figure S26. PBSn mean values for windows of 20 SNPs with five SNPs of overlap in Andean populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of windows PBSn mean (red line) that also present high values for xpEHH are labeled.

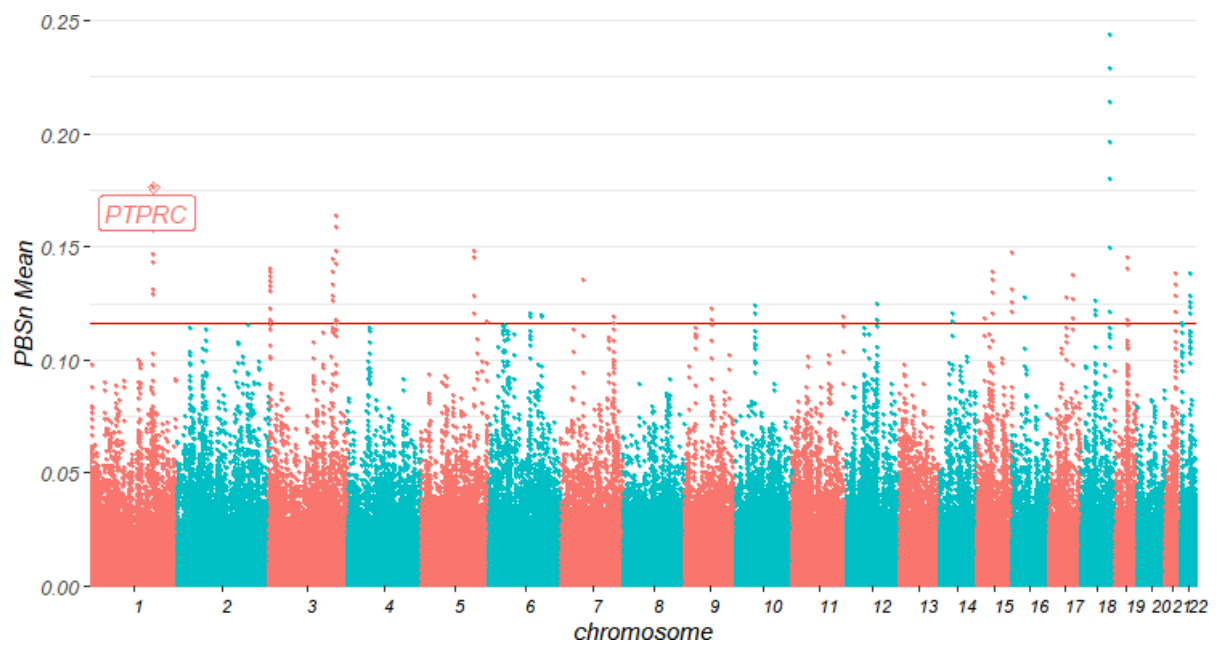


Figure S27. PBSn mean values for windows of 20 SNPs with 5 SNPs of overlap in Amazon populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of windows PBSn mean (red line) that also present high values for xpEHH are labeled.

SI Datasets

All supplementary Datasets are included in the webpage of our published paper: <https://www.pnas.org/content/117/51/32557/tab-figures-data>

SI Datasets Legends

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 S2: List of all samples included in the Native 500K dataset.

Dataset S3: List of all samples included in the Native 230K dataset.

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

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

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

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

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 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 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

SI References

1. W. B. Church, A. von Hagen, “Chachapoyas: Cultural Development at an AndeanCloud 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. Gneccchi-Ruscione, *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. Technol. 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).

Published article



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, Marília 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 Trujilloⁱ, Pedro O. Flores-Villanueva^c, Michael Deanⁱ, Sílvia Fuselli^k, Moara Machado^{a,j}, Pedro E. Romero^l, Francesca Tassi^k, Meredith Yeager^l, Timothy D. O'Connor^{m,n,o}, Robert H. Gilman^{l,p}, Eduardo Tarazona-Santos^{a,1,q,2,3}, and Heinner Guío^{c,h,r,2,3}

^aDepartamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil; ^bLaboratório de Bioinformática (LABINFO), Laboratório Nacional de Computação Científica, Petrópolis, RJ, 25651-076, Brazil; ^cLaboratorio de Biotecnología y Biología Molecular, Instituto Nacional de Salud, Lima 9, Peru; ^dHuman Genome and Stem Cell Research Center, Biosciences Institute, University of São Paulo, São Paulo, SP, 05508-090, Brazil; ^eBeagle, Belo Horizonte, MG, 31270-901, Brazil; ^fMosaico Translational Genomics Initiative, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil; ^gLaboratório de Genética Humana e Médica, Programa de Pós Graduação em Genética e Biologia Molecular, Universidade Federal do Pará, Belém, PA, 66075-110, Brazil; ^hCarrera de Medicina Humana, Facultad de Ciencias de la Salud, Universidad Científica del Sur, Lima, 150142, Peru; ⁱCentro Nacional de Salud Intercultural, Instituto Nacional de Salud, Lima 11, Peru; ^jDivision of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD 20892; ^kDepartment of Life Sciences and Biotechnology, University of Ferrara, Ferrara, 44121, Italy; ^lUniversidad Peruana Cayetano Heredia, Lima 31, Peru; ^mInstitute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201; ⁿProgram for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD 21201; ^oDepartment of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201; ^pDepartment of International Health, Johns Hopkins School of Public Health, Baltimore, MD 21205; ^qInstituto de Estudos Avançados Transdisciplinares, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil; and ^rFacultad 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-ASI* (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 inviral–host interaction.

Native Americans | human population genetics | natural selection | geneflow

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

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.

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.

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

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

³To 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.

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 Quechumaram 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

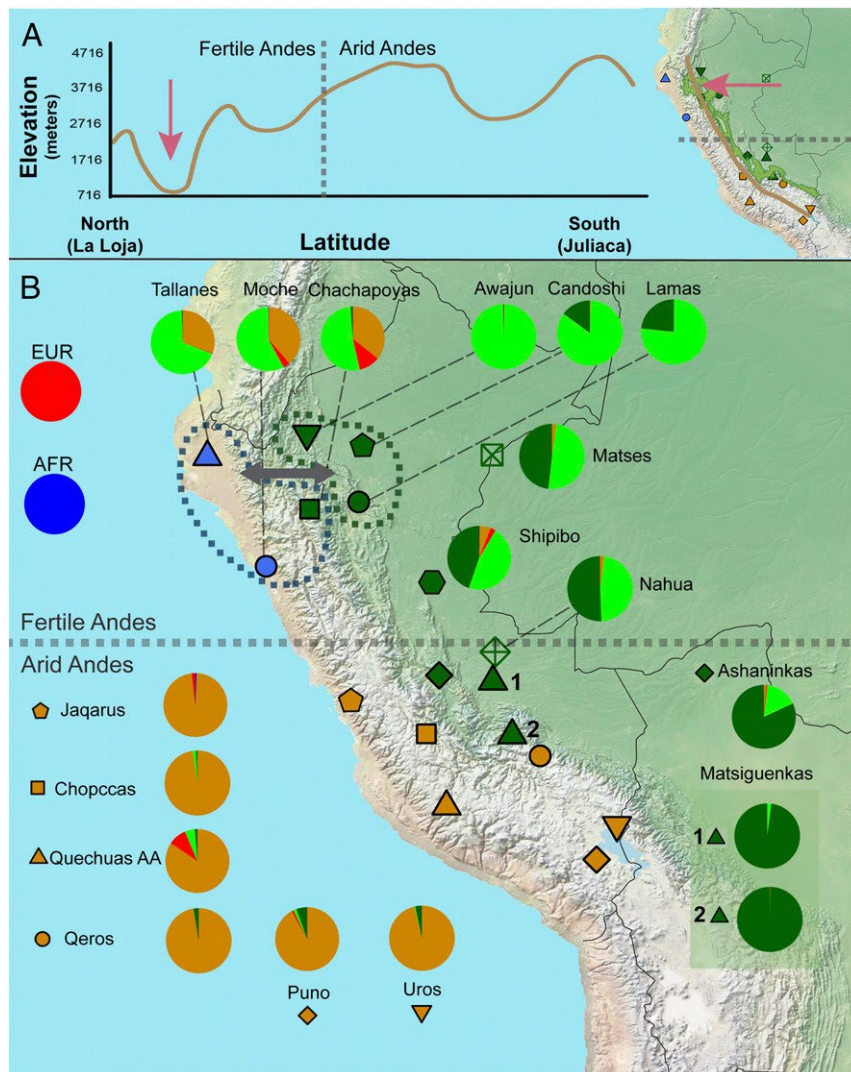


Fig. 1. Genetic and geographic landscape for western South American natives. (A) Elevation vs. latitude plot from a cross-section line in the Andean region from La Loja (Ecuador) to Juliaca (Peru). A vertical line indicates the division between Fertile and Arid Andes (16). The map shows the path used 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*). Matsigenkas 1 = Matsigenkas- Sepahua, Matsigenkas 2 = Matsigenkas-Shimaa. The gray horizontal dashed line in the center of the map shows the approximate division between the fertile Andes and arid Andes (16).

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. 1A). 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. 1B 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. 1B 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. 1B).

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. 1B 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/Globe-trotter methods) (33, 34) (Fig. 1B and SI Appendix, Figs. S11–S13 and section 2.1.3), statistical tests of treeness (35) (Fig. 1B 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

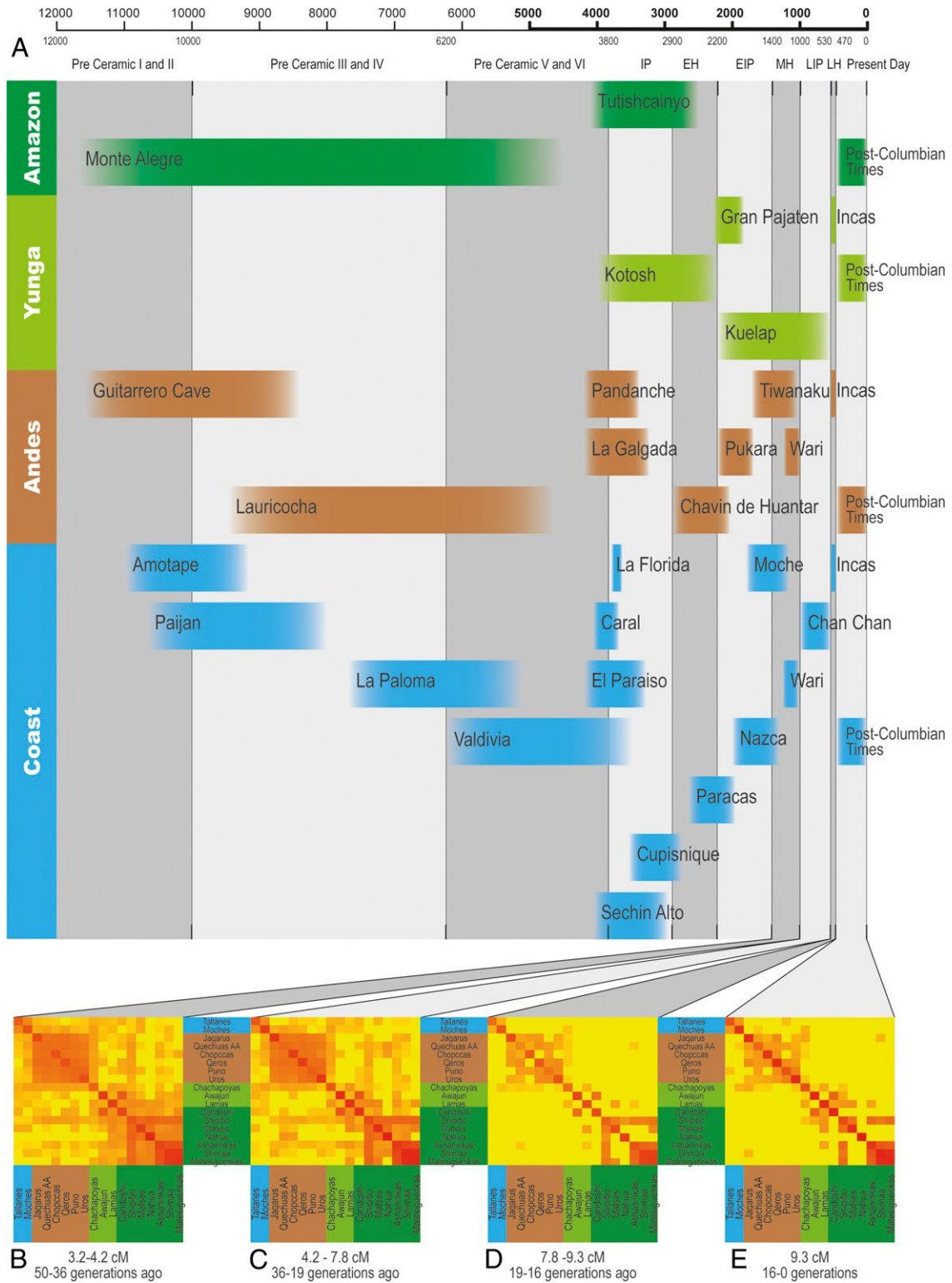


Fig. 2. Changes in IBD sharing over time between the Pacific coast, central Andes, Amazon Yunga, and Amazon and its relationship with the archaeological chronology of the Andes. (A) Key historical events (cultures and archaeological sites) of Peruvian history in four Peruvian longitudinal regions, coast, 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](https://creativecommons.org/licenses/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).

differentiation between the fertile north, where altitudes are lower, and the arid south, where the Andes altitudes are higher (Fig. 1A) 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. 1B 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. 1B) 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 (Chopceas, Quechuas AA, Qeros, Puno, Jaqarus, and Uros; $n = 102$) vs. Amazonian populations (Ashaninkas, Matsigenkas, 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-ASI* (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-ASI* 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-ASI* 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 *DUOX2A2*. 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

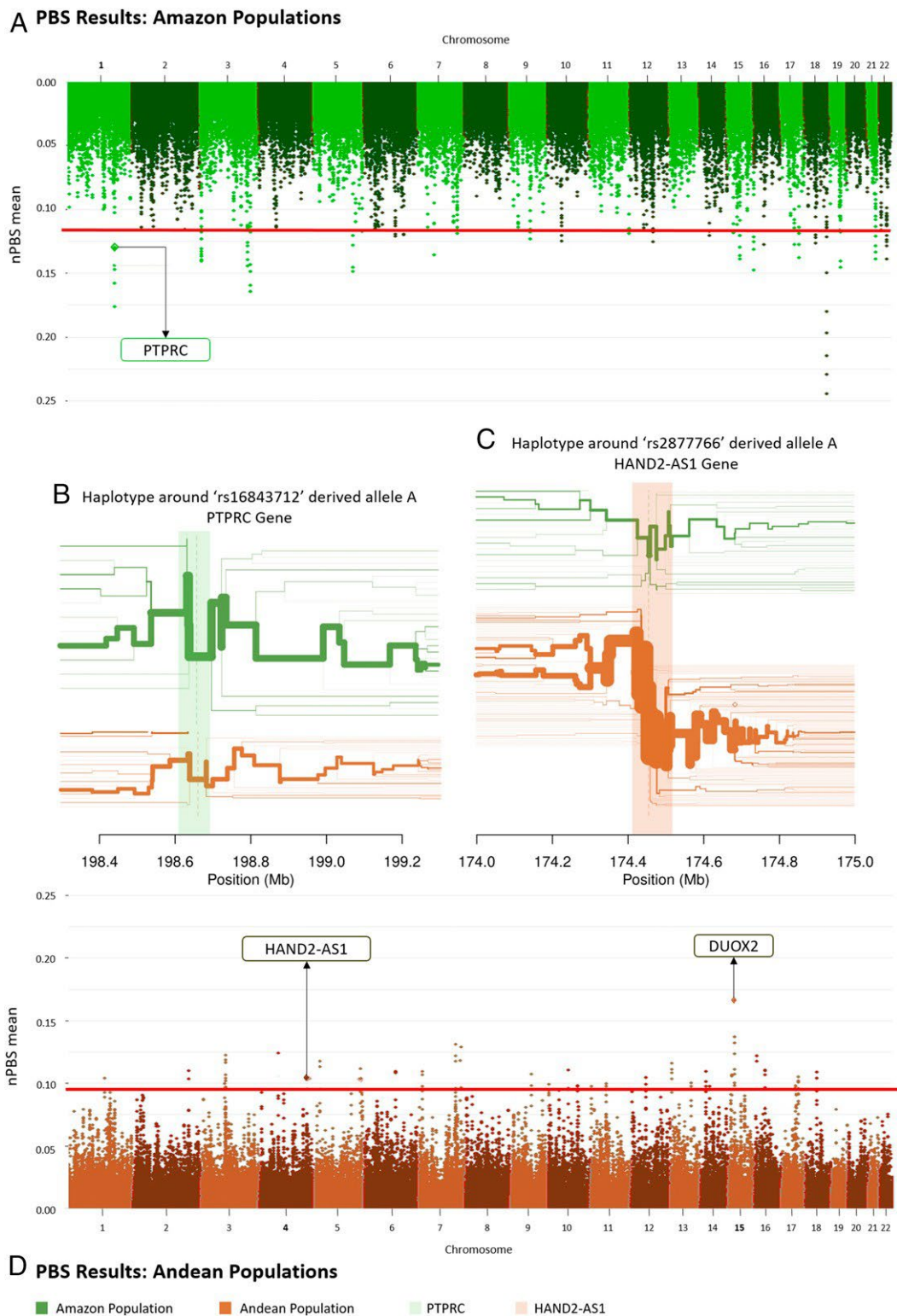


Fig. 3. Sandwich-like representation of natural selection signatures in the arid Andes and the Amazon tropical forest. Manhattan plots (the bread of the sandwich) correspond to the PBSn estimated from the sliding window in Amazon (A) and Andean (D) populations. The horizontal red line shows the 99.95th 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

(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 Diversidade Genética Humana, Mateus Gouveia, Kelly Nunes, Garrett Henthall, 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 Biotecnología y Biología 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 (PRINTE). 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

processed in the Sagarana HPC cluster at the Centro de Laboratórios Multi-usuários 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 populacional (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. Gnecci-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. Lumbereras, *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 through 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. Corvillain, 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).

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).

Chapter 3 - Identifying signatures of Natural Selection in Indian populations

Identifying signatures of natural selection in Indian populations

Marla Mendes^{1,2}, Manjari Jonnalagadda³, Shantanu Ozarkar⁴, Victor Borda Pua⁵, Christopher Kendall², Eduardo Tarazona-Santos¹, Esteban J. Parra²

Affiliations

1. Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil.
2. Department of Anthropology, University of Toronto - Mississauga Campus, 3359 Mississauga Rd, Mississauga, ON L5L 1C6, Canada.
3. Symbiosis School for Liberal Arts (SSLA), Symbiosis International University (SIU), Pune 411014, India.
4. Department of Anthropology, Savitribai Phule Pune University, Pune - 411007.
5. Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD.

Corresponding Author

Marla Mendes marlamendesdeaquino@gmail.com

Esteban J. Parra esteban.parra@utoronto.ca

Keywords

Positive selection, India, microarray data

Abstract

In this study, we present the results of a genome-wide scan for signatures of positive selection using data from four tribal groups (Kokana, Warli, Bhil, and Pawara) and two caste groups (Deshastha Brahmin and Kunbi Maratha) from West Maharashtra, as well as two samples of South Asian ancestry from the 1KG project (Gujarati Indian from Houston, Texas and Indian Telugu from UK). We used tests of positive selection based on different statistics, including PBS, xpEHH, iHS, CLR, Tajima's D, as well as two recently developed methods:

GROSS and ASMC. In order to minimize the risk of false positives, we selected regions that are outliers in all the samples included in the study using more than one method. We identified putative selection signals in 107 regions encompassing 434 genes. Many of the regions overlap with only one gene. The signals observed using microarray-based data are very consistent with our analyses using high-coverage sequencing data, as well as those identified with a novel coalescence-based method (ASMC). Importantly, at least 24 of these genomic regions have been identified in previous selection scans in South Asian populations or in other population groups. Our study highlights genomic regions that may have played a role in the adaptation of anatomically modern humans to novel environmental conditions after the out of Africa migration.

Introduction

South Asia was one of the first geographic areas colonized during the out-of-Africa migration of anatomically modern humans, and not surprisingly, the South Asian region is characterized by having one of the highest levels of genetic diversity outside of Africa (Xing, et al. 2010, Mallick et al., 2016, Paganı et al., 2016, Peter et al., 2019 and Bergström et al., 2020, Jain et al., 2020). This diversity has been shaped by different evolutionary and demographic factors including bottlenecks and genetic drift, multiple migration waves, endogamy and natural selection (Metspalu et al., 2011; Moorjani et al., 2013; Nakatsuka et al., 2017; Liu et al., 2017; Metspalu et al., 2018; Debortoli et al., 2020). Recent studies have highlighted three important migration events that contributed to the formation of present-day South Asian populations. Briefly, these three major events correspond to the I) original out of Africa migration that eventually gave rise to the *Ancient Ancestral South Indian* population (AASI), II) the migration of Neolithic farmers, primarily from the Iranian plateau, and III) the Bronze Age migration of the Yamnaya Steppe Pastoralists (Mellars et al., 2013; Lazaridis et al., 2016; Silva et al., 2017; Narasimhan et al., 2019). It has been proposed that an initial admixture process between AASI hunter-gatherers and Iranian-related farmers gave rise to a population that has been named the *Indus Periphery* group. Further admixture events of the *Indus Periphery* populations with AASI southeastern groups and northwestern groups with Steppe ancestry gave rise to the *Ancestral South Indian* (ASI) and *Ancestral North Indian* (ANI) populations, respectively, a process that probably occurred in the second millennium BCE (Narasimhan et al., 2019). Most of the modern human populations in South Asia show varying proportions of ASI and ANI ancestry (Reich et al., 2009; Moorjani et al., 2013; Metspalu et al., 2018; Narasimhan et al., 2019). It is important to note that there have been

many other documented demographic events in South Asia, including invasions from the Greeks, Kushans, Huns, Muslims, Moghuls, and the English (Kivisild et al., 2003).

Modern Indian populations are typically divided into tribal and non-tribal groups. Tribal groups are considered the Indigenous populations in this region, while non-tribal groups comprise social hierarchical endogamous castes and religious groups outside the caste system (Reddy et al., 2009). Tribes have had a smaller effective population size compared to castes and consequently have experienced more intensively the effects of genetic drift (Indian Genome Variation Consortium, 2008; Debortoli et al., 2020). Genetic studies have indicated that the shift to endogamous marriages which is characteristic of the caste system in India was relatively recent, perhaps around 2,000-1,500 years ago, as there is evidence of substantial admixture in this region prior to this time (Chaubey et al., 2006; Reddy et al., 2009; Reich et al., 2009; Moorjani et al., 2013; Basu et al., 2016; Narasimhan et al., 2019).

In contrast with the recent advances in our understanding of the demographic history of the South Asian continent as a result of genetic studies in modern and ancient samples (Reich et al., 2009; Juyal et al., 2014; Haak et al., 2015; Basu et al., 2016; Silva et al., 2017; Mccoll et al., 2018; Shinde et al., 2019; Narasimhan et al., 2019; Debortoli et al., 2020), there have been limited attempts to explore the potential role of positive selection on South Asian populations. As humans migrated out-of-Africa, they adapted to novel environments and it is of interest to identify the genomic regions that were targeted during these adaptation processes. There have been many efforts to identify selection signatures in European, East Asian and African populations (Mathieson et al., 2015; Suo et al., 2011; Patin et al., 2017; Perdomo-Sabogal and Nowick, 2019), but just a few studies have focused exclusively on South Asian samples (Metspalu et al. 2011, Karlsson et al., 2013; Juyal et al. 2014; Jonnalagadda et al. 2017).

Here, we present the results of a genome-wide scan for signatures of positive selection using data from four tribal groups (Kokana, Warli, Bhil, and Pawara) and two caste groups (Deshastha Brahmin and Kunbi Maratha) from West Maharashtra, as well as two samples of South Asian ancestry from the 1000 Genome Project (Gujarati Indian from Houston, Texas and Indian Telugu from UK). In order to identify putative genomic regions under positive selection, we used tests of positive selection based on different statistics, including Population Branch Statistic (PBS), Cross-population Extended Haplotype Homozygosity (xpEHH), Integrated Haplotype Score (iHS), Composite Likelihood Ratio (CLR), Tajima's D, as well as two recently developed methods: Graph-aware Retrieval of Selective Sweeps (GRoSS) - that uses admixture graphs to infer signatures of selection in specific branches of

the graphs (Refoyo-Martínez et al., 2018), and Ascertained Sequentially Markovian Coalescent (ASMC) - a coalescence-based method (Palamara et al., 2018).

Materials and Methods

Datasets

In this study, we used two different datasets to enable a deeper understanding of selection signatures in South Asia populations. The first dataset is genome-wide data from six West Maharashtra (WM) populations, belonging to the Indo-European language family. Those populations include four tribal populations (collected from Jawhar at 19.918N, 73.238E and Dhadgaon at 21.828N, 74.228E): Kokana, Warli, Bhil, and Pawara; and two caste groups (collected close to Pune city at 18.538N, 73.878E): Deshastha Brahmins, and Kunbi Marathas (Jonnalagadda, 2015) (Supplementary Figure 1). The sampled individuals are 480 volunteers who provided informed consent and information about their place of origin, clan, age, and gender along with 5–8ml of whole blood, collected in EDTA vials with the approval from the Institutional Ethics Committee (IEC) at the Savitribai Phule Pune University (Ethics/2012/16). The DNA extraction was performed using the phenol-chloroform method (Sambrook et al. 1989) and the quantification with Eppendorf BioPhotometer plus. The genotyping was carried out with Applied Biosystem's Axiom TM Precision Medicine Research Array (PMRA) at Imperial Life Sciences Pvt Ltd. Laboratory (Gurgaon, Haryana, India) using standard protocols. This array includes approximately 900,000 genetic markers.

The second dataset is from the 1KGP Phase 3 data (1KGP samples, Auton, 2015), of which we use the samples with Indian ancestry: GIH (Gujarati Indian from Houston, Texas), and ITU (Indian Telugu from UK); the European CEU sample (Utah Residents (CEPH) with Northern and Western European Ancestry); and the African YRI sample (Yoruba in Ibadan, Nigeria) samples. We used this dataset in two ways: 1) To carry out diverse tests of selection based on the SNPs that overlap with the microarray-based sample from West Maharashtra described above, and 2) To carry out diverse tests of selection based only on the high coverage 1000 genome data (~70,7M autosomal SNPs) to validate the previous results.

Quality Control

The first QC in the WM samples was done with the Axiom Analysis Suite program which retained ~522,125 polymorphic markers and 478 samples (Jonnalagadda et al. 2019).

We did additional QC steps to remove samples based on: 1) sex discrepancies, 2) outliers for heterozygosity, 3) missing call rates <0.95 , 4) related individuals ($\pi\text{-hat} > 0.25$), and 5) samples that were outliers in Principal Component Analysis (PCA) plots. We also remove markers with: 1) genotype call rate <0.95 , 2) Hardy-Weinberg (HW) p-values $<10^{-6}$, 3) minor allele count <4 , 4) Insertion/Deletion (Indel) markers, 5) markers not present in the 1000 Genomes reference panel, or that did not match the chromosome, position, or alleles information, 6) A/T or G/C SNPs, 7) allele frequency differences $> 20\%$ between the study sample and the 1000 Genomes South Asian reference sample, 8) SNPs without chromosome information and 9) duplicated SNPs. After all QC steps, the dataset contained

$\sim 365,152$ variants and 456 samples. Similar QC steps were carried out in the 1KG sample, obtaining a final sample with $\sim 54,2\text{M}$ variants.

Natural Selection Analysis

Our approach to identifying putative regions of positive selection is based on the application of several methods that focus on different aspects of the genomic data in order to identify “outlier” regions based on the empirical distribution of test statistics across the genome. We carried out an initial scan based on the SNPs that overlap between the WM and the 1KG samples ($\sim 283\text{K}$ SNPs) using six different approaches: PBS, xpEHH, iHS, CLR, Tajima’s D and GRoSS (Supplementary Figure 2).

We developed custom python scripts to select the top 0.1%, 0.5% and 1% regions for each method (the greater values for PBS, xpEHH, iHS, CLR and the lowest values for Tajima’s D and GRoSS P-values). We then annotated these genomic regions using the UCSC (University of California Santa Cruz) Table Browser tool (<https://genome.ucsc.edu/>), which searches for the specific genes found in each genomic region.

In order to minimize false-positive results, as an additional step, we filtered the results using two strategies: 1/ We only selected genomic regions that are outliers for any given method in the four samples of India included in our study (WM-Castes, WM-Tribes, WM-All and 1KG-India and 2/ We only selected genomic regions that are outliers for at least two different methods (Figure 1). We also evaluated if the regions identified in this analysis show unusual coalescence patterns in Indian populations using the recently developed ASMC method (Palamara et al., 2018). Finally, we also evaluated the results of the selection tests based only on the high coverage 1KG Indian samples as a strategy to validate the results.

Methods Based on Population Differentiation

PBS

To identify changes in the allele frequencies of a target population since its divergence from an ancestral population we performed a PBS test. This statistic is based on the comparison of the allele frequency differences measured with F_{ST} values among three groups: 1) a target population; 2) a sister population, and 3) an outgroup (Yi et al. 2010). This method can identify signatures of natural selection mainly between 75K and 50K years ago and is sensitive to both selections on standing variation and *de novo* mutation (Rees et al. 2020).

As a QC-specific step for this test, we applied a MAF filter (Minimum Allele Frequency > 0.05) where ~82,291 variants were removed. The F_{ST} values were computed using 4P software (Benazzo et al. 2014) and the PBS formula was applied as follows (Yi et al. 2010):

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

Where the $F_{ST}T$ values correspond to the F_{ST} computed with 4P and transformed according to Cavalli-Sforza (1969):

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

Therefore:

$F_{ST}T1$: transformed F_{ST} between the target population and the sister population. $F_{ST}T2$:

transformed F_{ST} between the target population and the outgroup.

$F_{ST}T3$: transformed F_{ST} between the sister population and the outgroup.

The PBS values were normalized following the formula (Crawford et al. 2017):

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

Where:

PBS_n : normalized PBS.

$PBS1$: estimated PBS when the PBS is calculated for the target population. $PBS2$:

estimated PBS when PBS is focused on the sister population.

$PBS3$: estimated PBS when PBS is focused on the outgroup.

In all PBS analyses we used the CEU as the sister population and YRI as the outgroup, and we performed the test for each of the following target populations: 1) India WM, 2) Caste WM, 3) Tribe WM, and 4) India 1KGP populations together (ITU and GIH). We identified

putatively selected regions showing extreme PBS values using *in-house* scripts. For these analyses, we used bins of 20 SNPs with 5 SNPs of overlap.

GRoSS

To incorporate the genetic history of the Indian populations in the inference of natural selection events, we applied the Graph-aware Retrieval of Selective Sweeps (GRoSS) software (Refoyo-Martínez et al. 2018). This method uses complex admixture graphs to infer signatures of natural selection along the branch of the graph, using a modified version of the Q_b statistic (Racimo et al. 2018), known as S_b that requires data frequency information and in this case, the population history graph topology.

To infer the admixture graph we use qpGraph (Patterson et al. 2012) with two configurations (Supplementary Figure 3): one with the final leaf being the Indian population (Figure 3-A) and the other with the final leaves being Tribe and Caste groups (Figure 3-B). Then, we run GRoSS with the output “.dot” from qpGraph and the data formatted in “.gross” from 1) India (WM), 2) Caste (WM) and Tribe (WM) and 3) India 1kqp.

To analyze the GRoSS results, we focus on the branch between Europe2 and the target population, India (WM), India 1kqp (Figure 3-A), Caste (WM), Tribe (WM) (Figure 3-B).

Regions showing a strong deviation of neutrality will have low P-values (Refoyo-Martínez et al. 2018). In our analytical pipeline, we annotated the top genomic regions identified with GRoSS by creating windows spanning 100Kb before and after the SNP with the lowest P-value for those regions.

Methods Based on Linked Variation

This group of methods focuses on more recent *de novo* mutations and is particularly powered to identify selective events that happened approximately less than 30,000 years ago (Sabeti et al. 2006; Rees et al. 2020). To cover this time span, we apply xpEHH (Sabeti et al, 2002) and iHS (Voight et al. 2006). The principle of those methods is based on the fact that a positive selection event increases the frequency of a variant and of the variants close to it, faster than the recombination or mutation process breaks those haplotypes, generating a high-frequency long-range haplotype (Sabeti et al, 2002).

The xpEHH method incorporates the calculation of the EHH for all SNPs in 1MB of distance forwards and backwards for two target populations, in this case, India and CEU (Sabeti et al,

2002), while iHS tracks the decay of homozygosity in the target haplotype concerning to the ancestral and derived haplotypes extending from a specific site (Szpiech and Hernandez, 2014).

For both analyses we apply the software Selscan (Szpiech and Hernandez, 2014) with default parameters, in our data, phased with Sanger Imputation Service, using EAGLE2 (Loh et al. 2016) and Shapeit4 (Delaneau et al., 2019) using the GRCh37 genetic map and the MCMC parameters: `-mcmc-iterations 10b,1p,1b,1p,1b,1p,1b,1p,10m`, which perform 10 burn-in iterations, followed by four paired runs of pruning and burn-in, and, finally, 10 main iterations of sampling. Both results were normalized with the extension “norm” from Selscan, by 20 equally sized allele frequency bins. In the iHS inference, we used polarized data. After identifying the SNPs with the highest values for iHS and xpEHH, we annotated the genomic regions by creating windows spanning 100 Kb before and after the selected SNPs.

Methods Based on Site Frequency Spectrum

These methods can detect older natural selection events, ~80,000 years ago, and although Tajima’sD is used mostly for sequence data, we apply it in this study combined with other methods, to identify regions whose frequency spectra are strongly different from the bulk of the genome, suggesting the influence of selection (Sabeti et al. 2006, Voight et al. 2006).

This test compares the average number of pairwise differences and the average number of total segregating sites (Tajima, 1989). Strong negative Tajima’sD values suggest an excess of rare alleles, which may be indicative of positive selection or population expansion (Vitti et al. 2013). To estimate Tajima’s D, we used vcfTools with the “TajimaD” flag, for 100 kb windows (Bigham et al. 2009), and identified the regions with the most negative values as regions under putative selective pressure.

Composite Methods

To reduce the ratio of false positives, this approach combines test scores from diverse sites across a contiguous region (Vitti et al. 2013). In this study, we computed the Composite likelihood ratio (CLR) (Kim and Stephan, 2002) with SweeD, which calculates this test using the relation between the likelihood of a sweep at a certain position in the genome by the product of the empirical site frequency spectrum over all SNPs (Pavlidis et al, 2013). We ran SweeD with the phased and polarized data, in a resolution of 200Kb windows by

chromosome, which corresponds to an average of 20 SNPs by each window, but we just consider the windows with more than 10 SNPs. We considered the higher values as an indicator of positive selection.

ASMC

The rapid rise in frequency of a beneficial allele due to a recent positive natural selection event provokes the coalescence of all individuals with the beneficial allele to a more recent common ancestor than expected under a neutral model. Thus, we checked if our results showed an unusually high density of very recent inferred TMRCA events using the Ascertained Sequentially Markovian Coalescent (ASMC) method (Palamara et al. 2018).

To run ASMC our first step was to produce the files needed for the ASMCprepareDecoding.jar script: the “.demo” file made with SMC++ (Terhorst et al. 2016); the “.freq” file produced with Plink and the “.disc” file provided together with ASMC package. This script provides the “.decodingQuantities” file that is necessary to run the ASMC program, together with the “.haps” files from our phased data. Then, we merge, normalize and plot our results with the tools suggested by the authors.

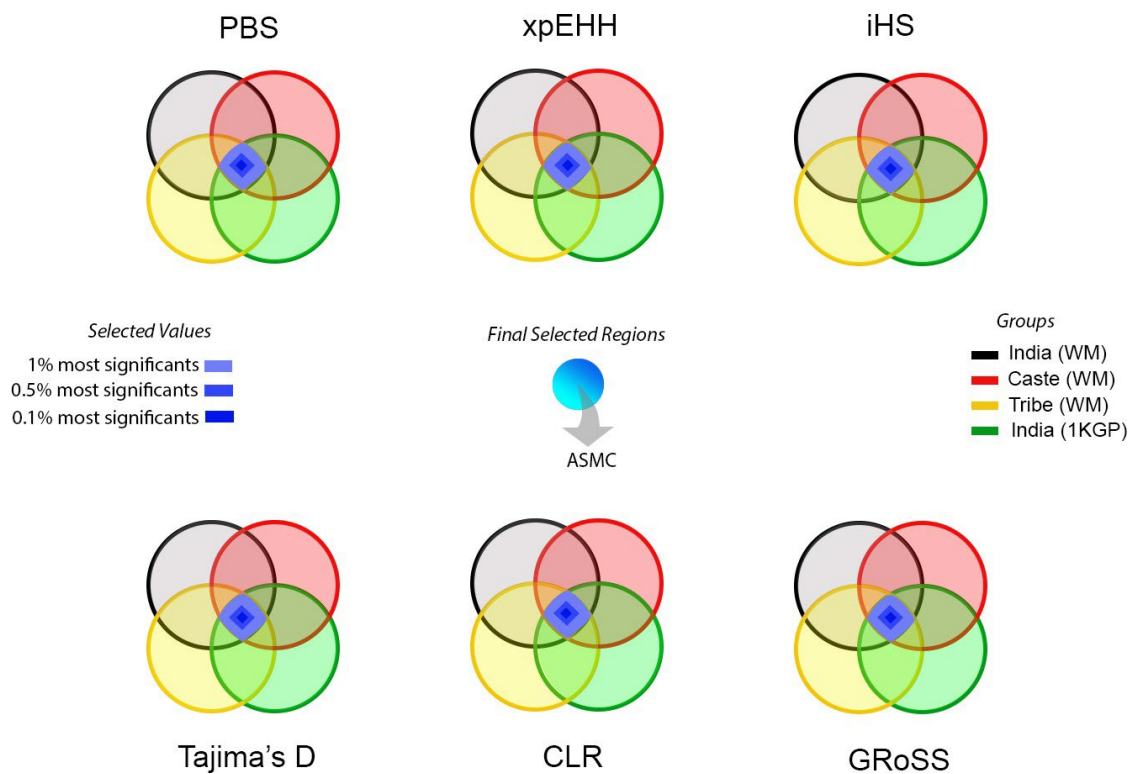


Figure 1. Schematic representation of the approach to identify putative selective regions. We applied six different methods to identify outliers (top 1%, 0.5% and 0.01% results) and selected regions that were observed in all population groups and were outliers for at least two independent methods. Additionally, we performed analyses using a novel coalescence-based method implemented in the program ASMC

Results

To achieve our aim to detect putative signatures of natural selection in the South Asian population with the minimum of false-positive results, we applied six different approaches in a dataset including ~283K SNPs markers that overlap between the WM and 1KG-India samples (Figure 1) and we report regions that were identified with at least two independent methods.

Our analysis identified a total of 107 genomic regions overlapping 434 genes distributed across all autosomal chromosomes, (Figure 2 - shows the number of regions (Figure 2A), the number of genes found by 2 or 3 independent methods (Figure 2B) and the number of genes in the top 1%, 0.5% and 0.1% of the distribution (Figure 2C). Supplementary Table 1 reports

these signals, including information about the methods for which the regions were identified as outliers. These regions are also reported in a condensed form in Supplementary Table 2).

Chromosome 16 is particularly rich in the number of genes, primarily due to a single region of 2.6 Mb on chromosome 16 that alone contains 78 genes. This region also shows enrichment in recent inferred TMRCA events using the ASMC program (Supplementary Figure 4).



Figure 2. Overview of our results. A) Distribution of the number of regions identified for each chromosome, for all thresholds (1%, 0.5% and 0.1%); B) Distribution of the number of genes located within putative selective regions using 1%, 0.5%, and 0.1% thresholds for each chromosome; C) Distribution of the number of genes located within putative selective regions identified with two or three methods for each chromosome, for all thresholds (1%, 0.5% and 0.1%); D) Percentage of signals in the top 1% shared by pair of populations (WM Castes, WH Tribes, WH full sample, and India IKG). The reference group is indicated in the Y-axis; for example, for the PBS method, 8.72% of the signals found in the WM Caste group are also found in the WM Tribe group, but just 7.78% of the signals identified in the WM Tribe group are found in the WM Caste group; E) Percentage of signals in the top 1% shared by different methods. The reference method is indicated in the Y-axis; for example, 33.5% of the signals identified using PBS are also observed with xpEHH.

We analyzed for each method what is the proportion of signals that are shared by each pair of populations. Figure 2D shows in matrix format the percentage of signals in the top 1% that

is shared by each pair of populations for each method. The method showing the largest overlap was Tajima's D, where the WM full sample shared 70.31% of the top 1% signals with the Indian 1KGP samples. In contrast, for the xpEHH method, these groups only shared 3.31% of the top 1% signals. We also evaluated what is the proportion of signals that are shared by different methods. This is presented in graphical format in Figure 2E. The methods showing the largest overlap were PBS and GRoSS. In contrast, there was no overlap in the signals identified by CLR and PBS.

Next, we compared the results identified using these six methods with those based on a recently described coalescence-based approach implemented in the package ASMC (Palamara et al. 2018). The graphs with the results corresponding to this method are presented for each chromosome as supplementary material (Supplementary Figure 4). We observed considerable congruence between our original signals and the results obtained with ASMC. Many of the putative genomic regions identified in our initial scan also show recent inferred times to the most recent common ancestor (TMRCA) using ASMC, as expected under recent positive selection. The ASMC method tends to show higher resolution than the other methods, with narrower regions that include a smaller number of genes.

There have been previous efforts to identify signatures of selection using exclusively South Asian samples (Metspalu et al. 2011, Karlsson et al., 2013; Juyal et al. 2014, Jonnalagadda et al. 2017), or including South Asian samples in the analyses (Liu et al., 2017). In Table 1, we highlight the genomic regions identified in our study that have been also reported in previous studies.

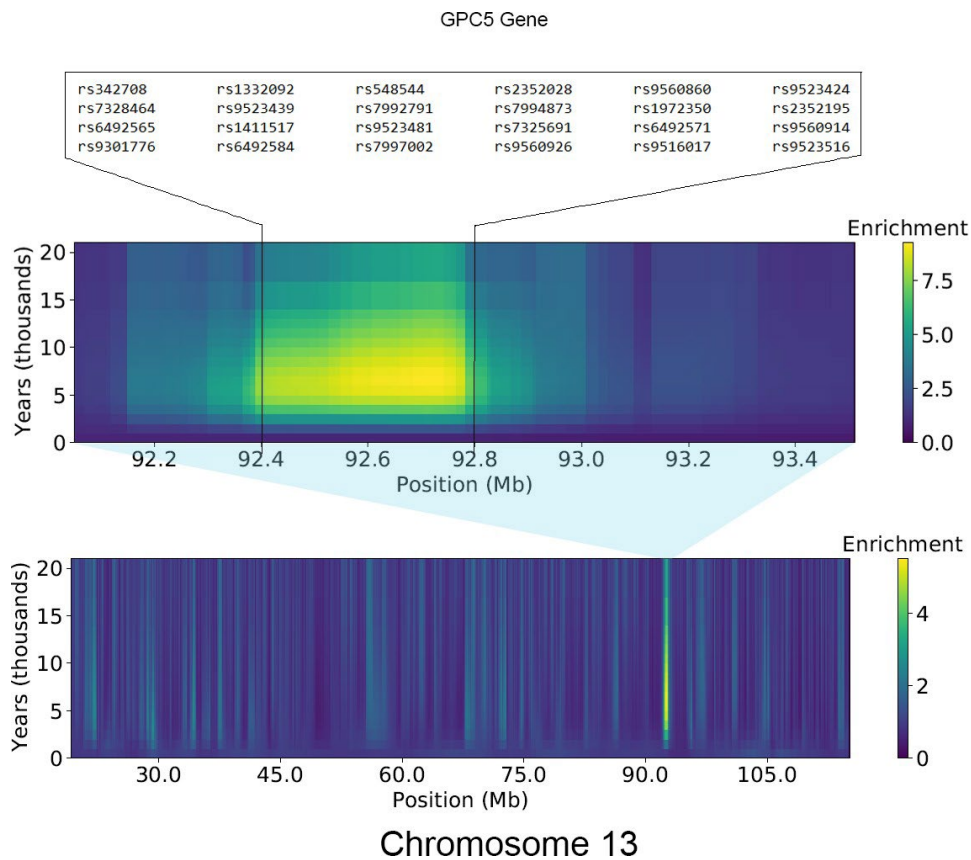


Figure 3. ASMC results for chromosome 13, showing with a greater resolution the region including the gene *GPC5* where the highest enrichment in recent coalescence events is concentrated on this chromosome.

Discussion

Genome-wide scans for signatures of selection can provide very useful insights about the role of natural selection driving the adaptation of our species after the out of Africa migration of anatomically modern humans. In this context, very few studies have specifically focused on South Asian populations (Metspalu et al, 2011, Karlsson et al., 2013; Juyal et al. 2014, Jonnalagadda et al. 2017), which have some of the highest genetic diversity observed outside Africa (Xing, et al. 2010, Mallick et al., 2016, Pagani et al., 2016, Peter et al., 2019 and Bergström et al., 2020, Jain et al., 2020).

In this study we have carried out genome-wide scans to identify putative signatures of natural selection in South Asia, using microarray-based data from several tribal and caste groups of West Maharashtra, India as well as high coverage sequencing data available from the 1KGP samples from India (GIH and ITU). We applied methods based on different strategies,

including haplotype length, population differentiation, site frequency spectrum, as well as recently developed methods based on admixture graphs and locus-specific pairwise coalescence times. In order to minimize the risk of false-positive signals, we only selected regions that were identified by at least two independent methods and were present in all the samples analyzed in the study (Figure 1).

Based on these analyses, we identified 107 genomic regions comprising 434 genes as potential candidates of positive selection in Indian populations. The average number of genes per region is 4.08, and the median is 1. The results are presented in supplementary tables 1 and 2. To our knowledge, at least 24 of the 107 regions identified in our study have been reported before in previous efforts to identify putative signals of natural selection in South Asian and East Asian populations (Metspalu et al., 2011; Suo et al., 2012; Karlsson et al., 2013; Liu et al., 2017; Perdomo-Sabogal and Nowick, 2019). These regions are reported in Table 1.

We did not observe any systematic excess of sharing of signals between the two groups of West Maharashtra (tribes and castes) with respect to the comparisons of these groups with the 1KG Indian samples, which may be reflective of the varied evolutionary and demographic events witnessed by these West Maharashtra caste and tribal groups. It is important to note that in a previous study (Debortoli et al., 2020) we showed that in Principal Component plots, the WM-Tribes and WM-Castes clustered separately from each other, whereas the WH-Castes were located closer to the 1KG South Asian samples. Additionally, Identity By Descent (IBD) analyses indicated that the WM-Tribes have had smaller effective population sizes and have been under stronger influence of genetic drift than the WM-Castes.

When comparing the results for each method in the different groups (WM-Castes, WH- Tribes, WM-All and India-1KG, Figure 2D), we observed some variation in the percentage of shared signals depending on the method. The lowest overlap between groups was observed for the PBS, xpEHH and iHS methods, and the highest overlap for the CLR and Tajima's D approaches, with the GroSS method showing intermediate values. With respect to the comparison of the putative selective signals between methods (Figure 2E), we observed that most of the signals identified by iHS, xpEHH and PBS were also detected using GroSS (between 88.9% and 93.4%, Figure 2E). More than 50% of the signals detected with CLR were also observed with GroSS. In contrast, only 13% of the signals identified with Tajima's D were also present in the GroSS output. We also observed that, aside from what was

reported for GRoSS, methods based on similar strategies tend to share more signals than they do with other methods. For example, the three methods based on population differentiation (PBS, xpEHH and GRoSS, for which we primarily carried out comparisons of European and South Asian samples) share more signals than they do with other methods (e.g. CLR or Tajima's D), and similar trends were observed for the two methods based on the site frequency spectrum (CLR and Tajima's D), with respect to the other approaches (PBS, xpEHH, iHS, CLR). Also, not surprisingly, the largest overlap for the iHS signals (with the exception of GRoSS) is with xpEHH. Both methods are haplotype-based methods, and are particularly well powered to identify selection on *de novo* mutations (SDN) that occurred less than 30,000 years ago. Previous studies have also reported limited overlap between different approaches used to identify signatures of selection (Biswas and Akey, 2006; Akey, 2009). This is not surprising given that these methods are based on different characteristics of the data and have different sensitivities to detect selective events depending on factors such as time and type of selection (Hancock and Di Rienzo, 2008; Oleksyk et al, 2010). Our strategy has been to select outliers identified with more than one method in all population groups in order to more reliably identify putative genomic regions under selection.

Additionally, we analyzed the data using a recently developed method based on locus-specific pairwise coalescence times (ASMC), and observed that in general this method shows concordant results with respect to those observed based on the other approaches, and in some cases provides a higher level of resolution. Supplementary Figure 4 shows the graphs generated by the program ASMC for each chromosome, including additional information about some specific regions (e.g. gene information and overlap with results described in previous studies). The genomic regions with the highest enrichment of recent coalescence events in our ASMC analysis (higher than 7) were located on chromosomes 1 (~248Mb), 4 (~80Mb), 6 (~29.5Mb), and 16 (~1Mb). The genes overlapping these regions are highlighted in Supplementary Figure 4.

We observed a substantial overlap of our signals with those reported in a selection study using samples of Bengali ethnicity from Bangladesh (BEB) that applied the Composite of Multiple Signals (CMS) method (Karlsson et al., 2013). This study explored the relationship of selection signals with selective pressure due to cholera. The authors reported that a number of genes identified in the putative selected regions were associated with cholera susceptibility in two separate cohorts. The region with the strongest signal of selection, located on chromosome 2 and encompassing five genes (*NRNP200*, *CIAO1*, *ITPRIPL1*, *NCAPH*, and

TMEM127) was also the region showing the strongest association with cholera, with the top associated SNPs located between the genes *NRNP200* and *ITPRIPL1*. We identified a very large region on chromosome 2 that spans almost 2 Megabases (from 96.9 Mb to 98.8 Mb) showing signatures of selection, which includes these two genes (Table 1, Supplementary Tables 1 and 2). It is important to note that Karlsson et al. (2013) described three independent putative selected regions in this genomic interval, the first from positions 96.2 to

96.4 Mb (including the *NRNP200* and *ITPRIPL1* genes), the second from positions 97.5 to 97.7 Mb (including *COX5B*, *ACTR1B* and *ZAP70*), and the third from positions 98.1 to 98.4 Mb (including *VWA3B* and *CNGA3*). All of these genes were identified in our initial analysis, and this broad region also shows the largest ASMC enrichment on chromosome 2 (Supplementary Figure 5). The region from chromosome 2 from 97.1 to 98.4 was also identified by Liu et al. (2017) in three samples from South Asia. In summary, there is strong evidence of positive selection acting on this genomic region in South Asian populations, but further studies will be required to elucidate the specific target/s of selection and the selective factors involved. Karlsson et al. (2013) also reported associations with cholera in three additional putative selected regions when focusing on the most severe cholera cases, encompassing the potassium ion transport genes *KCNH7* and *KCNH5*, and the ribosomal protein kinase gene *RPS6KB2*. Two of these three genes (*KCNH5* and *RPS6KB2*) were also identified in our study. In our analysis, *KCNH5* was the only gene present in a relatively narrow interval on chromosome 14 from 63.1 to 63.5 Mb (Table 1). A broader region on chromosome 14 (from 61.6 Mb to 64 Mb) including *KCNH5* was also identified as a putative selective signal in a study by Metspalu et al. (2011) in South Asian samples and this region shows strong enrichment in our ASMC results (Supplementary Figure 4N). In contrast, in our analyses, *RPS6KB2* is one of many genes identified in a very broad region on chromosome 11 spanning more than 3 Mb (from 65.4 to 68.8 Mb, Table 1). This region also shows strong enrichment in our ASMC analyses (Supplementary Figure 4K).

In addition to the three regions described above, many other regions reported by Karlsson et al. (2013) were also identified in our study (Table 1). Most of these regions overlap with less than 4 genes in our analyses, including a region on chromosome 7 (from 119.9 to 120.3 Mb) encompassing the gene *KCND2* (also identified in Liu et al. (2017) in South Asian samples and showing a strong ASMC enrichment), a short region on chromosome 10 (from 320 to 735 kb) encompassing the gene *DIP2C*, a region on chromosome 13 (from 92.0 to 93.5 Mb) encompassing the gene *GPC5* (also showing a strong ASMC enrichment), a region on chromosome 17 (from 58.7 to 59.4) encompassing the gene *BCAS3* (also identified in

Metspalu et al. (2011) study in South Asian samples), a region on chromosome 22 (from 46.7 to 46.9 Mb) encompassing the gene *CELSRI* (also identified in Metspalu et al. (2011) study in South Asian samples), a region on chromosome 2 (from 72.3 to 73.0 Mb) encompassing the genes *CYP26B1*, *EXOC6B*, *SNORD78* (also reported in Liu et al., 2017 in multiple Asian samples, including South Asian samples), a region on chromosome 2 (from 241.6 to 242.0 Mb), encompassing the genes *KIF1A*, *AGXT*, *C2orf54* and *SNED1*, a region on chromosome 7 (from 111.3 to 111.5 Mb) encompassing the *DOCK5* and *BC043243* genes (also showing extreme ASMC enrichment), a region on chromosome 9 (from 123.7 to 124.1 Mb) encompassing the genes *C5*, *CNTRL*, *RAB14* and *GSN* (also reported in Metspalu et al., (2011) in South Asian samples), and a region on chromosome 10 (from 118.1 to 118.3 Mb) encompassing the *PNLIPRP3* and *JA611286* genes. Figure 3 shows the ASMC results for Chromosome 13, clearly showing a strong enrichment of recent coalescence events in a relatively narrow genomic interval including the *GPC5* gene. While there is very strong evidence pointing to the action of positive selection in these regions, an evaluation of the associations reported in the GWAS catalog (<https://www.ebi.ac.uk/gwas/>) indicates that most of these genes have pleiotropic effects and are associated with multiple traits in GWAS studies, so it is challenging to determine the specific selective factors driving these signals.

In contrast to the regions described above, which include a small number of genes, one of the regions identified in our study and also previously identified in South Asian (Karlsson et al., 2013) and East Asian samples (Perdomo-Sabogal and Nowick 2019) is an extremely gene-rich region spanning around 2.5 Mb located on Chromosome 16 between positions ~29.46 and 32.1 Mb (Table 1, Supplementary Tables 1 and 2, Supplementary Figure 6). This region includes 78 genes and is characterized by the presence of gene regulatory factors (GRFs), including zinc-finger (ZNF) genes with a Krüppel-associated box (KRAB-ZNF). KRAB-ZNF genes have undergone extensive expansion in mammals and have been rapidly evolving in primates, and several of these genes are considered to be human-specific (Nowick et al., 2010; Perdomo-Sabogal and Nowick, 2019). Perdomo-Sabogal and Nowick (2019) speculated that positive selection may have influenced diversity in several classes of GRF genes, thus playing an important role in local adaptation of human populations, and in their study, they identified numerous KRAB-ZNF clusters exhibiting evidence for positive selection in three human populations (the CEU, CHB and YRI 1KG samples). One of these GRF clusters overlaps with the chromosome 16 region identified in our study and it is possible that these regulatory genes have been the target of positive selection. However, it should be mentioned that this broad region also includes many non-GRF genes, thus making

it difficult to pinpoint the target of selection. Interestingly, this region also includes the *VKORC1* gene, a very important pharmacogene that encodes a key enzyme in the vitamin K cycle and plays a key role in the coagulation pathway (Owen et al., 2010). *VKORC1* is the pharmacological target of warfarin and previous studies have reported that positive selection may have played a role in the variability of anticoagulant response in humans (Ross et al., 2010; Patillon et al., 2012).

In addition to the signals shared between our study and Karlsson et al. (2013) scan of positive selection in South Asians, we identified other outlier regions that have been reported in other studies (Table 1). These include a region on Chromosome 11 (from 61.4 to 62.6 Mb) encompassing many genes previously reported by Suo et al. (2012), a very large region on chromosome 11 (from 126.2 to 132.2 Mb) encompassing the genes *KIRREL3*, *DJ031150* and *NTM* previously reported by Metspalu et al. (2011), a region on chromosome 20 (from 53.0 to 53.3 Mb) encompassing the *BCAS3* gene previously reported by Metspalu et al. (2011) in South Asian populations and Sabeti et al. (2007) in European populations, a narrow region spanning a few kilobases on chromosome 22 (from 35.46 to 35.48 Mb) encompassing the homeobox *ISX* gene previously reported by Metspalu et al., (2011), which has been reported to be a critical molecular mediator of the cross-talk between diet and immunity (Widjaja-Adhi et al., 2017), and a region located on the short arm of chromosome 6 (from 29.5 to 33.1 Mb) previously reported by Liu et al (2017) and Suo et al. (2012). This corresponds to the Major Histocompatibility Complex (MHC) region, which is a well-known target of selection in the human genome (Prugnolle et al., 2005; Meyer et al., 2017). In our ASMC analysis, we observed the largest enrichment in the region around ~29.9Mb, which includes 7 genes (*SNORD32B*, *OR2H2*, *GABBRI*, *MOG*, *ZFP57*, *HLA-F*, *HLA-F-ASI*) (Supplementary Figure 7).

Conclusion

In this study, we highlight numerous genomic regions that may have been under positive selection in South Asian populations. We used tests of positive selection based on different statistics, including PBS, xpEHH, iHS, CLR, Tajima's D, as well as two recently developed methods: GRoSS and ASMC. In order to minimize the risk of false positives, we selected regions that are outliers in all the samples included in the study using more than one method. We identified putative selection signals in 107 regions encompassing 434 genes. At least 24

of these genomic regions have been identified in previous selection scans in South Asian populations or in other population groups. Many of the regions overlap with only one gene.

It is important to consider some of the limitations of this study. The first limitation is that our initial analysis was based on microarray-based data (approximately 300,000 markers) and not Whole Genome Sequencing (WGS) data, which is the ideal type of data to use for this type of studies. However, we compared the output of our analyses with the results obtained using the high-coverage genome sequencing data from two Indian samples of the 1KG Project (GIH and ITU), with highly consistent results: More than 90% of the regions identified in the microarray-based analysis are also outliers in the WGS analysis (Supplementary Table 2). The second limitation is that strategies based on the identification of outliers in the empirical distribution of the relevant parameters cannot fully guarantee that all these regions have been under the influence of positive selection. We tried to minimize the risks of false positives by selecting regions that are outliers in all the samples included in the study (WM-Tribes, WM-Castes, WM-All, 1KG-India) using more than one method. The microarray-based signals are very consistent with the WGS-based analysis and with our independent analysis using the ASMC method. It is also important to note that many of the regions identified in our study have been also reported in previous efforts to identify signatures of selection and from this perspective, the regions highlighted in Table 1 have particularly strong support. The third limitation is that, although we have identified regions that have been putatively under selection in South Asian populations, in many cases the regions include multiple genes, and it is not possible to identify which gene has been the target of selection. Similarly, even for the regions overlapping with only one gene, it is challenging to know what selective factors may have been involved as most of the genes are pleiotropic and have been associated with a broad range of traits.

Despite these limitations, our study provides important insights on genomic regions that may have been under positive selection in South Asian populations, and further research may identify the polymorphisms in these genomic regions that drove adaptation to the novel conditions anatomically modern humans encountered after the migration out of Africa.

Table 1. List of the regions with putative signatures of natural selection in our study that have been described in other studies, with a particular emphasis in studies in South Asian populations or other Asian groups. 1: Metsupalu et al. 2011, 2: Suo et al. 2012, 3: Karlsson et al. 2013, 4: Liu et al. 2017, 5: Perdomo-Sabogal and Nowick, 2019. In Blue, we show regions

with results in the 0.5% of the most significant values for at least one method, and in red the results in the top 0.1% most significant results for at least one method. We also highlight the regions that also have significant results in the 1kbp high coverage data (1kbp_HC), and in the PopHuman Browser with iHS.

chr	Start	End	SNPs	Genes	Shared with other studies	Shared with 1KGP_HC	Shared with PopHuman Browser (iHS)
1	234663636	235491532	118	LOC100506795,TOMM20,SNORA14B,RBM34,ARID4B,MIR4753	3	PBS(0.1%)	GIH,ITU
2	72356366	73053177	20	CYP26B1,EXOC6B,SNORD78	3,4	Tajimas'D(0.5%),	GIH,ITU
2	96940073	98858761	56	SNRNP200,ITPRIPL1,NCAPH,ARID5A,KANSL3,FER1L5,ANKRD39,SEM A4C,FAM178B,FAHD2B,ANKRD36,ANKRD36B,COX5B,ACTR1B,LOC7 28537,ZAP70,VWA3B	3,4	PBS(0.1%), xpEHH(0.1%)	GIH,ITU
2	241662829	242033643	60	KIF1A,AGXT,C2orf54,SNED1	3		GIH,ITU
4	39289068	39529218	26	RFC1,KLB,RPL9,LIAS,LOC401127,UGDH	3		GIH,ITU
6	29550028	33086926	3545	SNORD32B,OR2H2,GABBR1,MOG,ZFP57,HLA-F,HLA-F-	2,4	PBS(0.5%)	GIH,ITU
7	111366163	111461829	12	DOCK4,BCO43243	3	PBS(0.5%)	GIH,ITU
7	119913721	120390387	19	KCND2	3,4	Tajimas'D(0.5%),	GIH,ITU
9	123714613	124095120	18	C5,CNTRL,RAB14,GSN	1,3	PBS(0.1%)	GIH,ITU
10	320129	735608	23	DIP2C	3	Tajimas'D(1%)	GIH,ITU
10	118187423	118261387	9	PNLIPRP3,JA611286	3	Tajimas'D(0.5%)	
11	61447904	62622555	144	DAGLA,MYRF,DKFZP434K028,BCO20196,TMEM258,MIR611,FEN1,F ADS1,MIR1908,FADS2,FADS3,RAB31L1,BEST1,FTH1,BC132896,SNO KAT5,RNASEH2C,AP5B1,SNX32,CFL1,MUS81,EFEMP2,CTSW,FIBP,CC DC85B,FOSL1,KLC2,RAB1B,AK125412,CNIH2,YIF1A,TMEM151A,CD 248,RIN1,BRMS1,B3GNT1,SLC29A2,AX747485,NPAS4,MRPL11,LOC 100130987,POLD4,CLCF1,RAD9A,PPP1CA,TBC1D10C,CARNS1,RPS6 KB2,PTPRCAP,CORO1B,GPR152,CABP4,TMEM134,AIP,PITPNM1,CD K2AP2,CABP2,C11orf24,LRP5,MRGPRF,BC039516,TPCN2	2	PBS(0.1%), xpEHH(0.1%)	GIH,ITU
11	65479472	68846261	198		3	CLR(0.1%),Tajimas'D(0.5%), xpEHH(0.1%), PBS(0.5%)	GIH,ITU
11	126293395	132206716	884	KIRREL3,DJO31150,NTM	1	PBS(0.5%)	GIH,ITU
13	92050934	93519487	139	GPC5	3	xpEHH(0.1%), PBS(1%)	GIH,ITU
14	63173944	63511955	35	KCNH5	1,3	xpEHH(0.1%), PBS(0.5%)	GIH,ITU
16	29464909	32077476	122	BOLA2,KIF22,MAZ,AB209061,AK097472,PRRT2,PAGR1,BC029255, MVP,CDIPT,CDIPT- AS1,SEZ6L2,ASPHD1,KCTD13,TMEM219,TAOK2,HIRIP3,LOC595101, CD2BP2,TBC1D10B,MYLPP,SEPT1,ZNF48,SEPT2,ZNF771,DCTPP1,SE PHS2,ITGAL,MIR4518,ZNF768,ZNF747,AK056973,ZNF764,ZNF688, ZNF785,ZNF689,PRR14,FBRS,LOC730183,SRCAP,SNORA30,LOC100 862671,PHKG2,C16orf93,RNF40,ZNF629,BCL7C,MIR4519,BC0739 28,MIR762,CTF1,FBXL19- AS1,FBXL19,ORAI3,SETD1A,HSD3B7,STX1B,STX4,BC039500,ZNF668 ,ZNF646,PRSS53,VKORC1,BCKDK,KAT8,PRSS8,PRSS36,FUS,TL5/FUS- ERG,PYCARD,C,16orf98,TRIM72,PYDC1,ITGAM,DL489986,ITGAX,IGH V3-07,IGH	3,5	Tajimas'D(0.1%), PBS(0.1%), xpEHH(0.1%)	GIH,ITU
16	46760587	47735434	15	MYLK3,C16orf87,GPT2,ITFG1,PHKB	3	Tajimas'D(0.5%),	
16	87117167	87457487	57	AK125749,C16orf95,FBXO31,MAP1LC3B,ZCCHC14	3	PBS(0.1%)	GIH,ITU
17	17876126	18011299	4	LRRC48,ATPAF2,BC150162,GID4,DRG2	3	PBS(0.5%)	GIH,ITU
17	58755212	59470192	58	BCAS3	1,3	Tajimas'D(0.1%), CLR(0.1%), PBS(1%)	GIH,ITU
20	53092265	53267710	13	DOK5	1	Tajimas'D(0.1%), CLR(0.1%), PBS(0.1%)	
22	35462129	35483380	3	ISX	1		
22	46756730	46933067	49	CELSR1	1,3	Tajimas'D(0.5%)	

Works Cited

Akey, Joshua M. “Constructing Genomic Maps of Positive Selection in Humans: Where Do We Go from Here?” *Genome Research*, vol. 19, no. 5, 2009, pp. 711–722.,

doi:10.1101/gr.086652.108.

Basu, Analabha, et al. “Genomic Reconstruction of the History of Extant Populations of India Reveals Five Distinct Ancestral Components and a Complex Structure.” *Proceedings*

- of the National Academy of Sciences*, vol. 113, no. 6, 2016, pp. 1594–1599.,
doi:10.1073/pnas.1513197113.
- Bergström, Anders, et al. “Insights into Human Genetic Variation and Population History from 929 Diverse Genomes.” *Science*, vol. 367, no. 6484, 2020, doi:10.1126/science.aay5012.
- Bigham, Abigail W., et al. “Identifying Positive Selection Candidate Loci for High-Altitude Adaptation in Andean Populations.” *Human Genomics*, vol. 4, no. 2, 2009,
doi:10.1186/1479-7364-4-2-79.
- Biswas, Shameek, and Joshua M. Akey. “Genomic Insights into Positive Selection.” *Trends in Genetics*, vol. 22, no. 8, 2006, pp. 437–446., doi:10.1016/j.tig.2006.06.005.
- Cavalli-Sforza, Luigi Luca. “‘Genetic Drift’ in an Italian Population.” *Scientific American*, vol. 221, no. 2, 1969, pp. 30–37., doi:10.1038/scientificamerican0869-30.
- Chaubey, Gyaneshwer, et al. “Peopling of South Asia: Investigating the Caste–Tribe Continuum in India.” *BioEssays*, vol. 29, no. 1, 2006, pp. 91–100., doi:10.1002/bies.20525.
- Crawford, Jacob E., et al. “Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans.” *The American Journal of Human Genetics*, vol. 101, no. 5, 2017, pp. 752–767., doi:10.1016/j.ajhg.2017.09.023.
- Debortoli, Guilherme, et al. “Novel Insights on Demographic History of Tribal and Caste Groups from West Maharashtra (India) Using Genome-Wide Data.” *Scientific Reports*, vol. 10, no. 1, 2020, doi:10.1038/s41598-020-66953-3.
- Delaneau, Olivier, et al. “Accurate, Scalable and Integrative Haplotype Estimation.” *Nature Communications*, vol. 10, no. 1, 2019, doi:10.1038/s41467-019-13225-y.
- Haak, Wolfgang, et al. “Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe.” *Nature*, vol. 522, no. 7555, 2015, pp. 207–211., doi:10.1038/nature14317.

- Hancock, Angela M., and Anna Di Rienzo. "Detecting the Genetic Signature of Natural Selection in Human Populations: Models, Methods, and Data." *Annual Review of Anthropology*, vol. 37, no. 1, 2008, pp. 197–217., doi:10.1146/annurev.anthro.37.081407.085141.
- Jain, Abhinav, et al. "IndiGenomes: a Comprehensive Resource of Genetic Variants from over 1000 Indian Genomes." *Nucleic Acids Research*, 2020, doi:10.1093/nar/gkaa923.
- Jonnalagadda, Manjari, et al. "A Genome-Wide Association Study of Skin and Iris Pigmentation among Individuals of South Asian Ancestry." *Genome Biology and Evolution*, vol. 11, no. 4, 2019, pp. 1066–1076., doi:10.1093/gbe/evz057.
- Jonnalagadda, Manjari, et al. "Identifying Signatures of Positive Selection in Pigmentation Genes in Two South Asian Populations." *American Journal of Human Biology*, vol. 29, no. 5, 2017, doi:10.1002/ajhb.23012.
- Jonnalagadda, Manjari, et al. "Skin Pigmentation Variation among Populations of West Maharashtra, India." *American Journal of Human Biology*, vol. 28, no. 1, 2015, pp. 36–43., doi:10.1002/ajhb.22738.
- Juyal, Garima, et al. "Population and Genomic Lessons from Genetic Analysis of Two Indian Populations." *Human Genetics*, vol. 133, no. 10, 2014, pp. 1273–1287., doi:10.1007/s00439-014-1462-0.
- Karlsson, Elinor K., et al. "Natural Selection in a Bangladeshi Population from the Cholera-Endemic Ganges River Delta." *Science Translational Medicine*, vol. 5, no. 192, 2013, doi:10.1126/scitranslmed.3006338.
- Kim, Yuseob, and Wolfgang Stephan. "Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome." *Genetics*, vol. 160, no. 2, 2002, pp. 765–777., doi:10.1093/genetics/160.2.765.
- Kivisild, T., et al. "The Genetic Heritage of the Earliest Settlers Persists Both in Indian Tribal and Caste Populations." *The American Journal of Human Genetics*, vol. 72, no. 2,

2003, pp. 313–332., doi:10.1086/346068.

Lazaridis, Iosif, et al. “Genomic Insights into the Origin of Farming in the Ancient Near East.”

Nature, vol. 536, no. 7617, 2016, pp. 419–424., doi:10.1038/nature19310.

Liu, Xuanyao, et al. “Characterising Private and Shared Signatures of Positive Selection in 37 Asian

Populations.” *European Journal of Human Genetics*, vol. 25, no. 4, 2017, pp.

499–508., doi:10.1038/ejhg.2016.181.

Loh, Po-Ru, et al. “Reference-Based Phasing Using the Haplotype Reference Consortium Panel.” 2016,

doi:10.1101/052308.

Mallick, Swapan, et al. “The Simons Genome Diversity Project: 300 Genomes from 142 Diverse

Populations.” *Nature*, vol. 538, no. 7624, 2016, pp. 201–206., doi:10.1038/nature18964.

Mathieson, Iain, et al. “Genome-Wide Patterns of Selection in 230 Ancient Eurasians.”

Nature, vol. 528, no. 7583, 2015, pp. 499–503., doi:10.1038/nature16152.

Mccoll, Hugh, et al. “The Prehistoric Peopling of Southeast Asia.” *Science*, vol. 361, no.

6397, 2018, pp. 88–92., doi:10.1126/science.aat3628.

Mellars, P., et al. “Genetic and Archaeological Perspectives on the Initial Modern Human Colonization

of Southern Asia.” *Proceedings of the National Academy of Sciences*, vol. 110, no. 26, 2013, pp.

10699–10704., doi:10.1073/pnas.1306043110.

Metspalu, Mait, et al. “Shared and Unique Components of Human Population Structure and

Genome-Wide Signals of Positive Selection in South Asia.” *The American Journal of Human*

Genetics, vol. 89, no. 6, 2011, pp. 731–744., doi:10.1016/j.ajhg.2011.11.010.

Metspalu, Mait, et al. “The Genetic Makings of South Asia.” *Current Opinion in Genetics&*

Development, vol. 53, 2018, pp. 128–133., doi:10.1016/j.gde.2018.09.003.

Meyer, Diogo, et al. “A Genomic Perspective on HLA Evolution.” *Immunogenetics*, vol. 70, no. 1,

2017, pp. 5–27., doi:10.1007/s00251-017-1017-3.

Nakatsuka, Nathan, et al. “The Promise of Discovering Population-Specific

- Disease-Associated Genes in South Asia.” *Nature Genetics*, vol. 49, no. 9, 2017, pp.1403–1407., doi:10.1038/ng.3917.
- Narasimhan, Vagheesh M., et al. “The Formation of Human Populations in South and Central Asia.” *Science*, vol. 365, no. 6457, 2019, doi:10.1126/science.aat7487.
- Nowick, K., et al. “Rapid Sequence and Expression Divergence Suggest Selection for Novel Function in Primate-Specific KRAB-ZNF Genes.” *Molecular Biology and Evolution*, vol. 27, no. 11, 2010, pp. 2606–2617., doi:10.1093/molbev/msq157.
- Oleksyk, Taras K., et al. “Genome-Wide Scans for Footprints of Natural Selection.” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1537, 2010, pp. 185–205., doi:10.1098/rstb.2009.0219.
- Owen, Ryan P., et al. “VKORC1 Pharmacogenomics Summary.” *Pharmacogenetics and Genomics*, vol. 20, no. 10, 2010, pp. 642–644., doi:10.1097/fpc.0b013e32833433b6.
- Pagani, Luca, et al. “Genomic Analyses Inform on Migration Events during the Peopling of Eurasia.” *Nature*, vol. 538, no. 7624, 2016, pp. 238–242., doi:10.1038/nature19792.
- Palamara, Pier Francesco, et al. “High-Throughput Inference of Pairwise Coalescence Times Identifies Signals of Selection and Enriched Disease Heritability.” *Nature Genetics*, vol. 50, no. 9, 2018, pp. 1311–1317., doi:10.1038/s41588-018-0177-x.
- Patillon, Blandine, et al. “Positive Selection in the Chromosome 16 VKORC1 Genomic Region Has Contributed to the Variability of Anticoagulant Response in Humans.” *PLoS ONE*, vol. 7, no. 12, 2012, doi:10.1371/journal.pone.0053049.
- Patin, Etienne, et al. “Dispersals and Genetic Adaptation of Bantu-Speaking Populations in Africa and North America.” *Science*, vol. 356, no. 6337, 2017, pp. 543–546., doi:10.1126/science.aal1988.
- Patterson, Nick, et al. “Ancient Admixture in Human History.” *Genetics*, vol. 192, no. 3, 2012, pp. 1065–1093., doi:10.1534/genetics.112.145037.
- Pavlidis, Pavlos, et al. “SweeD: Likelihood-Based Detection of Selective Sweeps in

- Thousands of Genomes.” *Molecular Biology and Evolution*, vol. 30, no. 9, 2013, pp. 2224–2234., doi:10.1093/molbev/mst112.
- Perdomo-Sabogal, Álvaro, and Katja Nowick. “Genetic Variation in Human Gene Regulatory Factors Uncovers Regulatory Roles in Local Adaptation and Disease.” *Genome Biology and Evolution*, vol. 11, no. 8, 2019, pp. 2178–2193., doi:10.1093/gbe/evz131.
- Peter, Benjamin M, et al. “Genetic Landscapes Reveal How Human Genetic Diversity Aligns with Geography.” *Molecular Biology and Evolution*, vol. 37, no. 4, 2019, pp. 943–951., doi:10.1093/molbev/msz280.
- Prugnolle, Franck, et al. “Pathogen-Driven Selection and Worldwide HLA Class I Diversity.” *Current Biology*, vol. 15, no. 11, 2005, pp. 1022–1027., doi:10.1016/j.cub.2005.04.050.
- Racimo, Fernando, et al. “Detecting Polygenic Adaptation in Admixture Graphs.” *Genetics*, vol. 208, no. 4, 2018, pp. 1565–1584., doi:10.1534/genetics.117.300489.
- Reddy, B. Mohan, et al. “Molecular Genetic Perspectives on the Indian Social Structure.” *American Journal of Human Biology*, vol. 22, no. 3, 2009, pp. 410–417., doi:10.1002/ajhb.20983.
- Refoyo-Martínez, Alba, et al. “Identifying Loci under Positive Selection in Complex Population Histories.” 2018, doi:10.1101/453092.
- Reich, David, et al. “Reconstructing Indian Population History.” *Nature*, vol. 461, no. 7263, 2009, pp. 489–494., doi:10.1038/nature08365.
- Ross, Kendra A, et al. “Worldwide Allele Frequency Distribution of Four Polymorphisms Associated with Warfarin Dose Requirements.” *Journal of Human Genetics*, vol. 55, no. 9, 2010, pp. 582–589., doi:10.1038/jhg.2010.73.
- Sabeti, P. C., et al. “Positive Natural Selection in the Human Lineage.” *Science*, vol. 312, no. 5780, 2006, pp. 1614–1620., doi:10.1126/science.1124309.
- Sabeti, Pardis C., et al. “Detecting Recent Positive Selection in the Human Genome from Haplotype Structure.” *Nature*, vol. 419, no. 6909, 2002, pp. 832–837.,

doi:10.1038/nature01140.

Sabeti, Pardis C., et al. “Genome-Wide Detection and Characterization of Positive Selection in Human Populations.” *Nature*, vol. 449, no. 7164, 2007, pp. 913–918.,

doi:10.1038/nature06250.

Sambrook, John J., et al. *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor Laboratory Press, 1989.

Shinde, Vasant, et al. “An Ancient Harappan Genome Lacks Ancestry from SteppePastoralists or Iranian Farmers.” *Cell*, vol. 179, no. 3, 2019, doi:10.1016/j.cell.2019.08.048.

Silva, Marina, et al. “A Genetic Chronology for the Indian Subcontinent Points to HeavilySex-Biased Dispersals.” *BMC Evolutionary Biology*, vol. 17, no. 1, 2017,

doi:10.1186/s12862-017-0936-9.

Suo, Chen, et al. “Natural Positive Selection and North–South Genetic Diversity in East Asia.”

European Journal of Human Genetics, vol. 20, no. 1, 2011, pp. 102–110.,

doi:10.1038/ejhg.2011.139.

Szpiech, Z. A., and R. D. Hernandez. “Selscan: An Efficient Multithreaded Program to Perform

EHH-Based Scans for Positive Selection.” *Molecular Biology and Evolution*, vol. 31, no. 10,

2014, pp. 2824–2827., doi:10.1093/molbev/msu211.

Tajima, F. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNAPolymorphism.”

Genetics, vol. 123, no. 3, 1989, pp. 585–595., doi:10.1093/genetics/123.3.585.

Terhorst, Jonathan, et al. “Robust and Scalable Inference of Population History from Hundreds of

Unphased Whole Genomes.” *Nature Genetics*, vol. 49, no. 2, 2016, pp.303–309.,

doi:10.1038/ng.3748.

Vitti, Joseph J., et al. “Detecting Natural Selection in Genomic Data.” *Annual Review of Genetics*, vol.

47, no. 1, 2013, pp. 97–120.,

doi:10.1146/annurev-genet-111212-133526.

Voight, Benjamin F, et al. “A Map of Recent Positive Selection in the Human Genome.” *PLoS Biology*, vol. 4, no. 3, 2006, doi:10.1371/journal.pbio.0040072.

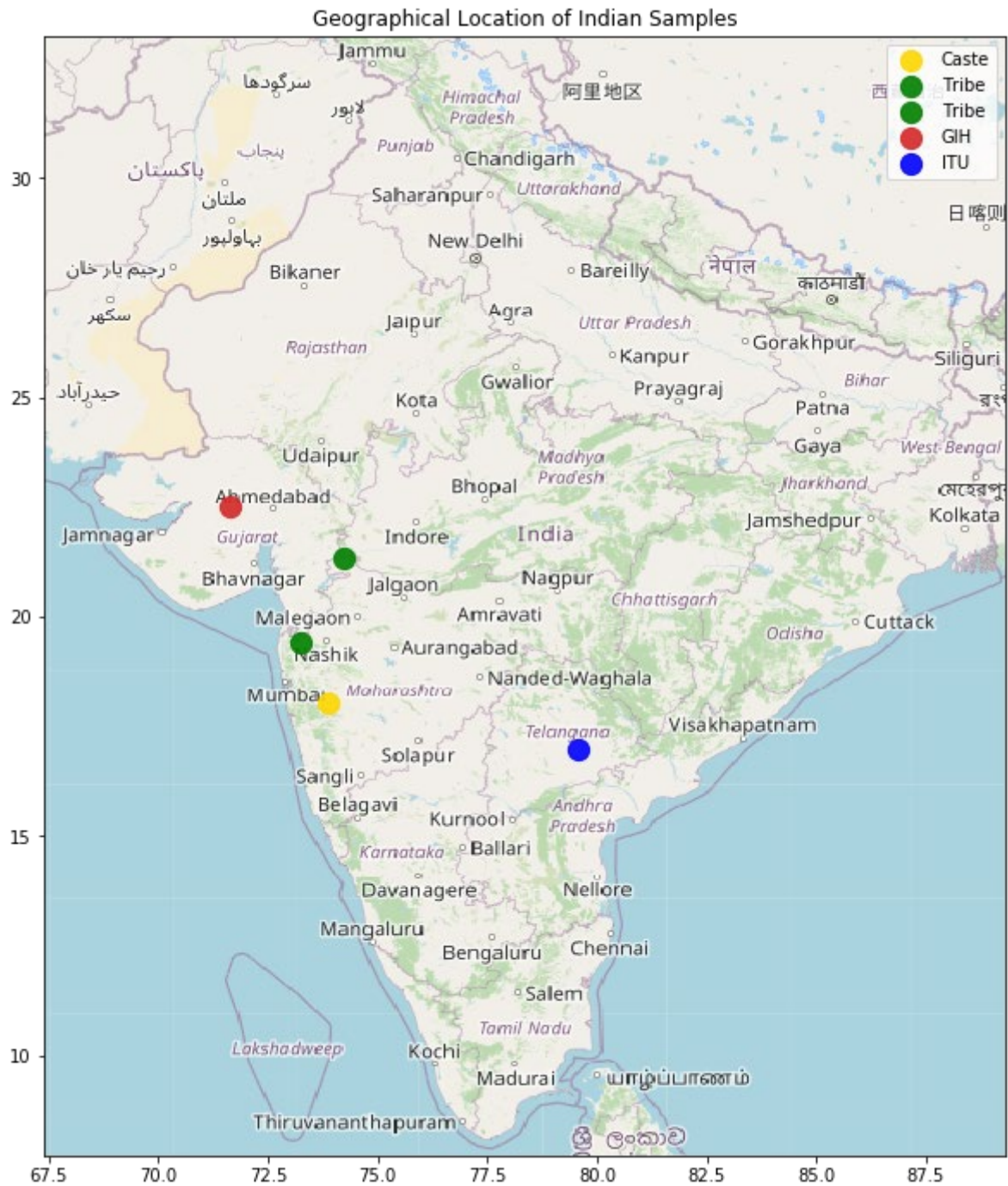
Widjaja-Adhi, Made Airanthi K., et al. “Transcription Factor ISX Mediates the Cross Talk between Diet and Immunity.” *Proceedings of the National Academy of Sciences*, vol.114, no. 43, 2017, pp. 11530–11535., doi:10.1073/pnas.1714963114.

Xing, Jinchuan, et al. “Genetic Diversity in India and the Inference of Eurasian Population Expansion.” *Genome Biology*, vol. 11, no. 11, 2010, doi:10.1186/gb-2010-11-11-r113.

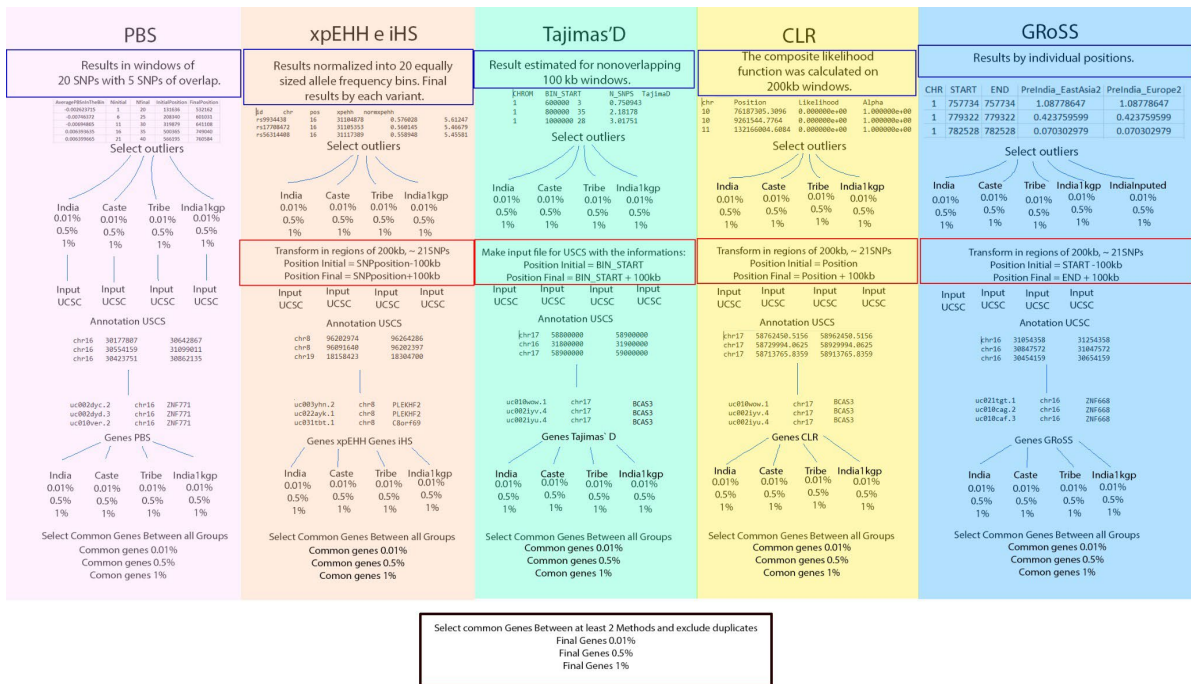
Yi, Xin, et al. “Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude.” *Science*, vol. 329, no. 5987, 2010, pp. 75–78., doi:10.1126/science.1190371.

Auton, Adam, et al. “A Global Reference for Human Genetic Variation.” *Nature*, vol. 526, no. 7571, 2015, pp. 68–74., doi:10.1038/nature15393.

Supplementary Material



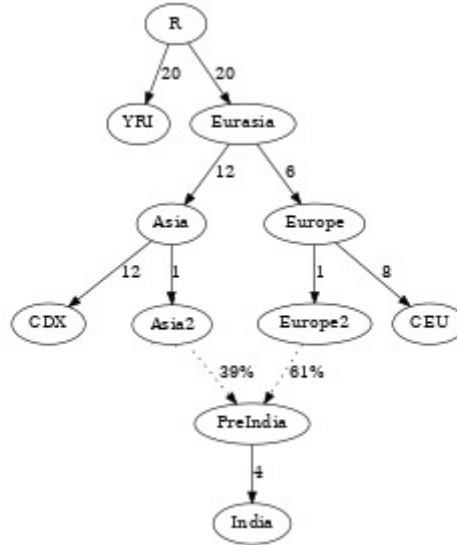
Supplementary Figure 1. The geographical location of the samples analyzed in this study.



Supplementary Figure 2. Depicts a schematic representation of our approach. Additional details of the methods used in our analyses are provided in the main text.

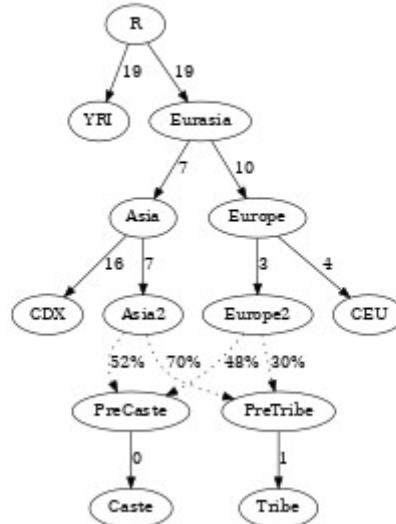
YRI CDX YRI CDX 0.062859 0.062869 0.000011 0.000461 0.023

A)



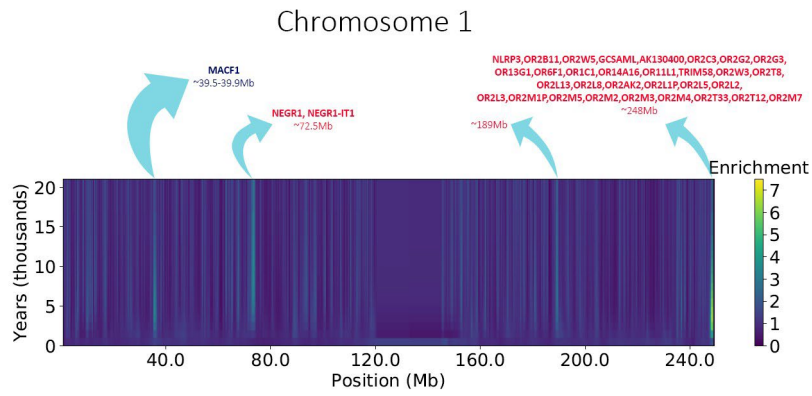
YRI CEU YRI Cas 0.043604 0.043697 0.000094 0.000353 0.266

B)

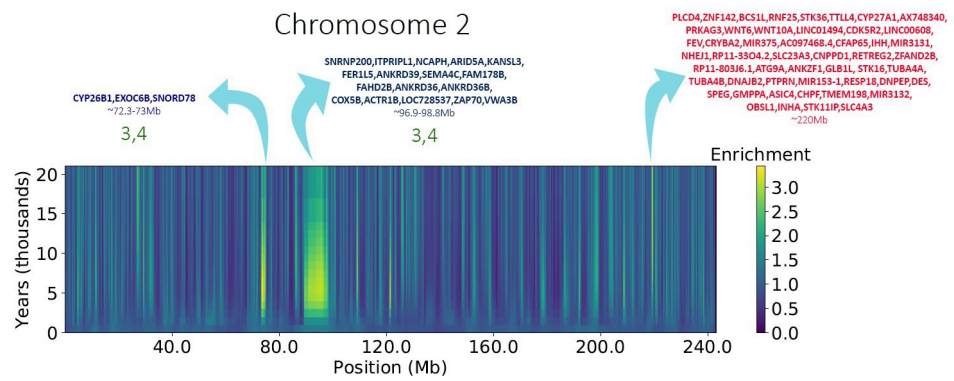


Supplementary Figure 3. Admixture graphs showing our two approaches. A) Admixture graph including a preIndia group resulting from admixture from a European and an Asian source. B) Admixture graph including preTribe and preCaste groups as a result of admixture between a European and Asian source.

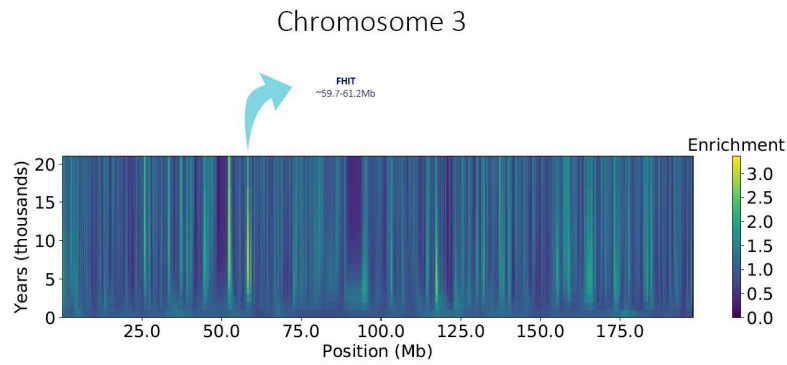
A)



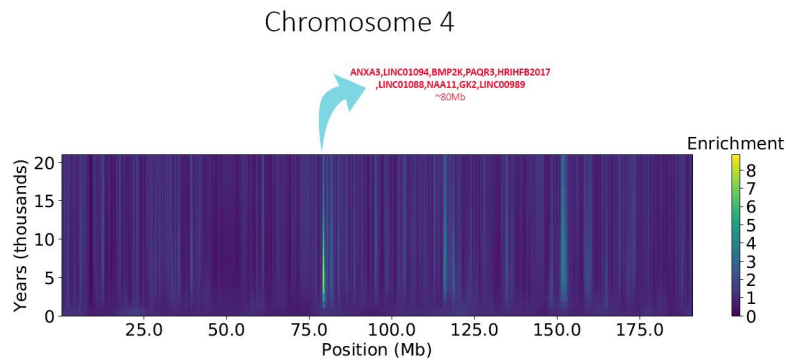
B)



C)

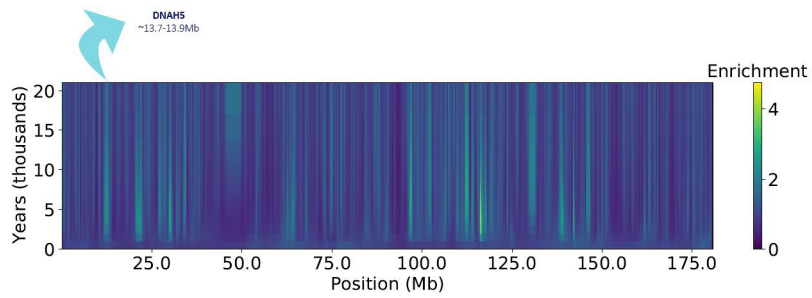


D)



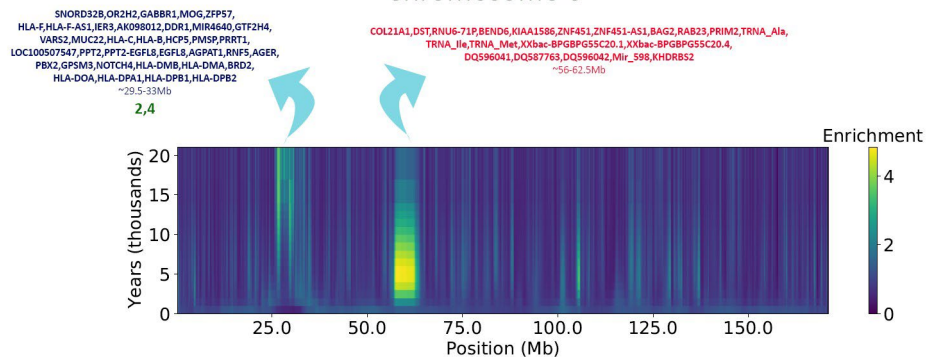
E)

Chromosome 5



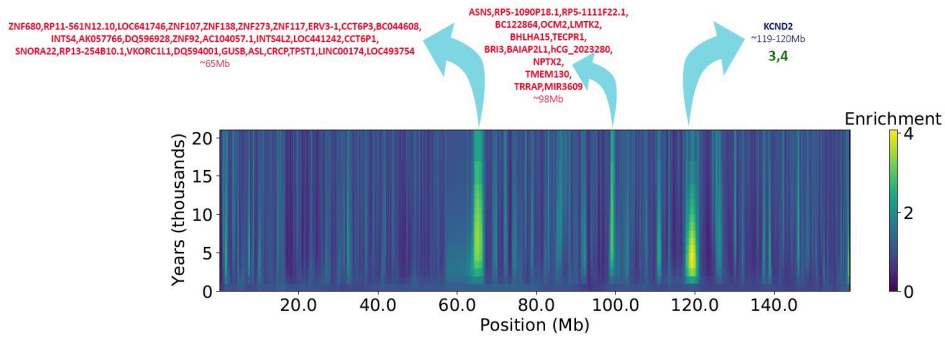
F)

Chromosome 6



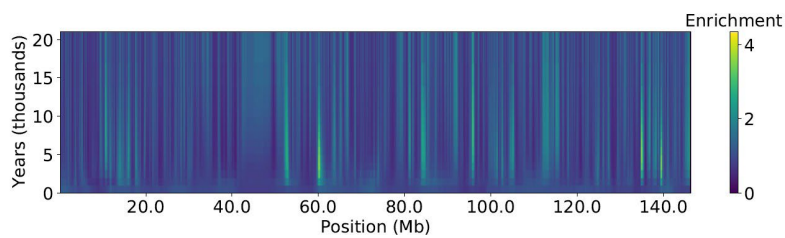
G)

Chromosome 7

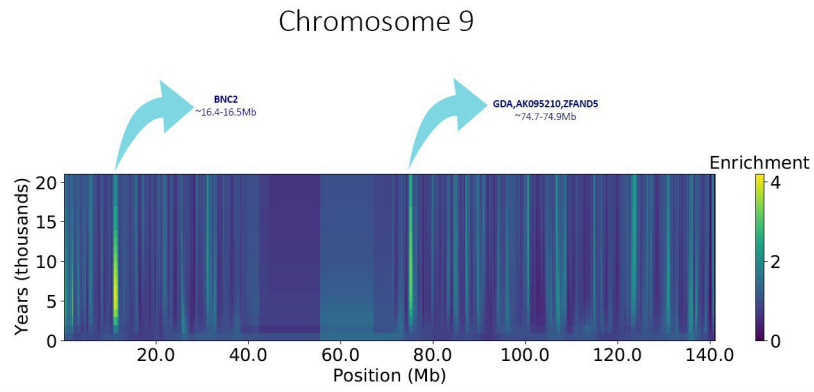


H)

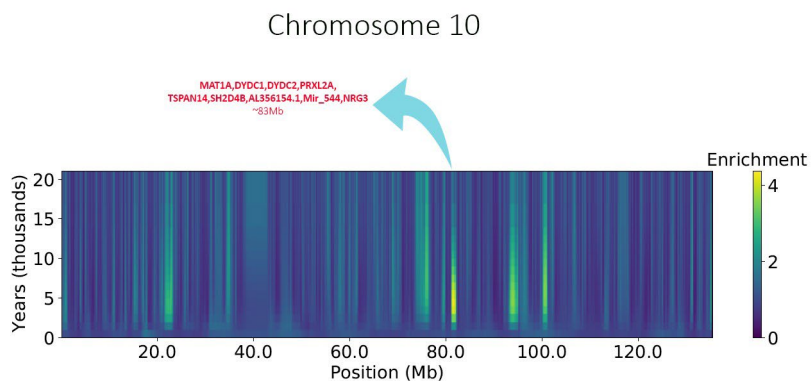
Chromosome 8



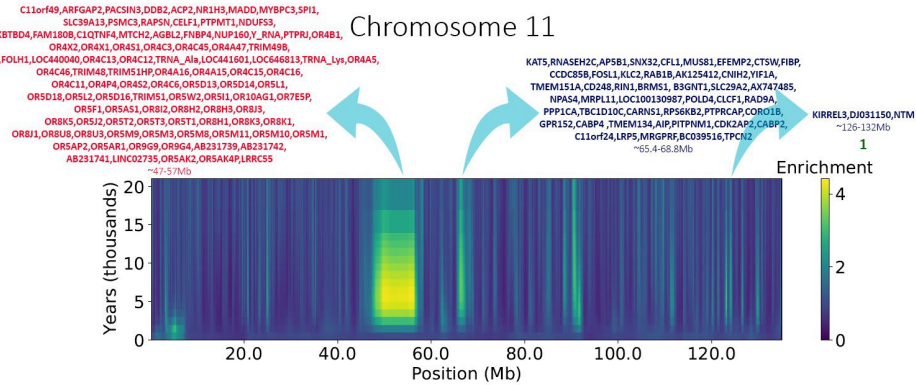
I)



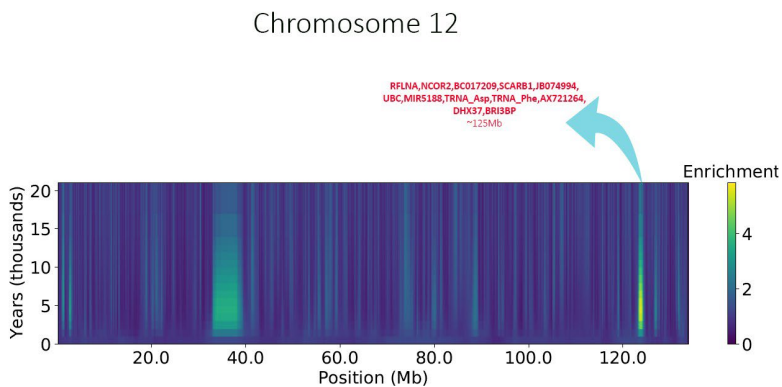
J)



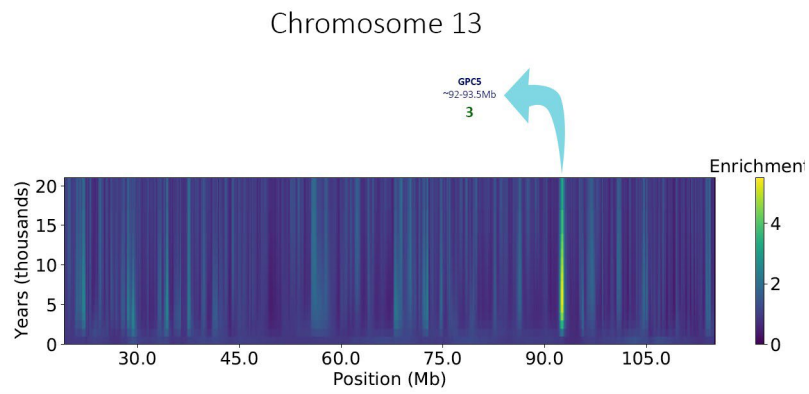
K)



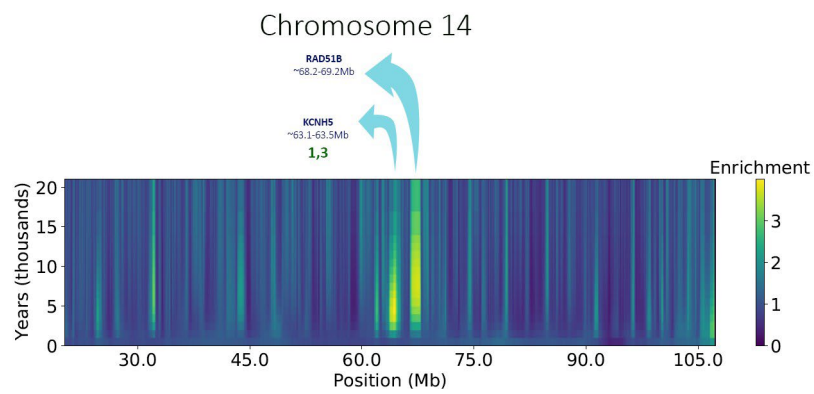
L)



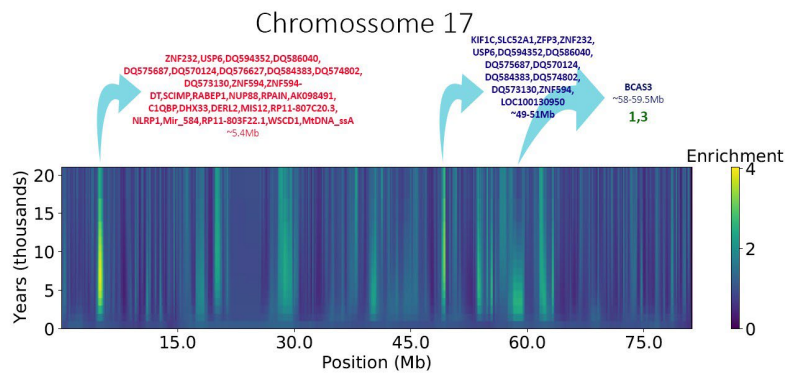
M)



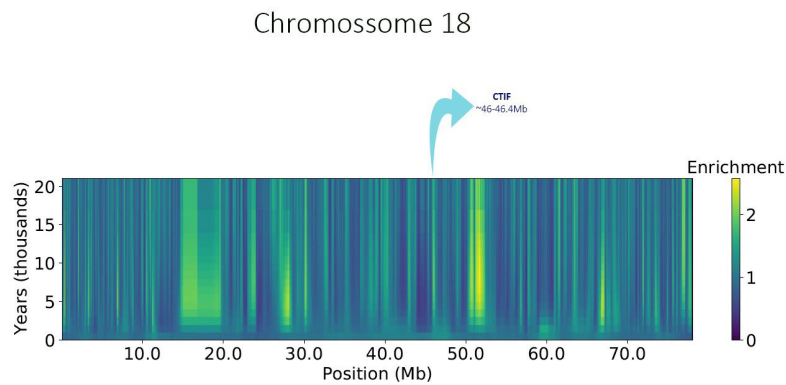
N)



Q)

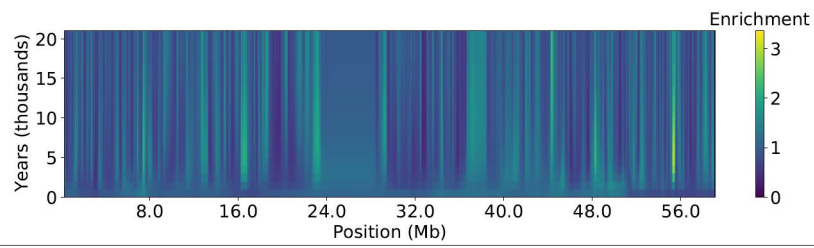


R)



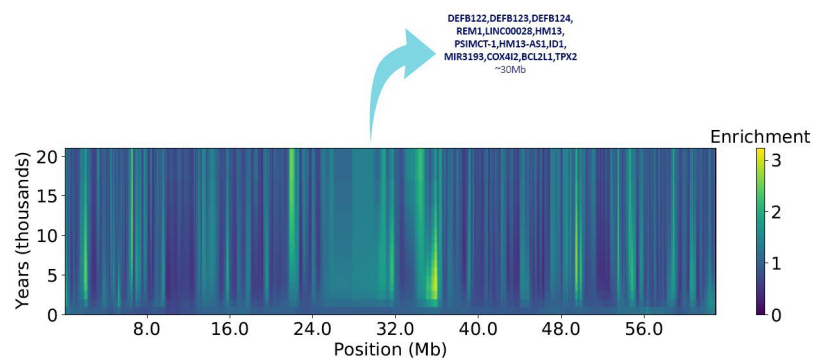
S)

Chromosome 19



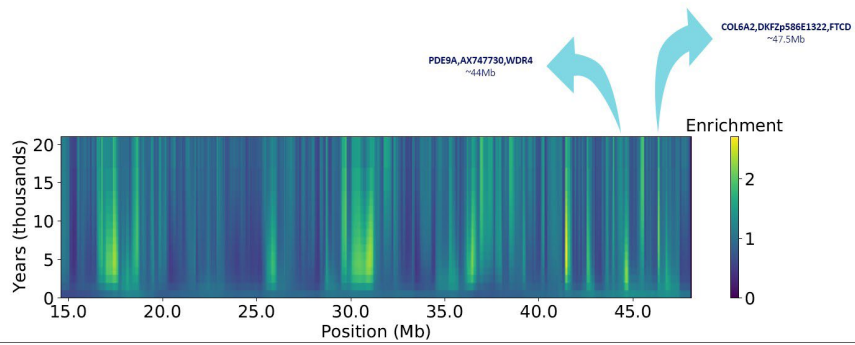
T)

Chromosome 20



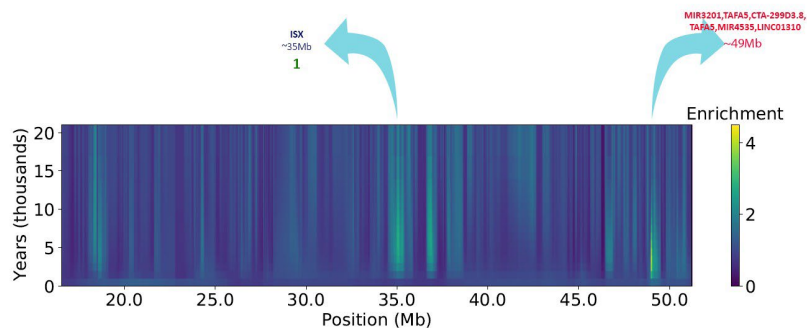
U)

Chromosome 21

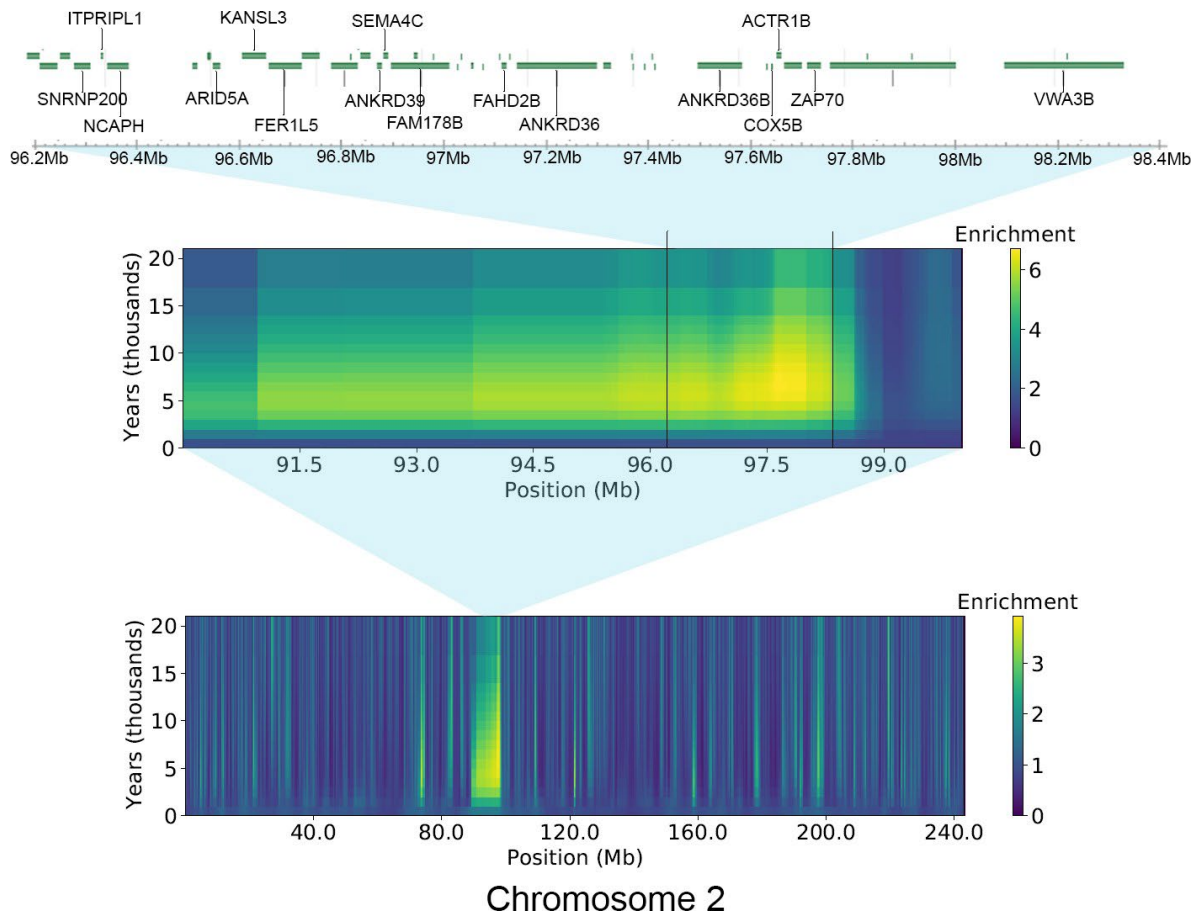


V)

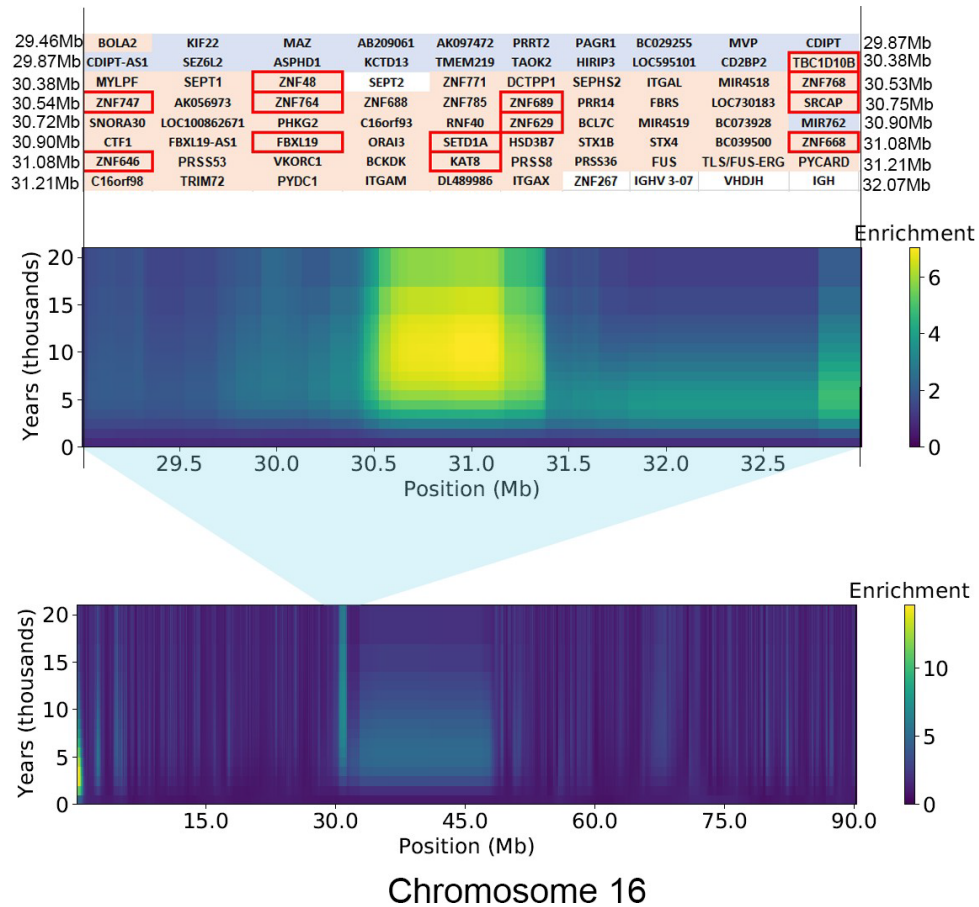
Chromosome 22



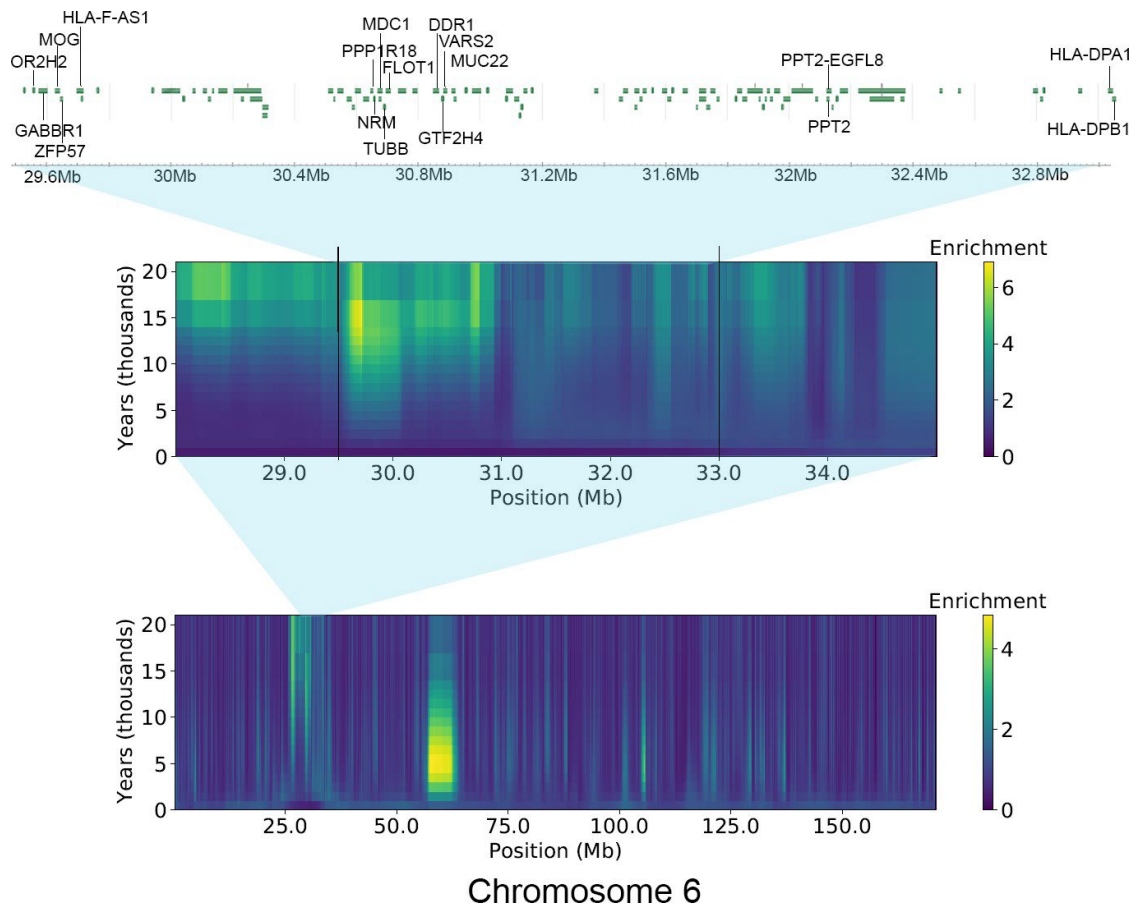
Supplementary Figure 4. ASMC, detailing in blue, regions found as possible signatures of natural selection in our study, the numbers in green indicate other studies where those regions were reported (1: Metsupalu et al. 2011, 2: Suo et al. 2012, 3: Karlsson et al. 2015, 4: Liu et al. 2017, 5: Perdomo-Sabogal et al, 2019). In red we show the genes present in regions with high enrichment but that was not found as an outlier in our study. A) chromosome 1, B) chromosome 2, C) chromosome 3, D) chromosome 4, E) chromosome 5, F) chromosome 6, G) chromosome 7, H) chromosome 8, I) chromosome 9, J) chromosome 10, K) chromosome 11, L) chromosome 12, M) chromosome 13, N) chromosome 14, O) chromosome 15, P) chromosome 16, Q) chromosome 17, R) chromosome 18, S) chromosome 19, T) chromosome 20, U) chromosome 21, V) chromosome 22.



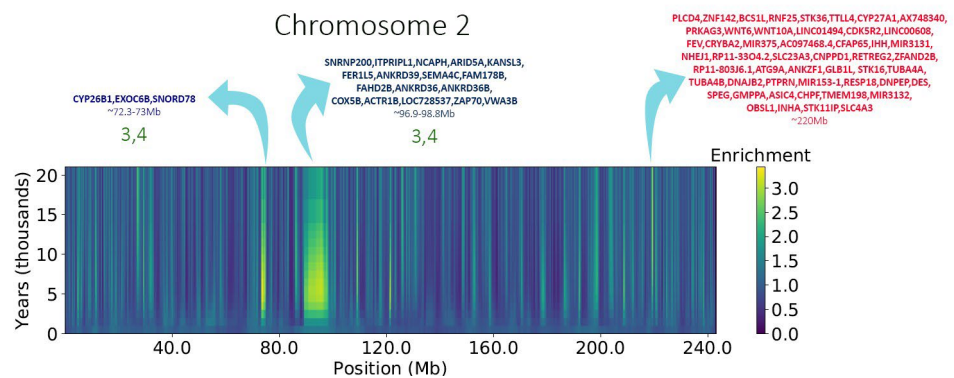
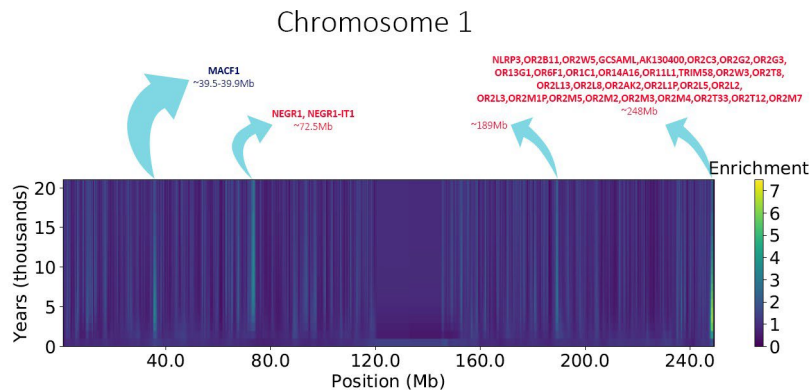
Supplementary Figure 5. ASMC results for chromosome 2, showing a zoom in the region with the biggest enrichment of recent coalescence events (96.2Mb to 98.4Mb), on the top we describe all the genes located within this region.



Supplementary Figure 6. ASMC results for chromosome 16, showing at the top, the genes identified in the region between 29.46Mb and 32.07Mb. In blue we list the genes within the top 0.5% signals and in red the genes within the top 0.1% signals identified for at least one method. The red squares highlight genes that have also been identified in other studies, as detailed in supplementary table 1.



Supplementary Figure 7. ASMC results for chromosome 6, showing a zoom in the region from 29.6Mb to 32.8Mb, on the top we describe the genes located within this region.



Supplementary Figure 8. ASMC results for Chromosome 1 and 2, detailing in blue, regions found as possible signatures of natural selection in our study, the numbers in green indicate other studies where those regions were reported (1: Metsupalu et al. 2011, 2: Suo et al. 2012, 3: Karlsson et al. 2015, 4: Liu et al. 2017, 5: Perdomo-Sabogal et al, 2019). In red we show the genes present in regions with high enrichment but that was not found as an outlier in our study.

Supplementary Table 1: Details of the 435 genes results of our scanning for natural selection signatures. Showing also the shared signals with 1kbp high coverage data and the overlap with other studies.

Supplementary Table 2: Details of the 97 regions results of our scanning for natural selection signatures. Showing also the shared signals with 1kbp high coverage data and the overlap with other studies.

Because of the size and dynamics, those Supplementary tables 1 and 2 are in the following link:

<https://docs.google.com/spreadsheets/d/1J3xtV31hLrswFd2dOU8zgua9FJJCnIsk293IF6YVOvw/edit?usp=sharing>

Chapter 4 - Application of Polygenic Risk Scores (PRS) on admixed Brazilian Populations

Application of Polygenic Risk Scores (PRS) on admixed Brazilian Populations

Marla Mendes^{1,2}, Livia Metzker¹, Camila Zolini¹, Eduardo Tarazona-Santos¹

Affiliations

1 Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil.

2 Department of Anthropology, University of Toronto. Mississauga Campus, 3359 Mississauga Rd, Mississauga, ON L5L 1C6, Canada.

Corresponding Author

Eduardo Tarazona-Santos (edutars@gmail.com)

Keywords

Admixture, PRS, ancestry, genetic diversity, neglected populations.

Abstract

Introduction

Polygenic Risk Scores (PRS) is a calculation that aims to predict disease risk or phenotypes at an individual level, based on the summary statistics obtained with genome-wide association studies (GWAS) (Wray *et al.* 2007). For this, the PRS for a specific phenotype uses the cumulative effect of multiple SNPs, which was determined in several GWAS approaches (Khera *et al.* 2018). Because of its capacity to estimate the probability of an individual developing or not a quantitative genetic characteristic, and in some cases it may even identify clinically actionable levels of risk, the PRS has implications in the future of personalized medicine and has been the target of researcher interest in many scientific groups (Fries 2020).

Despite the potential to become a routine in clinical practices because of the possibility to increase the individual preventive measures, the calculation of PRS today has significant

limitations. Besides all the ethical issues involved, one of the main concerns of the clinical implementation of PRS is a concern surrounding all human genetics in the present day: the lack of representativeness of our human genetic diversity. The effect of variants has been largely calculated based on individuals predominantly of European ancestry (Popejoy and Fullerton 2016), and when we take into account the genetic mosaic that is the American continent populations, for example, the accuracy of Polygenic Risk Scores can be doubted (Bitarello and Mathieson 2020, Marnetto *et al.* 2020, Mostafavi *et al.* 2020, Wang *et al.* 2020, Lewis and Vassos 2020, Chande *et al.* 2020).

Theoretically, the biological reasons that influence the decrease of PRS prediction accuracy in non-European populations include differences in linkage disequilibrium (LD), differences in data collection, phenotype definition, allele frequencies across populations, differences in causal or marginal effect sizes and their estimators using different regression models, rare polymorphisms, and epistatic or gene-environment interactions (Novembre and Barton 2018). But these theoretical constations still have a gap of empirical observations in real data, in the sense of actually testing the effect of these variables in decreasing the predictive power of PRS in non-European data (Bitarello and Mathieson 2020).

Given the fact that admixed populations constitute a significant part of our global community, not just today, but also, the growth of heterogeneous populations seems to be inevitable in the future (Hall *et al.* 2016), it is vital to include these groups in the progress of personalized medicine. In this context, here we apply several PRS statistical tests in Brazilian populations, which constitutes a classical model for population genetics studies on admixture. For this, we used cohorts that are part of the EPIGEN Brazil Initiative (Kehdy *et al.* 2015). One of the main features of this data is that each of the three cohorts has a specific percentage of European ancestry, so the cohort of Salvador has 42.9%, the cohort of Bambuí 78.5% and the cohort of Pelotas 76.1%. Thus, we can measure how this difference in European ancestry affects the accuracy of the predictive power of the PRS.

Materials and Methods

Sampling

Base Data

GIANT Consortium

To quantify the influence of the calculation of PRS in admixed populations using effect size of variants calculated in the European population we used the summary statistics of The Genetic Investigation of ANthropometric Traits (GIANT) consortium. Specifically, we used

Meta-analysis of genome-wide association studies for body mass index in ~700,000 individuals of European ancestry (Yengo *et al.* 2018), which includes ~450,000 UK Biobank participants of United Kingdom (UK) ancestry plus ~250,000 European adults participants from a large GWAS study of body mass index (Locke *et al.* 2015). This last one, includes samples from a Cohort of London-based civil servants (Marmot and Brunner 2005), British individuals (The Wellcome Trust Case Control Consortium 2007), and white European descent (Biosocial Surveys 2008). This data has information about ~2,336,270 SNPs in the build hg19, including the marginal SNP effect size (Beta), the P-Value measuring the significance of the marginal effect, the frequency of the tested allele, the standard error of the effect size (SE), and the sample size.

The database from the GIANT consortium used on this study is publicly available on:

https://portals.broadinstitute.org/collaboration/giant/images/c/c8/Meta-analysis_Locke_et_al_%2BUKBiobank_2018_UPDATED.txt.gz .

PAGE Study

To compare how the precision of the PRS calculation changes when we change a base data in a relatively homogeneous European population by a base data from a heterogeneous group, we used the data provided by Wojcik *et al.* (2019). This data corresponds to The Population Architecture through Genomics and Environment (PAGE) study and includes the effect size of variants involved in the Body Mass Index calculated using an US multi-ethnic group that includes 17,127 African American individuals, 21,955 US-Hispanic/Latino individuals, 4,647 Asian ancestry individuals, 3,936 Native Hawaiian ancestry individuals, and 645 Native American ancestry individuals. The information about the association of ~34,605,623 SNPs (genotyped and imputed) and Body Mass Index and other traits are reported in this study that also shows the importance of diverse, multi-ethnic participants in large-scale genomic studies. This data also has the necessary information for PRS calculation, like the allele effect and frequency, the effect estimate (Beta), the P-Value, and also the information score from IMPUTE2 (Howie *et al.* 2009) of imputation certainty, that we use as a filter during the Quality Control process.

The database from PAGE study used here can be found publicly available on:

http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST008001-GCST009000/GCST008025/ .

Target Data

EPIGEN-Brazil Initiative

Contrasting with the majority of genetic studies that are made in relatively homogeneous populations, the EPIGEN-Brazil Initiative aims to study the association between genetic variants found in complex diseases and the Brazilian population, whose main characteristic is its high level of admixture. This data includes ~2.5 millions SNPs and 6,487 individuals from three Brazilian cohorts: Salvador (n=1309), Bambuí (n=1442) and Pelotas (n=3736); which presents different percentages of the three main ancestries in Brazil (European, African and Native American) (Kehdy *et al.* 2015).

The specific ancestry percentage of each cohort is i. Salvador: 50.8% African ancestry, 42.9% European ancestry, and 6.4% Native American ancestry. ii. Bambuí: 78.5% of European, 14.7% of African, and 6.7% of Native American ancestries. iii. Pelotas: 76.1% European, 15.9% African, and 8% Native American (Figure1). Those differences in the ancestry components made EPIGEN-Brazil a data where we can test the prediction power of PRS.

It is important to emphasize that despite the significance of this study, the genetic diversity in Brazil can be even more diversified, mainly in urban centers, such as Rio de Janeiro or São Paulo, that received different migrations from Asia, as Japanese that form in Brazil the greatest community outside Japan.



Figure 1. Continental admixture of the EPIGEN Brazil populations, adapted from Kehdy *et al.* (2015): Brazilian regions, the studied populations, and their continental individual ancestry bar plots. N represents the numbers of EPIGEN individuals in the Original Dataset.

Quality Control

Base Data

For the quality control process, we follow the instructions of Choi *et al.* (2020). So we choose the BMI as our target trait after checking that the heritability is greater than 20% (Locke *et al.* 2015), and that all summary statistics downloaded by us have the necessary information for the PRS calculation (Effect allele size, P-value, Effective size, frequency of the tested allele, the standard error of the effect size (SE), and the sample size). We also selected just base data in the build hg19, which is the original reference for our target data. With awk commands, we excluded the information about SNPs with $MAF < 0.01$ and $INFO\text{-score} < 0.8$, we also change the SNP ids for “chr:position” to facilitate the posterior merge with the target data. To avoid possible sex bias, for all those tests we chose to analyze just the autossomic information.

Target Data (EPIGEN-Brazil)

For the basic steps in a cleaning data process, we use the automated script in python MosaiQC.py (<https://github.com/ldgh/Smart-cleaning-public>), which includes removal of chr 0, duplicate data, missing data, ambiguous variants; the annotation of variants for dbSNP ID and, if necessary, the LiftOver process. The relatedness was removed using the software NAToRA (https://github.com/ldgh/NAToRA_Public) after the calculation of kinship coefficients on REAP (Thornton *et al.* 2012). Using plink we also filter the data with $geno > 0.99$, $mind < 0.02$, $HWE P > 1 \times 10^{-6}$, $MAF > 0.01$.

PRS calculation and p-value cutoff

To merge the base and the target data we use a script *in-house* on R (MergingBaseTarget.R) that takes into account the following scenarios: (i) perfect allele matches with the same strand, (ii) perfect allele matches with opposite strands, (iii) switched allele 1 and allele 2 calls with the same strand and (iv) switched allele 1 and allele 2 calls with the opposite strand.

Clumping and Thresholding

One of the most used methods for PRS calculation is clumping plus thresholding (C+T), so before calculating PRS, the variants are first clumped, and variants that are weakly correlated (r^2) or not correlated with one another are retained. This step prunes redundant correlated effects caused by linkage disequilibrium (LD) between variants in the target dataset. We make the clumping on the process using PLINK with the following parameters

--clump-p1 1 (significance threshold of the index SNP), --clump-p2 1 (significance threshold of the tag/clumped SNP), --clump-r2 0.10 (LD threshold for clumping) --clump-kb 250 (threshold of physical distance for clumping in kb).

Thresholding on the base data will remove variants with a p-value higher than a chosen level of significance (Choi *et al.* 2020).

We format the input file for generating polygenic risk scores with another script *in-house* on R (WriteInput_for_RangeOfpvalueThresholds.R). The polygenic risk scores were generated using a range of p-value thresholds in plink (--score --q-score-range), with the values of 0.001, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 and 1.0. We then will have a file for each p-value threshold with the following information: Family ID, Individual ID, phenotype, number of SNPs used for scoring, number of named (effect) alleles for each subject, and the polygenic risk score for each subject. So, we could plot the prediction R-squared values vs the p-value

thresholds, and select the best value of p-value thresholds for PRS calculation in each case, 1.GIANT X Bambuí, 2.GIANT x Pelotas, 3.GIANT x Salvador, 4. PAGE X Bambuí, 5.PAGE x Pelotas, 6. PAGE x Salvador (Figure 2).

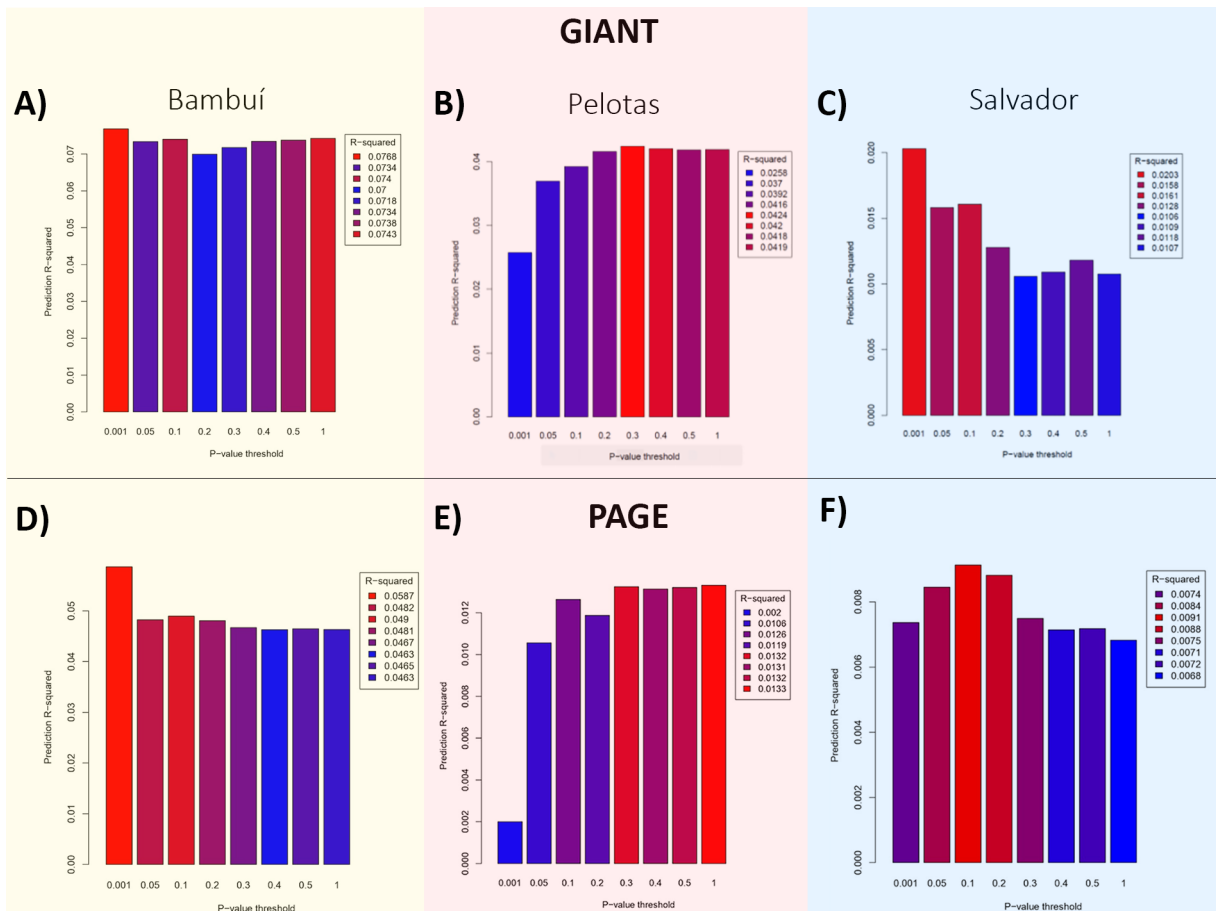


Figure 2. Barplot of prediction R-squared values vs p-value thresholds for each test. Colour gradient from blue for lowest R-squared value to red for highest R-squared value. The legend shows the R-squared values in the order of the p-value thresholds. A) GIANT x Bambuí. B) GIANT x Pelotas C) GIANT x Salvador D) PAGE x Bambuí E) PAGE x Pelotas F) PAGE x Salvador.

Statistical tests

We used statistical tests to measure the correlation between the ancestry of the base data (European) and the PRS calculation for the Body-Mass Index (BMI) trait in the Brazilian people. In this sense, the coefficients by Pearson (Pearson 1903), Kendall (Kendall 1938) and Spearman (Spearman 1904) have been used.

Pearson's correlation was chosen to measure if the variables BMI and PRS were directly proportional and generate a rate to show that. It's a parametric test, If the result is positive and closed to 1, the correlation is directly and most proportional. If the result is negative and closed to -1, the factors are opposites, and if the value is zero the variables are not correlated. To calculate the Pearson's correlation, the following formula was used:

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Where:

r_{xy} = Pearson r correlation coefficient between x and y. n = number of observations.

x_i = value of x (for ith observation)

y_i = value of y (for ith observation)

Kendall's correlation is associated with the interdependence between the variables. It's a non-parametric test and is more used than Spearman's when the base data is not linear. Kendall's correlation measures the power of the dependence on PRS associated with BMI in our tests. If the result is positive and closed to 1, the correlation shows a big dependence between the variables, while a result closed to zero indicated the absence of correlation between variables. If the result is negative and closed to -1, the factors are inversely related (Kendall 1938). To calculate Kendall's correlation, the following formula was used:

$$\tau = \frac{n_c - n_d}{\frac{1}{2} n(n-1)}$$

Where:

τ = tau (Kendall correlation coefficient) n = sample size

n_c = number of concordant

n_d = Number of discordant

Spearman's correlation is a non-parametric test and measures the association between the ordinal variables. If the result is positive and closed to 1, the correlation shows an association between the variables. If the result is negative and closed to -1, the factors are unrelated (Spearman 1904). To calculate the Spearman's correlation, the following formula was used:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d^2}{n(n^2 - 1)}$$

Where:

ρ = rho (Spearman rank correlation)

d = the difference between the ranks of corresponding variables
 n = number of observations

To calculate those correlations, we used a script *in-house* on R.

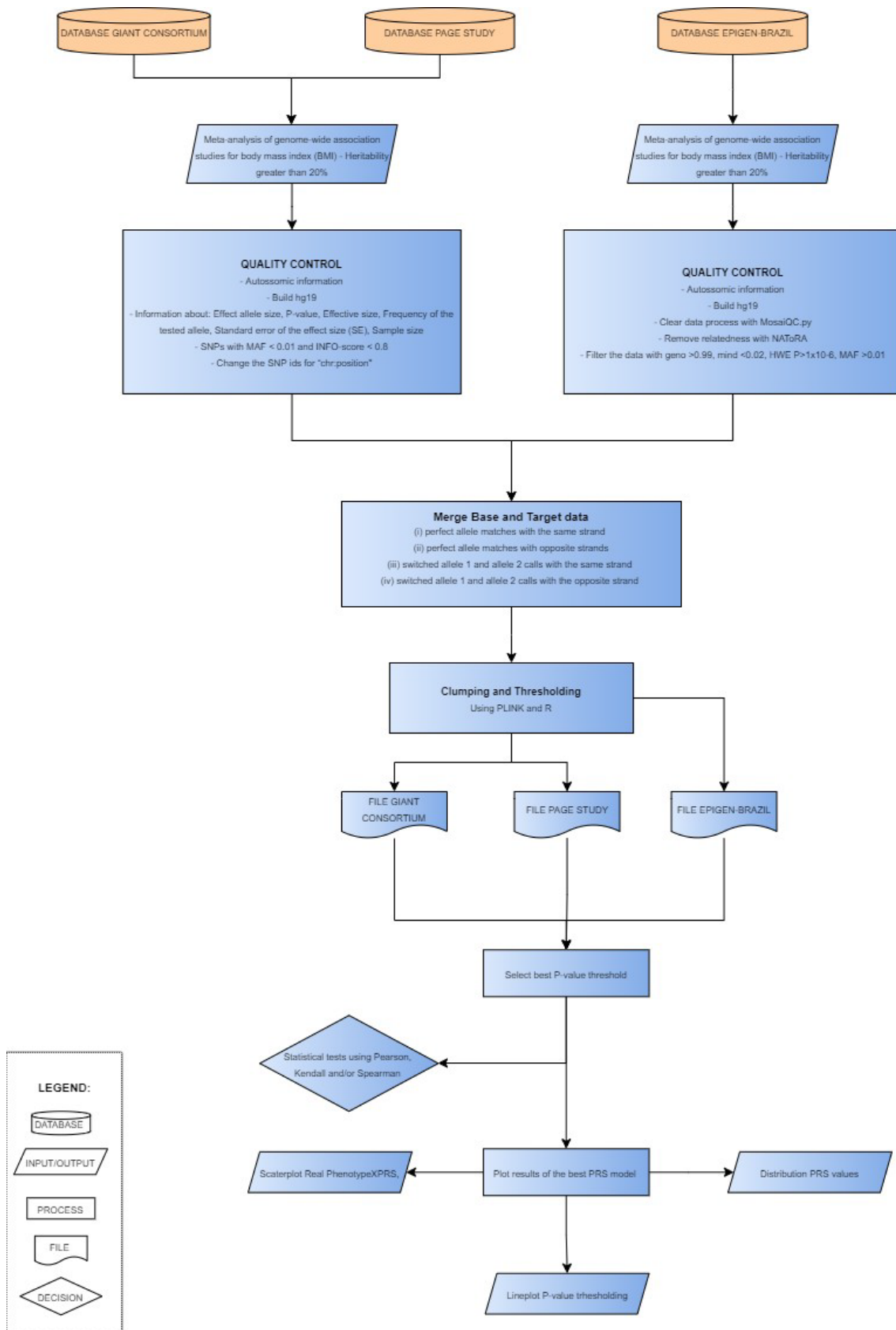


Figure 3: Workflow of our polygenic risk score pipeline.

Results

To measure the influence of ancestry on the accuracy of PRS calculation, we applied it on Brazilian populations that have different percentages of components from the three main parental populations in the country, European, African and Native American. For this we use two base data, one made just in the European population (GIANT consortium), and the other made in trans-ethnic samples (PAGE Study). So, we will have six tests:

- 1) GIANT x Bambuí: PRS calculated with base data 100% European (GIANTconsortium) vs target data 78.5% European (Bambuí).
- 2) GIANT x Pelotas: PRS calculated with base data 100% European (GIANTconsortium) vs target data 76.1% European (Pelotas).
- 3) GIANT x Salvador: PRS calculated with base data 100% European (GIANTconsortium) vs target data 42.9% European (Salvador).
- 4) PAGE x Bambuí: PRS calculated with a trans-ethnic base data (PAGE Study) vs target data 78.5% European (Bambuí).
- 5) PAGE x Pelotas: PRS calculated with a trans-ethnic base data (PAGE Study) vs target data 76.1% European (Pelotas).
- 6) PAGE x Salvador: PRS calculated with a trans-ethnic base data (PAGE Study) vs target data 42.9% European (Salvador).

We then plotted histograms of the distribution frequency of PRS for each group (Supplementary Figure 1) and scatterplots showing the relationship between the observed BMI and the inferred PRS for each test, colouring the individuals by their European ancestry percentage from blue to red, respecting the best p-value threshold (Figure 1). All histograms (Supplementary Figure 1) show a normal pattern different among them mainly in the occupied area on X axis, with some of them with most of the values concentrated close to the value of zero (Supplementary Figure 1 - E), this pattern is also found in Figure 4 - H.

In the scatter plots in Figure 4, we could check that the relation between calculated PRS and real BMI will always be directly proportional unless we join the different cohorts into a single group. In the plot with all groups together we will have the greater PRS values for the less European individuals for the GIANT data, and for PAGE with this same target group, we will have most of the results around zero.

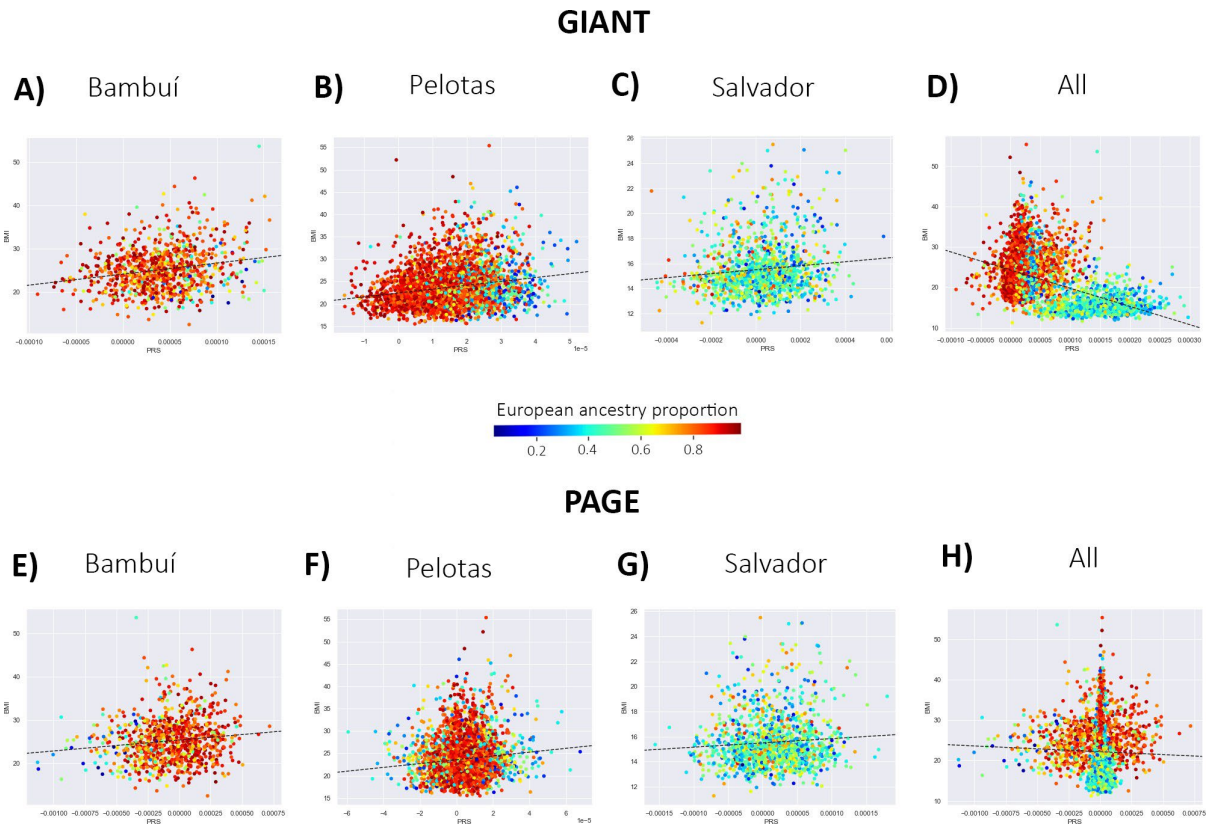


Figure 4. Scatterplot between observed Body-Mass Index (BMI) in the X-axis vs inferred PRS (Y-axis), for each test with the European ancestry proportion showed by the gradient colour of the points. A) GIANT x Bambuí, B) GIANT x Pelotas, C) GIANT x Salvador, D) Giant x All, E) PAGE x Bambuí, F) PAGE x Pelotas, G) PAGE x Salvador, H) PAGE x All.

In the final PRS values, we applied three different correlation tests, Pearson, Kendall and Spearman, and compared them to see if the correlation is stronger when the greater is the correspondence between the ancestry in Base and target data (Table 1). In addition, we merge the population data to see how the elements behave together and we have obtained the worst correlation, including negative values (Table 1). The highest correlation value was obtained for Bambuí, with a value of 0.207 for Pearson correlation using the base data 100%European (GIANT), while the lowest correlation value was -0.0289 for the Pearson between the PAGE base data and the All Groups Cohort.

X		BAMBUÍ	PELOTAS	SALVADOR	ALL GROUPS
GIANT	Pearson	0.2072633	0.2043198	0.115779	0.04245133
	Kendall	0.1129542	0.1324258	0.07530684	0.07040623
	Spearman	0.1686397	0.1974555	0.1117767	0.08609802
PAGE	Pearson	0.1320147	0.1131346	0.07854995	-0.02891785
	Kendall	0.08916491	0.07460167	0.04779395	-0.01643453
	Spearman	0.1344934	0.1116115	0.07134883	-0.02807944

Table 1. Correlation results for each analyzed test. On the cells in red and green, we have the lowest values and highest values respectively.

Discussion

Given the constant advancement of precision medicine technologies, the huge sampling gap between European populations and commonly neglected populations, such as Asians, Africans, Native Americans and admixed populations, is worrying. From the 2,166 polygenic risk scores available on the PGS catalogue (<https://www.pgscatalog.org/>), 91 (4.2%) uses GWAS data that includes Hispanic or Latin American samples, and just 3 included those populations in the score development, besides 50 that uses those populations for the PGS Evaluation. To work to improve this scenario, we use data from ~2.5 million SNPs from the largest initiative in population genomics from Latin America, which covers three different Brazilian populations that vary in the degree of their ancestry from the three main parental populations in Brazil, Europeans, Africans and Native Americans. What makes this data a good model to test the influence of ancestry and admixture in the PRS calculation.

We have shown in real data with different tests the existence of a pattern between the accuracy of the PRS score and the correspondence between the ancestry of target and base data. Overall all our PRS calculation has a positive correlation with the real BMI, and when we tested the PRS calculation for our three cohorts, for both base data, as expected the best

results always will be for the cohorts with greater European ancestry, Bambuí and Pelotas, while the worst correlations will be always for Salvador with just ~48% of European ancestry. But when we join all cohorts the correlation will be even worst, being inversely proportional for all results that used PAGE as the base data. One possible explanation is the heterogeneity of this new group, showing the necessity of covariants in the PRS formula, such as age and PCA components. In this case, is not just the ancestry that varies on these cohorts, but also the diets, environment, socioeconomic conditions, besides the average age for each cohort, Salvador for example, had predominant children and teenagers what makes a big difference in BMI values.

For the comparison between different base data, GIANT and PAGE, we saw that the correlation, in general, is worst for the PAGE consortium. This was not what we expected at first, because PAGE is a multiethnic cohort, different from GIANT that is almost exclusively European. A possible problematic characteristic of PAGE data is that a massive sampling could mask some information, and be confusing when we are testing a specific variable such as ancestry. But further tests need to be done to know exactly why this data is producing those low correlations values.

To deeply explore the relationship between the PRS calculation and the ancestry we made a scatter plot colouring each sample by their percentage of European ancestry, and the results from GIANT and PAGE have demonstrated different patterns (Figure 1). First for all of them we have a positive correlation except when we analyze the All Groups, which presented inversely proportional patterns for both GIANT and PAGE. And when we pay attention to Figure 1-D we see an intriguing pattern to have the greater PRS values for the less European individuals (the bluest samples). Curtis (2018) have found a similar pattern when calculating the PRS for schizophrenia in different populations from the 1000 Genome Project. In this study, Curtis found that the PRS calculated for the African population was ten times bigger than the PRS for European populations. One possible explanation cited by the author is the genetic susceptibility to schizophrenia and that the contributing alleles are suffering different effects of natural selection in Africans and Europeans. This would result in SNPs associated with schizophrenia being at different frequencies in Europeans and Africans, leading to a difference in PRS calculation. A good example that this could happen with BMI too is a work from Scliar *et al.* (2021) that shows an allele on SNP rs114066381, mapped in a potential regulatory region, that is significantly associated with BMI in females. This variant is rare in Europeans but with frequencies of ~3% in West Africa. This difference

will have an impact on the PRS calculated in Africans and in Europeans, which we need to explore in more detail. And actually, when we tried to test the correlations between the calculated PRS and different ancestry, we really found a bigger correlation for the African ancestry while the worst, and almost always inversely proportional correlation will be for the European ancestry (Supplementary Figure 4). When we check the PAGE results, the biggest difference is that when we focus on the plot with all groups together we have a big peak in the PRS value of zero. This is another thing that needs to be explored in the next steps of this work, and although all limitations this work serves to illustrate that ancestry impacts on PRS in a complex way that will need to be systematically studied.

Works Cited

“Biosocial Surveys.” 2007, doi:10.17226/11939.

Bitarello, Bárbara D., and Iain Mathieson. “Polygenic Scores for Height in Admixed Populations.” 2020, doi:10.1101/2020.04.08.030361.

Chande, Aroon T, *et al.* “The Phenotypic Consequences of Genetic Divergence between Admixed Latin American Populations: Antioquia and Chocó, Colombia.” *Genome Biology and Evolution*, vol. 12, no. 9, 2020, pp. 1516–1527., doi:10.1093/gbe/evaa154.

Choi, Shing Wan, *et al.* “Tutorial: a Guide to Performing Polygenic Risk Score Analyses.” *Nature Protocols*, vol. 15, no. 9, 2020, pp. 2759–2772., doi:10.1038/s41596-020-0353-1.

Curtis, David. “Polygenic Risk Score for Schizophrenia Is More Strongly Associated with Ancestry than with Schizophrenia.” 2018, doi:10.1101/287136.

Fries, Gabriel R. “Polygenic Risk Scores and Their Potential Clinical Use in Psychiatry: Are We There Yet?” *Brazilian Journal of Psychiatry*, vol. 42, no. 5, 2020, pp. 459–460., doi:10.1590/1516-4446-2020-0865.

“Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls.” *Nature*, vol. 447, no. 7145, 2007, pp. 661–678., doi:10.1038/nature05911.

- Hall, Matthew, *et al.* “Trajectories of Ethnoracial Diversity in American Communities, 1980-2010.” *Population and Development Review*, vol. 42, no. 2, 2016, pp. 271–297., doi:10.1111/j.1728-4457.2016.00125.x.
- Howie, Bryan N., *et al.* “A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies.” *PLoS Genetics*, vol. 5, no. 6, 2009, doi:10.1371/journal.pgen.1000529.
- “I. Mathematical Contributions to the Theory of Evolution. —XI. On the Influence of Natural Selection on the Variability and Correlation of Organs.” *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 200, no. 321-330, 1903, pp. 1–66., doi:10.1098/rsta.1903.0001.
- Kehdy, Fernanda S. G., *et al.* “Origin and Dynamics of Admixture in Brazilians and Its Effect on the Pattern of Deleterious Mutations.” *Proceedings of the National Academy of Sciences*, vol. 112, no. 28, 2015, pp. 8696–8701., doi:10.1073/pnas.1504447112.
- Kendall, M. G. “A New Measure Of Rank Correlation.” *Biometrika*, vol. 30, no. 1-2, 1938, pp. 81–93., doi:10.1093/biomet/30.1-2.81.
- Khera, Amit V., *et al.* “Genome-Wide Polygenic Scores for Common Diseases Identify Individuals with Risk Equivalent to Monogenic Mutations.” *Nature Genetics*, vol. 50, no. 9, 2018, pp. 1219–1224., doi:10.1038/s41588-018-0183-z.
- Lewis, Cathryn M., and Evangelos Vassos. “Polygenic Risk Scores: from Research Tools to Clinical Instruments.” *Genome Medicine*, vol. 12, no. 1, 2020, doi:10.1186/s13073-020-00742-5.
- Locke, Adam E., *et al.* “Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology.” *Nature*, vol. 518, no. 7538, 2015, pp. 197–206., doi:10.1038/nature14177.
- Locke, Adam E., *et al.* “Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology.” *Nature*, vol. 518, no. 7538, 2015, pp. 197–206., doi:10.1038/nature14177.

- Marmot, Michael, and Eric Brunner. "Cohort Profile: The Whitehall II Study." *International Journal of Epidemiology*, vol. 34, no. 2, 2005, pp. 251–256., doi:10.1093/ije/dyh372.
- Marnetto, Davide, *et al.* "Ancestry Deconvolution and Partial Polygenic Score Can Improve Susceptibility Predictions in Recently Admixed Individuals." *Nature Communications*, vol. 11, no. 1, 2020, doi:10.1038/s41467-020-15464-w.
- Mostafavi, Hakhamanesh, *et al.* "Variable Prediction Accuracy of Polygenic Scores within an Ancestry Group." *ELife*, vol. 9, 2020, doi:10.7554/elifelife.48376.
- Novembre, John, and Nicholas H Barton. "Tread Lightly Interpreting Polygenic Tests of Selection." *Genetics*, vol. 208, no. 4, 2018, pp. 1351–1355., doi:10.1534/genetics.118.300786.
- Popejoy, Alice B., and Stephanie M. Fullerton. "Genomics Is Failing on Diversity." *Nature*, vol. 538, no. 7624, 2016, pp. 161–164., doi:10.1038/538161a.
- Salkind, Neil J., and Kristin Rasmussen. *Encyclopedia of Measurement and Statistics*. SAGE Publications, 2007.
- Scliar, Marilia O., *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." *International Journal of Obesity*, vol. 45, no. 5, 2021, pp. 1017–1029., doi:10.1038/s41366-021-00761-1.
- Thornton, Timothy, *et al.* "Estimating Kinship in Admixed Populations." *The American Journal of Human Genetics*, vol. 91, no. 1, 2012, pp. 122–138., doi:10.1016/j.ajhg.2012.05.024.
- Wang, Ying, *et al.* "Theoretical and Empirical Quantification of the Accuracy of Polygenic Scores in Ancestry Divergent Populations." *Nature Communications*, vol. 11, no. 1, 2020, doi:10.1038/s41467-020-17719-y.
- Wojcik, Genevieve L., *et al.* "Genetic Analyses of Diverse Populations Improves Discovery for Complex Traits." *Nature*, vol. 570, no. 7762, 2019, pp. 514–518.,

doi:10.1038/s41586-019-1310-4.

Wray, Naomi R., *et al.* “Prediction of Individual Genetic Risk to Disease from Genome-Wide Association Studies.” *Genome Research*, vol. 17, no. 10, 2007, pp. 1520–1528.,

doi:10.1101/gr.6665407.

Yengo, Loic, *et al.* “Meta-Analysis of Genome-Wide Association Studies for Height and Body Mass Index in ~700000 Individuals of European Ancestry.” *Human Molecular Genetics*, vol. 27, no. 20, 2018, pp. 3641–3649., doi:10.1093/hmg/ddy271.

Supplementary Material

Application of Polygenic Risk Scores (PRS) on admixed Brazilian Populations

Marla Mendes^{1,2}, Livia Metzker¹, Camila Zolini¹, Eduardo Tarazona-Santos¹

Affiliations

1 Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil.

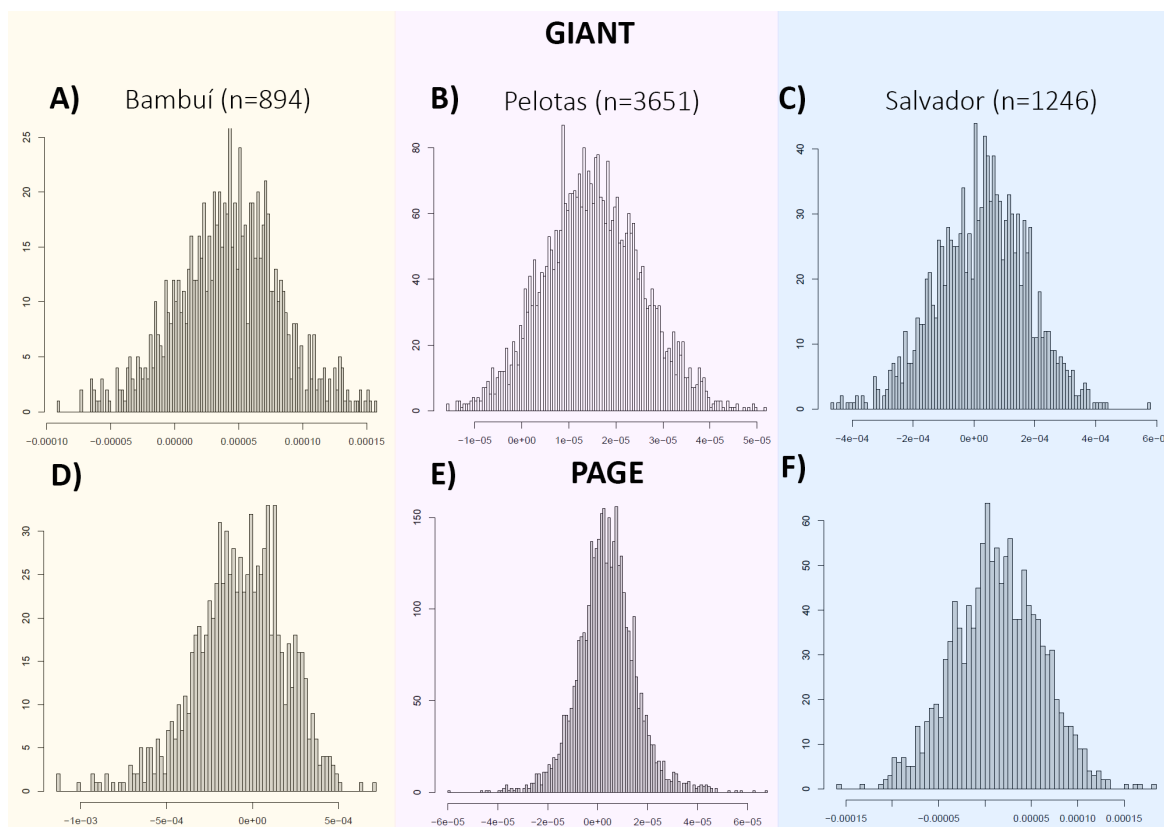
2 Department of Anthropology, University of Toronto. Mississauga Campus, 3359 Mississauga Rd, Mississauga, ON L5L 1C6, Canada.

Corresponding Author

Eduardo Tarazona-Santos (edutars@gmail.com)

Keywords

Admixture, PRS, ancestry, genetic diversity, neglected populations.



Supplementary Figure 1. Histograms of the frequency distribution of PRS values for each test. In the X-axis are the PRS values and in the Y-axis are the frequency. A) GIANT x Bambuí. B) GIANT x Pelotas C) GIANT x Salvador D) PAGE x Bambuí E) PAGE x Pelotas F) PAGE x Salvador.

X		Thirds test 1: Bambuí (n=298 in each part)			Thirds test 2: Pelotas (n=1217 in each part)			Thirds test 3: Salvador (n= 415 in each part)			Thirds test 4: all groups (n= in each part)		
		1/3 (1)	1/3 (2)	1/3 (3)	1/3 (1)	1/3 (2)	1/3 (3)	1/3 (1)	1/3 (2)	1/3 (3)	1/3 (1)	1/3 (2)	1/3 (3)
GIANT	Pearson	0.193	0.319	0.129	0.170	0.212	0.238	0.130	0.072	0.147	0.029	0.244	0.194
	Kendall	0.113	0.190	0.061	0.112	0.116	0.164	0.082	0.073	0.087	0.024	0.167	0.146
	Spearman	0.170	0.281	0.088	0.166	0.173	0.243	0.120	0.110	0.129	0.036	0.239	0.217
PAGE	Pearson	0.135	0.142	0.075	0.109	0.112	0.141	0.081	0.078	0.084	-0.135	0.010	0.037
	Kendall	0.109	0.095	0.034	0.079	0.052	0.089	0.057	0.036	0.044	-0.074	0.020	0.064
	Spearman	0.165	0.142	0.052	0.119	0.078	0.133	0.086	0.051	0.068	-0.112	0.029	0.094

Supplementary Table 1. Results of Pearson, Kendall and Spearman correlations compared to basedata from the GIANT Consortium and the PAGE Study. The tests were separated into three according to the percentage of European ancestry that increase from left to right. In Bambuí, each part has 298 individuals, in Pelotas, almost 1217 and 415 from Salvador.

X		PELOTAS+BAMBUÍ	BAMBUÍ+SALVADOR	SALVADOR+PELOTAS
GIANT	Pearson	0.2108575	0.1072303	-0.003244299
	Kendall	0.1383846	0.0762179	0.04686025
	Spearman	0.2054172	0.1092376	0.0493057
PAGE	Pearson	0.03837882	-0.1164324	-0.06509453
	Kendall	0.0568654	-0.06458905	-0.01220306
	Spearman	0.0839852	-0.1015286	-0.02596023

Supplementary Table 2. Correlation results for mixed datas in the analyzed test.

X		Thirds test 1: Pelotas+BambuÍ			Thirds test 2: BambuÍ+Salvador			Thirds test 3: Salvador+Pelotas		
		1/3 (1)	1/3 (2)	1/3 (3)	1/3 (1)	1/3 (2)	1/3 (3)	1/3 (1)	1/3 (2)	1/3 (3)
GIANT	Pearson	0.179	0.271	0.184	0.112	0.151	0.308	0.008	0.132	0.224
	Kendall	0.113	0.158	0.150	0.082	0.123	0.151	0.007	0.133	0.151
	Spearman	0.167	0.236	0.221	0.122	0.186	0.221	0.017	0.182	0.224
PAGE	Pearson	0.046	0.055	0.028	-0.446	-0.298	0.085	-0.045	-0.009	0.127
	Kendall	0.057	0.053	0.063	-0.075	-0.108	0.056	-0.029	0.026	0.077
	Spearman	0.084	0.079	0.093	-0.115	-0.171	0.085	-0.043	0.034	0.115

Supplementary Table 3. Results of Pearson, Kendall and Spearman correlations compared to basedata from the GIANT Consortium and the PAGE Study. The tests were separated into three according to the percentage of European ancestry that increase from left to right. In Pelotas+BambuÍ, each part has 1507 individuals, in Bambuí+Salvador almost 711 each, 1633 for the cohort from Salvador+Pelotas and 1920 for Salvador+Pelotas+BambuÍ.

X		EUR	AFR	NAT
GIANT	Pearson	-0.1210379	0.1218596	-0.001744258
	Kendall	-0.1854996	0.1850716	0.04566857
	Spearman	-0.2660399	0.2658231	0.06942105
PAGE	Pearson	0.04797083	-0.04604617	-0.01111381
	Kendall	-0.01841732	0.01361062	-0.01028001
	Spearman	-0.02978566	0.02497773	-0.01509043

Supplementary Table 4. Results of Pearson, Kendall and Spearman correlations comparing the PRS calculated and different ancestries.

Attachments

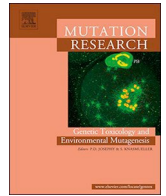
Collaborative Papers



ELSEVIER

Contents lists available at ScienceDirect

Mutat Res Gen Tox En

journal homepage: www.elsevier.com/locate/gentox

The Iberian legacy into a young genetic xeroderma pigmentosum cluster in central Brazil

L.P. Castro^a, M. Sahbatou^b, F.S.G. Kehdy^c, A.A. Farias^{d,e}, A.A. Yurchenko^f, T.A. de Souza^a, R.C.A. Rosa^g, C.T. Mendes-Junior^h, V. Bordaⁱ, V. Munford^a, É.A. Zanardo^j, S.N. Chehimi^j, L.D. Kulikowski^j, M.M. Aquino^k, T.P. Leal^k, E. Tarazona-Santos^k, S.C. Chaibub^l, B. Gener^{m,n}, N. Calmels^o, V. Laugel^o, A. Sarasin^{p,1}, C.F.M. Menck^{a,1,*}

^a Department of Microbiology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil

^b Foundation Jean Dausset – CEPH, Paris, France

^c Leprosy Laboratory, Oswaldo Cruz Institute, Oswaldo Cruz Foundation, Rio de Janeiro, Brazil

^d Human Genome and Stem-Cell Center, Institute of Biosciences, University of São Paulo (USP), São Paulo, Brazil

^e Department of Genetics and Evolutionary Biology, Biosciences Institute, University of São Paulo (USP), São Paulo, Brazil

^f Inserm U981, Gustave Roussy Cancer Campus, Université Paris Saclay, Villejuif, France

^g Department of Genetics, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil

^h Department of Chemistry, Forensic and Genomics Research Laboratory, Faculty of Philosophy, Sciences and Letters, University of São Paulo, Ribeirão Preto, Brazil

ⁱ National Laboratory for Scientific Computation (LNCC), Petropolis, Rio de Janeiro, Brazil

^j Cytogenomics Laboratory, Department of Pathology, School of Medicine, University of São Paulo (FMUSP), São Paulo, Brazil

^k Department of Genetics, Ecology and Evolution, Institute of Biological Sciences, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

^l General Hospital of Goiania, Goiania, Brazil

^m Osakidetza Basque Health Service, Cruces University Hospital, Department of Genetics, Bizkaia, Spain

ⁿ Biocruces Bizkaia Health Research Institute, Bizkaia, Spain

^o Laboratory of Medical Genetics, Institute of Medical Genetics of Alsace (IGMA), Strasbourg, France

^p UMR8200 CNRS, Gustave Roussy Institute, University Paris-Saclay, Villejuif, France

ARTICLE INFO

Keywords:

Xeroderma pigmentosum
Genetic cluster
Founder mutations
POLH(XPV)
DNA repair
Ancestry

ABSTRACT

In central Brazil, in the municipality of Faina (state of Goiás), the small and isolated village of Araras comprises a genetic cluster of xeroderma pigmentosum (XP) patients. The high level of consanguinity and the geographical isolation gave rise to a high frequency of XP patients. Recently, two founder events were identified affecting that community, with two independent mutations at the *POLH* gene, c.764 + 1 G > A (intron 6) and c.907 C > T; p.Arg303* (exon 8). These deleterious mutations lead to the xeroderma pigmentosum variant syndrome (XP-V). Previous reports identified both mutations in other countries: the intron 6 mutation in six patients (four families) from Northern Spain (Basque Country and Cantabria) and the exon 8 mutation in two patients from different families in Europe, one of them from Kosovo. In order to investigate the ancestry of the XP patients and the age for these mutations at Araras, we generated genotyping information for 22 XP-V patients from Brazil (16), Spain (6) and Kosovo (1). The local genomic ancestry and the shared haplotype segments among the patients showed that the intron 6 mutation at Araras is associated with an Iberian genetic legacy. All patients from Goiás, homozygotes for intron 6 mutation, share with the Spanish patients identical-by-descent (IBD) genomic segments comprising the mutation. The entrance date for the Iberian haplotype at the village was calculated to be approximately 200 years old. This result is in agreement with the historical arrival of Iberian individuals at the Goiás state (BR). Patients from Goiás and the three families from Spain share 1.8 cM (family 14), 1.7 cM (family 15), and a more significant segment of 4.7 cM within family 13. On the other hand, the patients carrying the exon 8 mutation do not share any specific genetic segment, indicating an old genetic distance between them or even no common ancestry.

* Corresponding author at: Department of Microbiology, ICB, USP, Av. Prof. Lineu Prestes, 1374, Ed. Biomédicas 2, São Paulo, SP 05508-900, Brazil.

E-mail address: cmmenck@usp.br (C.F.M. Menck).

¹ Both authors contributed equally to this work.

<https://doi.org/10.1016/j.mrgentox.2020.503164>

Received 28 December 2019; Received in revised form 22 February 2020; Accepted 25 February 2020

Available online 29 February 2020

1385-718/© 2020 Elsevier B.V. All rights reserved.

1. Introduction

In Brazil, it is common to find isolated communities with a high-frequency of inbreeding, resulting in genetic clusters [1]. One example is the genetic cluster for xeroderma pigmentosum (XP) in central Brazil, at Araras village, in Faina County, state of Goiás [2]. XP is a rare autosomal recessive disease. Patients develop a high frequency of skin tumors, ophthalmologic abnormalities, and, in some cases, neurological damage. Their skin and ocular problems are caused by exposure to sunlight and ultraviolet radiation. The manifestations depend mainly on the age of their clinical diagnosis and the adherence to photo-protection [3]. The molecular cause for XP is the lack of DNA repair of UV-induced DNA lesions or their error-prone replication. The clinical defects are directly proportional to the amount of life sunlight-exposure, and genetics is a determinant for the prognosis of the disorder. The syndrome can range from mild to severe symptoms according to the mutated gene and the type of mutation [4].

There are seven classic complementation groups: XP-A to XP-G and a Variant form, XP-V. The classical XP forms are caused by a defect in the nucleotide excision repair (NER). While in the XP-V, although NER-proficient, the translesion synthesis, carried out by DNA polymerase η (Pol η), is deficient due to bi-allelic mutations on the *POLH* gene [5]. The XP incidence over the world can be as low as one case per million in the USA, and Europe [6], with approximately 100 patients in the UK [3]. However, founder events in XP genes have been reported among Native Americans, Japanese, Jewish from Iraq, North Africans, Europeans, and Pakistanis [7–17]. Due to these effects, in addition to cultural consanguineous marriage, the frequency could be much higher at specific localities as it is for Japan (1 in 22,000), Tunisia (1 in 10,000), North Africa and the Middle East, where the frequency is 1 in

50,000 [7,8,11,18].

We estimate the existence of approximately 200 XP patients in Brazil, which would represent one case per million inhabitants. However, this frequency is higher at localities with founder effects, such as Araras village (Goiás State), where there is a high frequency of inbreeding. Considering the whole population from the municipality (Faina), 6,983 individuals reported in 2010 by the Institute of Geography and Statistics (IBGE), the XP patient incidence in that locality is one in 410 inhabitants.

From previous work, we identified at Araras, two mutations in the *POLH* gene OMIM #278,750. One at a splicing site in intron 6 rs772570523, c.764 + 1G > A, and the other at exon 8 rs759607901, c.908C > T; p.Arg303* [2]. Both mutations were deleterious for the Pol η activity and gave rise to the XP-V syndrome. These mutations were also described in European patients. The intron 6 mutation was reported in four families in the North of Spain [19]. The exon 8 mutation was identified in two patients, one, homozygote, from Kosovo and the other, a compound heterozygote, from France [5]. Both mutations were also observed in small frequencies (i.e. lower than 0.00005) in the GnomAD and TOPMED databases, but were not reported by the 1000 Genomes Project (phase 3).

The two mutations at *POLH*, at intron 6 and exon 8, were introduced in the community at different times. The intron 6 mutation probably came from Portugal (and in fact, their carriers present Portuguese family names), and was followed by the exon 8 mutation brought in 1965 by a family coming from Nova Fátima, in the municipality of Hidrolândia (still Goiás, 183 miles Southeast from Faina) (Appendix, Fig. A1). The members of this family tell that they were running away from the disease (XP) and the prejudice that was associated with life in the locality [20].

Table 1

Cohort with 22 XP-V patients for SNP array analyses; 15 from Brazil, 6 from Spain, and 1 from Kosovo. The number preceding the sample ID represents the family ID from each individual, in which samples with the same family ID represents full siblings.

Groups	Family_SampleID	Zygoty	Gender	Status	Village/City	State	Country
1	1_XP08GO.br	6/6	M	Alive	Faina	G	Brazil
	1_XP33GO.br	6/6	F	Alive	Faina	O	Brazil
	4_XP110GO.br	6/6	M	Alive	Araras	G	Brazil
	3_XP78GO.br	6/6	M	Alive	Araras	O	Brazil
	6_XP100GO.br	6/6	M	Alive	Araras	G	Brazil
						O	
						O	
						O	
						O	
2	8_XP06GO.br	6/6	M	Alive	Araras	G	Brazil
	7_XP11GO.br	6/6	F	Alive	Araras	O	Brazil
	2_XP25GO.br	6/6	M	Alive	Araras	G	Brazil
	2_XP52GO.br	6/6	M	Alive	Araras	O	Brazil
	5_XP85GO.br	6/6	M	Deceased ⁺	Araras	G	Brazil
					O		
						O	
						O	
						O	
3	12_XP03GO.br	6/8	F	Alive	Araras	G	Brazil
	12_XP04GO.br	6/8	F	Alive	Araras	O	Brazil
	12_XP39GO.br	6/8	M	Deceased ⁺⁺	Araras	G	Brazil
	9_XP56GO.br	6/8	M	Alive	Araras	O	Brazil
	10_XP88GO.br	8/8	M	Alive	Nova Fátima	G	Brazil
					O		
						O	
						O	
						O	
4	11_XP01KO.eu* ¹	8/8	F	Alive	–	–	Kosovo
5	14_XP01ES.eu* ²	6/6	F	Alive	–	Basque	Spain
	15_XP02ES.eu* ³	6/6	F	Alive	–	Basque	Spain
	15_XP03ES.eu* ⁴	6/6	F	Alive	–	Basque	Spain
	15_XP06ES.eu* ⁵	6/6	M	Alive	–	Basque	Spain
	13_XP04ES.eu* ⁶	6/6	M	Alive	–	Basque	Spain

13_XP05ES.eu*⁷

6/6

F

Alive

–

Basque

Spain

*Indicates patient ID, different from the reference article, adding .eu to denote their European origin. *¹ from Opletalova et al. 2014 [5] and *²-*⁷ from Calmels et al. 2016 [19]; *¹XP965V1, *²Patient #11, *³Patient #12, *⁴Affected brother of patient #12, *⁵Affected sister of patient #12, *⁶ Patient #13 and *⁷Affected sister of Patient #13. The zygosity is represented as 6/6 for homozygote for intron 6 mutation, 6/8 compound heterozygote and 8/8 homozygote for the exon 8 mutation. The Gender column is identified as male (M) and female (F). + Year of death 2018 (82 y). ++ Year of Death 2015 (40 y). GO: Goiás State. We described the five groups in the section 2.1.1.

The Brazilian ancestry profile determines the prevalence of genetic disorders, mainly in isolated regions [21–24]. Araras is the first XP genetic cluster described in Brazil. In order to study the global ancestry and the haplotype origin at Araras, we generated genome-wide information from 22 XP-V patients (15 families) from Brazil (Goiás), Spain (Basque Country) and Kosovo. Interestingly, we detected a shared identical-by-descendent (IBD) haplotype within the patients from Goiás and Basque Country. The time for the most recent common ancestor at Araras is approximately 200 years. The community is a young genetic cluster and is closely related to family 13 from Spain. However, no direct genetic relationship was detected for the patients from Goiás and Kosovo carrying the exon 8 mutation.

2. Material and methods

2.1. High-Throughput genotyping (SNP-array)

2.1.1. Samples description

The Infinium Omni-Express24 v1.1 Chip (713k SNPs) was used to perform this study (Illumina®, San Diego, CA, USA). Table 1 describes the genotype information and sample origin from the 22 XP-V individuals (13 males and 9 females), 15 from Goiás, Brazil (11 families), six from Basque Country in Spain (3 families), and one from Kosovo. From Goiás, there are mainly three families. The first family is from Araras village and the second from the city of Faina (30 miles from Araras), both at the municipality of Faina. The third family is from Nova Fátima village, in the municipality of Hidrolândia (Appendix, Fig. A1). For simplicity purposes, we will call patients either from Araras or Faina, as Araras' patients.

We organized the individuals into five groups considering their ancestry profile (Table 1). Groups 1 and 2 represent homozygous individuals mutated on intron 6. **Group 1** comprises the Afro-descendant individuals with black skin, considered black, according to IBGE: three patients from Araras and two from Faina. **Group 2** is composed of patients with white skin, although some of them present some Afro-descendant phenotype: five patients living in Araras. **Group 3** comprises five patients who carry the exon 8 mutation: only one is homozygote for this mutation, born in Hidrolândia and living in Araras, and four of his nephews and nieces, born in Araras, compound heterozygotes, also carrying the intron 6 mutation. **Groups 4 and 5** represent the European patients: one from Kosovo, homozygote for exon 8 mutation (Group 4) and six from Spain (Basque Country, Group 5), homozygote for intron 6 mutation. All saliva and blood samples were collected with written informed consent.

2.1.2. Quality control

We applied quality control (QC) before processing the data. In the QC filter, we considered the autosome variants with a minor allelic frequency greater than 1%, which were checked for the absence of Mendelian errors and Hardy-Weinberg equilibrium (HWE). We had no Mendelian errors and no SNPs with HWE = $10e-20$. SNPs with high missing call rates (>5%) were removed. The cleaning process also checked for strand's problems and position's conflicts. From the 636,208 genotyped SNPs, we had 592,705 variants passing filter with a call rate of 99.7 %.

2.1.3. Data processing

To infer the relationship (up to 4th-degree), we used KING software v2.2.4 [25]. From Fsuite software v1.0.4, we estimated the inbreeding coefficient (F-Median) and consanguinity level from each individual [26,27]. The software creates 100 random submaps with SNPs in minimal linkage disequilibrium (LD) and calculates probabilities to be inbred or not [27].

To study population structure, ADMIXTURE methods [28] and PCA were applied [29]. For ADMIXTURE, we used the data from HGDP-CEPH database (Human Genome Diversity Project) [30], 1000 genomes project (1KGP) [31] and Peruvian Native Americans Populations (Ashaninkas and Shimaa) [32]. We merged these three databases with the genotyping results from the XP patients, as described in Kehdy et al. [32]. From the merged data, there were 273,166 SNPs in common, and LD-pruning kept 169,099 SNPs. For PCA, we used PLINK 2 [33] with the genotypic data merged with HGDP-CEPH data (323,842 variants in common), and LD-pruning (threshold of $r^2 \geq 0.1$) kept 54,216 SNPs for analysis.

Local ancestry was inferred with RfMix [34]. We used a window of 0.2 cM, two iterations of expectation maximization, and standard forward-backward with PopPhased option. Due to the limited number of genetic markers, we used three classes as references: European, African and admixed Native American to infer local ancestry at chromosome 6. Also, as RfMix pipeline is computational and time-consuming, instead of using HGDP-CEPH, we used the database from the 1KGP, as described previously [35]. From 1KGP, we used individuals from Yoruba, representing Africans, Iberians representing Europeans and Peruvians representing Admixed Native Americans.

Runs of homozygosity (ROH) segments were generated by PLINK 2 and haplotypes were inferred after phasing the data with SHAPEIT 2 software [36]. To estimate the mutation-age, we assumed a correlated genealogy, according to Gandolf et al. 2014 [37]. The dating analyses were based on segment lengths of the conserved region comprising the mutation, in which we excluded full siblings. The method calculates the time since the most recent common ancestor (MRCA) individual, assuming a correlated genealogy with a confidence interval (CI) parameter of 95 %.

Measurement of identical-by-descent (IBD) sharing was performed with the phased data (SHAPEIT2) using RefinedIBD software [38]. We used the phased data from Chromosome 6 (chr 6) merged with the HGDP-CEPH database. The parameters were a sliding window of 40 variants, LOD score of 2.0, minimal IBD length of 2 cM, and 0.1 cM for trimming the end of a shared haplotype.

Table 2

Relationship inference by estimation of kinship coefficients using KING software.

Individual 1	Individual 2	InfType
1_XP33GO.br	1_XP08GO.br	FS
12_XP04GO.br	12_XP39GO.br	FS
12_XP04GO.br	12_XP03GO.br	FS
12_XP39GO.br	12_XP03GO.br	FS
2_XP52GO.br	2_XP25GO.br	FS
13_XP05ES.eu	13_XP04ES.eu	FS
15_XP06ES.eu	15_XP02ES.eu	FS
15_XP06ES.eu	15_XP03ES.eu	FS
15_XP02ES.eu	15_XP03ES.eu	FS
12_XP04GO.br	10_XP88GO.br	2nd
12_XP04GO.br	7_XP11GO.br	2nd
12_XP39GO.br	10_XP88GO.br	2nd
12_XP39GO.br	7_XP11GO.br	2nd
12_XP03GO.br	10_XP88GO.br	2nd
12_XP03GO.br	7_XP11GO.br	2nd
9_XP56GO.br	10_XP88GO.br	2nd
6_XP100GO.br	4_XP110GO.br	2nd
12_XP39GO.br	9_XP56GO.br	3rd
12_XP04GO.br	9_XP56GO.br	3rd
9_XP56GO.br	12_XP03GO.br	3rd
6_XP100GO.br	7_XP11GO.br	4th

Abbreviations: InfType: Kinship Inferred Type; FS: Full Siblings.

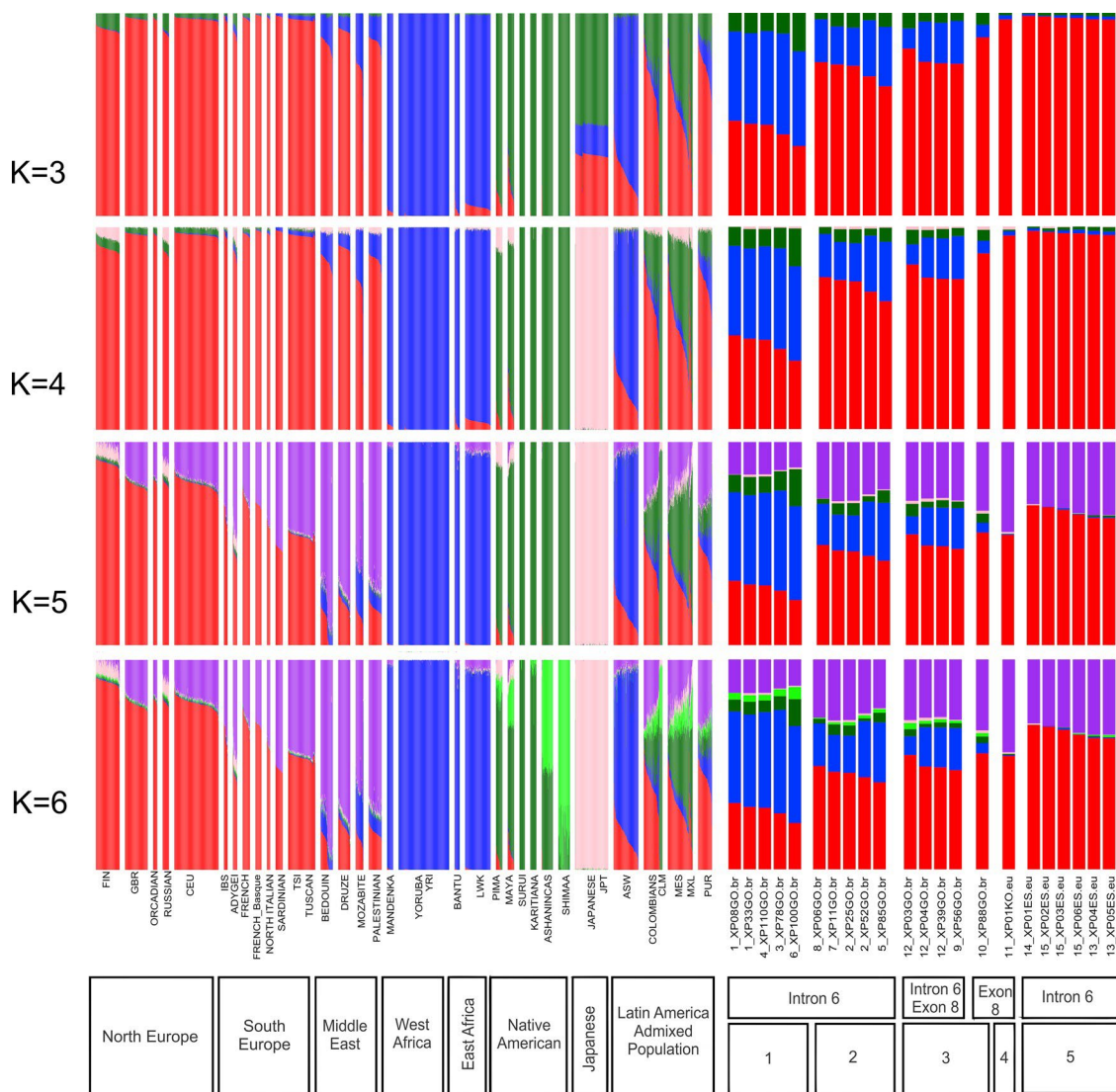


Fig. 1. ADMIXTURE analysis of the XP-V and their continental individual ancestry bar plots from $K = 3$ to $K = 6$. The color representation from the bars is explained in section 3.2 and Groups from the patients, which are identified in section 2.1. Abbreviations: FIN, Finnish in Finland; GBR, British in England and Scotland; CEU, Utah residents with Northern and Western European ancestry, USA; IBS, Iberian population in Spain; TSI, Tuscany in Italy; YRI, Yoruba in Ibadan, Nigeria; LWK, Luhya in Webuye, Kenya; JPT, Japanese in Tokyo. ASW, Americans of African ancestry in the USA; CLM, Colombians from Medellin, Colombia; MSL, Mende in Sierra Leone; MXL, Mexican ancestry from Los Angeles, USA; PUR, Puerto Ricans from Puerto Rico.

3. Results

3.1. Family structure

We checked for individuals' kinship and consanguinity levels from the 22 XP-V patients (15 families). The only patients with no family ties up to 4th-degree were 3_XP78GO.br and 11_XP01KO.eu. For the others, the data confirmed the full siblings (FS), as reported by themselves (Table 2).

The kinship between the patients from Goiás, that were not siblings, was also in agreement with the genealogy reported from Munford and Castro et al. 2017 [2], also shown here in Appendix Fig. A2. First, the three siblings from family 12 (compound heterozygotes) had a 2nd-degree kinship with patient 7_XP11GO.br, from Araras (intron 6 mutation homozygote), and patient 10_XP88GO.br (exon 8 mutation homozygote) from Hidrolândia. Second, the patient 9_XP56GO.br (compound

heterozygote) presented a 2nd-degree kinship with 10_XP88GO.br and 3rd-degree with family 12.

The reported kinship from Group 1 needed further investigation. Patient 3_XP78GO.br has never met his father and his mother had died. Indeed, KING did not identify a kinship up to the 4th-degree with other individuals from Goiás. The kinship between 4_XP110GO.br and 6_XP100GO.br, was confirmed by KING as 2nd-degree. Additionally, KING calculated the 4th-degree kinship between patients 6_XP100GO.br and 7_XP11GO.br (Table 2 and Appendix, Fig. A2). The two siblings from Faina, 1_XP08GO.br and 1_XP33GO.br (both homozygotes for intron 6 mutation) reported an old family tie with the Araras community (Appendix, Fig. A2). Indeed, their kinship is probably more distant than 4th-degree, as KING did not identify any closer familial relationship.

From Fsuíte, we estimated the inbreeding coefficient (f) (Appendix, Table A1). We detected a high consanguineity rate among the patients

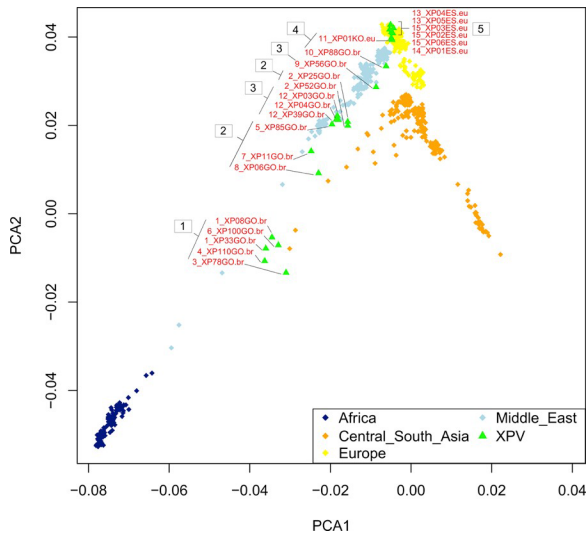


Fig. 2. Principal component (PCA) representation from PLINK software, considering the HGDP-CEPH data set as reference. Groups 1 to 5 were assigned as described in section 2.1.1.

from Araras, 10 out of 15 were considered as inbred ($pLRT_MEDIAN < 0.05$), with inbreeding coefficient rate (F_Median) ranging from 0.01 to 0.11. The compound heterozygotes and the patient 6_XP100GO.br were considered as outbred. However, for the families from Spain, families 13 and 15, there were some inconsistencies with the calculated inbreeding. FSuite detected 13_XP05ES.eu as likely to be second cousin's marriage (2C) and did not detect significant inbreeding for her brother 13_XP04ES.eu, although, they were detected as full siblings by KING software. Also, for family 15, patient 15_XP06ES.eu had a consanguinity rate extremely close to zero, differently from his sisters, but the data on Table 2 confirm these three individuals as full siblings.

3.2. Global and local ancestry inference

To study the global ancestry from the XP-V patients, we used ADMIXTURE [28], increasing the number of ancestry clusters (K) from K = 3 to K = 6 (Fig. 1). Analyses with K = 3 was important to identify the main genetic components from XP-V patients, in which red is geographically-associated with Europeans, blue with Africans and green with Native Americans (Fig. 1). Group 1 presented a mean proportion of 0.42 of European ancestry while group 2 and group 3 0.71 and 0.79, respectively. For the European patients, distributed across Groups 4 and 5, this proportion is 0.97 and 0.98. For the African component, the groups presented proportions of 0.46, 0.23, 0.15, 0.02 and 0.01 respectively (column K = 3 at Table A2, Appendix). The ancestry mean proportion among Brazilian patients from Europe, Africa, and Native America is respectively 0.64, 0.28 and 0.08. Indeed, this is highly similar to the genetic structure from the southeastern Brazil: approximately 0.7 European, 0.25 African and 0.05 Amerindians [39,40].

As shown in Fig. 1, ADMIXTURE analyses with a higher number of ancestor clusters (K = 4 to K = 6) identify more geographically-associated substructures. In Table A2, we detailed the proportions from ADMIXTURE with K = 5 and K = 6. We can identify, from K = 3, K = 5 and K = 6, that the African substructure has a higher proportion in Group 1, a lower proportion in Groups 2 and 3 and is close to zero in groups 4 and 5 (Fig. 1 and Table A2). For the European component, it is the opposite; it increases from Group 1 to Group 5. ADMIXTURE with K = 5 identifies a new substructure, depicted in purple, associated with the Middle East (ME). A new cluster appears with K = 6 (lighter green) and it did not change the European and African proportions from the patients. In fact, this new cluster separates Native Peruvians (Ashaninkas and Shimaa) from other Latin American Natives. As expected, the genetic ancestry from Native Americans for the Brazilian patients (Groups 1, 2 and 3) appears in a small proportion and it is zero among the European patients.

We can also identify a similar substructure separation in the PCA analysis (Fig. 2). Group 1 (the individuals with black skin) is displaced towards the African cluster direction (blue dots). As expected, patients from Groups 2 and 3 are scattered between Africans and Europeans,

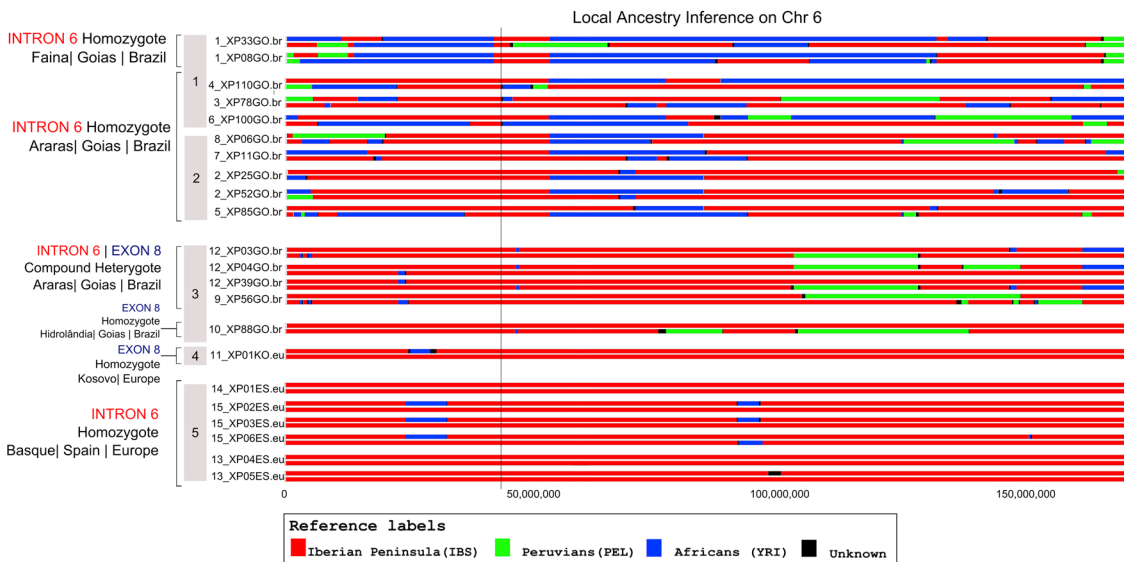


Fig. 3. Local Ancestry Inference (LAI) at chromosome 6 (Chr 6). The black vertical line indicates the localization of the *POLH* gene. IBS, Spanish individuals from the Iberian Peninsula; PEL, Peruvians from Lima; YRI, from Yoruba in Ibadan, Nigeria.

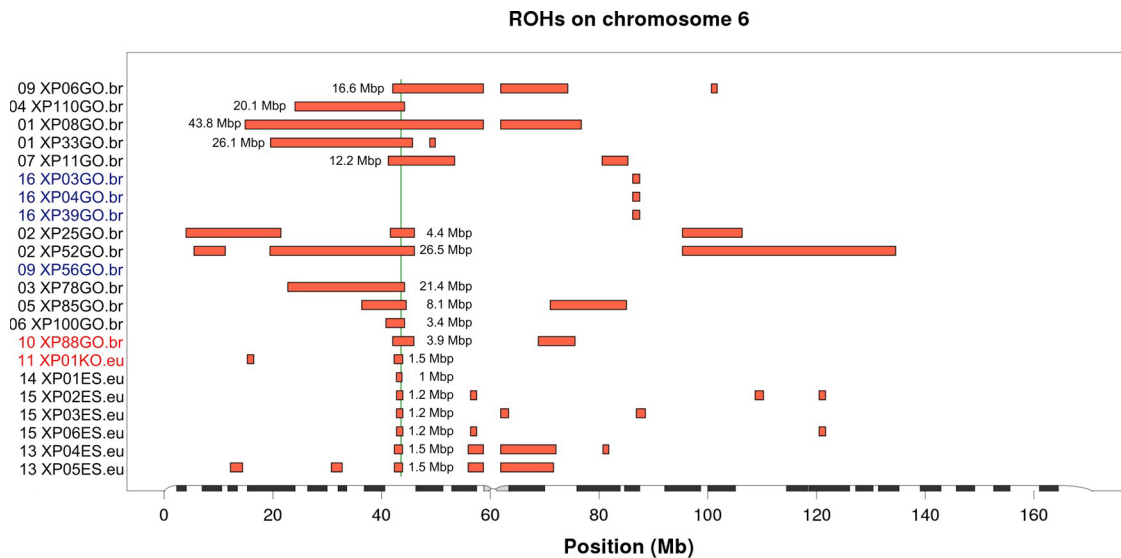


Fig. 4. Plot from Runs of Homozygosity (ROHs) burden in chromosome 6. Homozygosity mapping performed by PLINK and the plot generated by Fsuite using a window of 1 Mbp. The green line in the middle of the graph indicates the *POLH* locus. The individuals written in black are homozygotes for the intron 6 mutation, in red homozygote for exon 8 mutation and blue compound heterozygote for both mutations (intron 6 and exon 8).

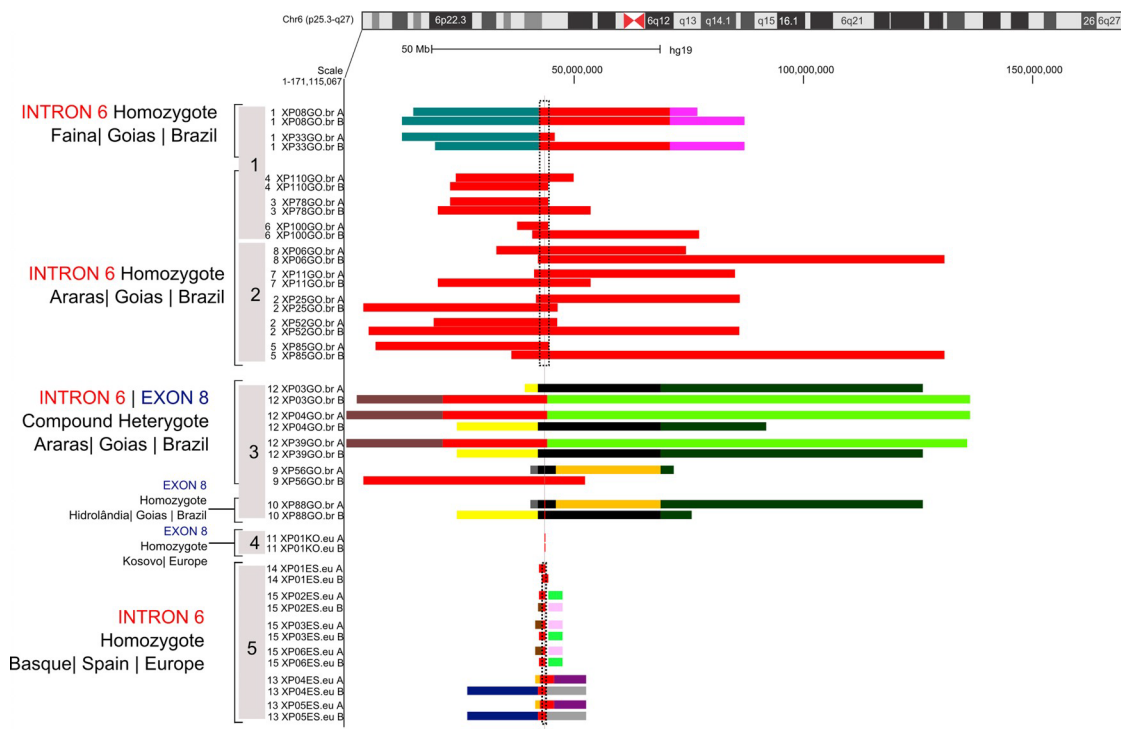


Fig. 5. Representation from the haplotype region shared with Araras around the *POLH* gene. As a graphic interface, we used the Genome Browser custom tracks tool. We compared, from each individual, the SNP haplotypes at the genomic region at chr 6: 1-171,115,067 (represented at the top of the figure). The shared haplotype was determined from the phased data at chromosome 6 (chr 6) and patients from Araras were used as a reference. Each color represents the same haplotype shared between the individuals. The red track represents the haplotype shared with Araras' patients, the only haplotype that is common between all individuals. The grey line crossing all individuals represents the *POLH* gene genomic location.

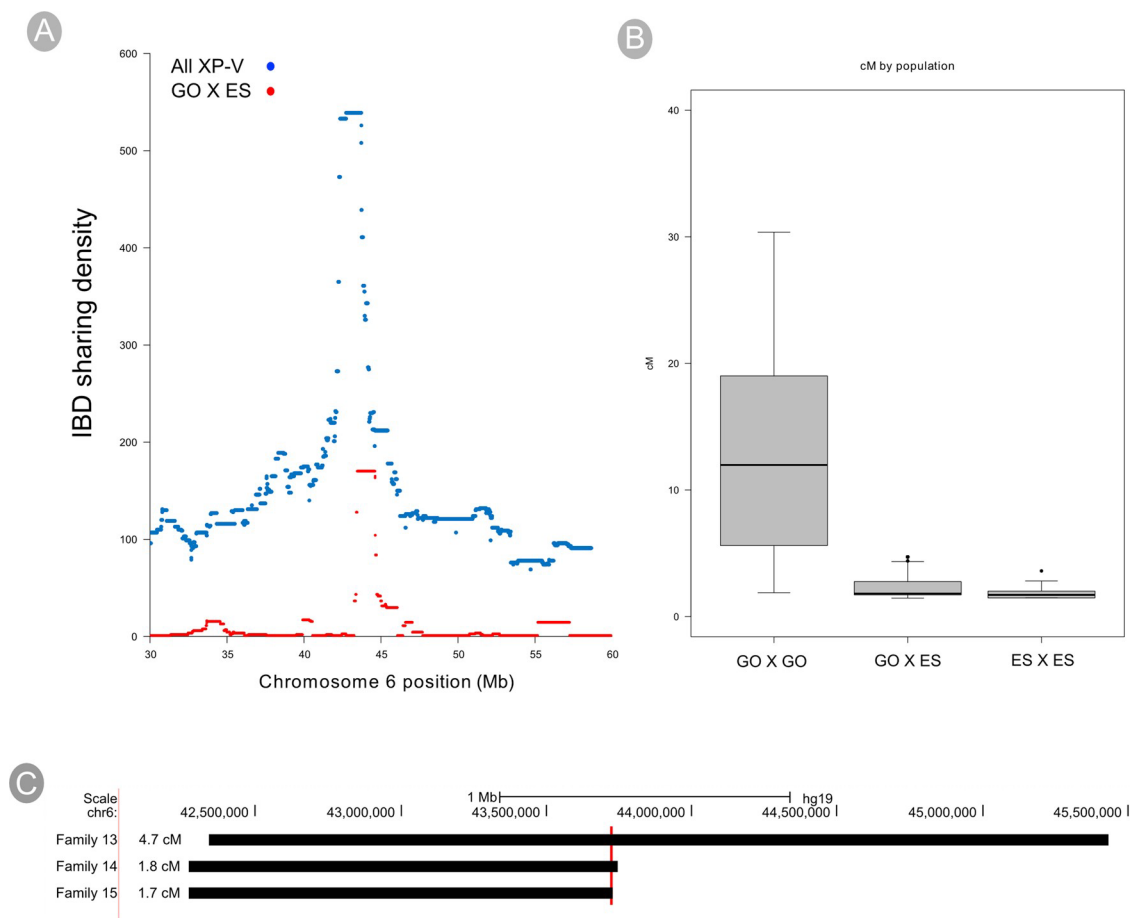


Fig. 6. Identical by descent (IBD) sharing analysis. **A.** Plot with the density of IBD segments shared at chromosome 6. In blue, the sharing segments between all XP-V patients and in red the segments shared only between patients from Goiás (GO) and Spain (ES). *POLH* gene is located at region 43,543,878-43,588,260 (GRCh37/ hg19). **B.** IBD segment sizes (cM), comprising the *POLH* gene, shared between patients from Goiás (GO X GO), Goiás and Spain (GO X ES) and only between Spanish (ES X ES). **C.** Shared IBD segment from each Spanish family with all patients from the community, except the patient 10_XP88GO.br that does not share any IBD segment at the *POLH* locus with the patients from Araras or Spain, neither with the patient from Kosovo.

overlapping with ME individuals (grey dots). Finally, for the patients from Basque Country and Kosovo, ADMIXTURE and PCA indicated predominantly European ancestry.

To infer the genetic local ancestry surrounding the mutated *POLH* gene, we performed the Local Ancestry Inference (LAI) analyses at Chromosome 6 (Fig. 3). As the Global Ancestry analyses showed, LAI also indicated an increased proportion of African component (blue) among the Brazilian patients that compose Group 1, and in a reduced proportion among European patients. However, all patients harbor European ancestry at the *POLH* locus, represented by the vertical black line crossing the chromosome 6 in Fig. 3. This result corroborates the hypothesis that the origin of this mutation is associated with the Iberian genetic background and is in agreement with the fact that patients harboring these mutations were described in Southern Europe [5,19].

3.3. Origin and time of introduction of both mutated alleles in Araras

To evaluate the possible existence of identical haplotype segments inherited from both parents, we detected runs of homozygosity regions (ROHs) (Fig. 4). All the individuals, except for the four compound heterozygotes, displayed in blue, shared an ROH region at the mutation

site. Individuals from Araras exhibited an increased burden of ROH regions, ranging from 3.4 Mbp (6_XP100GO.br) to 43.8 Mbp (1_XP08GO.br), compared to European patients, from 1 to 1.5 Mbp.

From the phased data, we identified the haplotype segments shared between Brazilian and European patients and plotted them as custom tracks at the UCSC Genome Browser (<https://genome.ucsc.edu/cgi-bin/hgCustom>). Tracks with the same color represent the same haplotype (Fig. 5). The red track includes a haplotype segment shared between all chromosomes harboring the intron 6 mutation, while a black track includes a haplotype segment shared by all chromosomes harboring the exon 8 mutation. The grey line, crossing all tracks, represents the location from *POLH* gene.

The analyses indicate a common haplotype segment of 2 Mbp in homozygosity, shared within Groups 1 and 2 (delimited by the black dotted box crossing the chromosomes from these Groups at Fig.5). However, if we consider each allele separately it is possible to identify larger shared regions, as seen for individual 5_XP85GO.br (one of the oldest patients from Araras). He shared more than 50 % of the entire chromosome 6 with individual 8_XP06GO.br.

We also identified larger haplotype segments specific for the family 1 from Faina (Group 1) and families 9, 12 and 10 (Group 3). The two

individuals from Faina (1_XP08GO.br and 1_XP33GO.br) share exclusive haplotype segments in both chromosomes (see the pink and dark grey tracks from family 1), indicating inbreeding among their ancestors. For Group 3, the three individuals from family 12 and the individual 9_XP56GO.br share the haplotype segment harboring intron 6 from Araras and another haplotype segment that came from Hidrolândia. Although the Hidrolândia's chromosomes from families 9 (grey, black, orange and dark green colors) and 12 (yellow, black and dark green colors) are quite different from each other, both chromosomes are observed within the individual 10_XP88GO.br.

The two patients, homozygotes for exon 8 mutation, 10_XP88GO.br and 11_XP01KO.eu from Kosovo, only share a small size of 0.2 Mb and 0.1 Mb, respectively, from both alleles. The absence of a more significant shared haplotype indicates no common ancestor or a long genetic distance separating apart the family ties between these two patients.

For the three Spanish families, we identified one Mbp of shared haplotype segment in homozygosity (Fig. 5, delimited by the black dotted box crossing Group 5). Camels and collaborators also studied the haplotypes from these families and indeed identified this one Mbp shared within the three families and suggested a common ancestor of 500–1000 years ago [19]. Interestingly, the entire one Mbp region in homozygosity was also identified in all patients carrying the intron 6 mutation in Araras. However, slightly larger haplotype segments are shared between the Spanish and Brazilian patients. As shown in a zoom, presented in Fig. A3, the family 13 shares 3.1 Mbp and 1.7 Mbp from each allele with the Brazilian patients, the family 15, 1.5 Mbp and 1 Mbp from each allele and the individual from family 14, 1.5 Mbp from both alleles. The larger shared haplotype segment from family 13 indicates that these patients have a closer family link with Araras patients.

We also performed the identical-by-descent (IBD) sharing analyses in order to identify the precise ancestral segment originated from the Basque Country at Araras. From the IBD output, we plotted the sharing density at chromosome 6 (Fig. 6A) within all XP individuals (in blue) and shared between patients from Goiás and Spain (in red). The peak, around the 45 Mbp genomic position, indicates a higher density of shared IBD segments at the mutation site within all XP-V patients. In addition, both patients homozygous for exon 8 mutation, 10_XP88GO.br and 11_XP01KO.eu, do not share any IBD segment at this mutation site.

The size of shared IBD segments (Fig. 6B) were also analysed. All patients from Goiás, homozygotes for intron 6 mutation, share an IBD segments varying between from 4 and 30 cM, that comprise the mutation. We compared the segments shared between Brazilian (GO X GO), within Brazilian and Spanish (GO X ES) and shared among Spanish patients (ES X ES). The higher density and larger segments shared within the patient from Goiás (GO X GO) are expected as the majority of these patients share a 2nd or 3th-degree kinship. Indeed, the consanguinity effect inflates these values [41].

Interestingly, Goiás and Spanish patients (GO X ES), shared on average, a larger segment compared to the segments shared between Spanish patients (ES X ES). As previously indicated by haplotype analyses, the higher average on the segment size shared between GO X ES is largely due to family 13 (Fig. 6C). Indeed, haplotype segments from this family are, apparently, more closely related to those of Araras patients than to those of the other two Spanish families. The IBD analyses identified the exact 3.1 Mbp genomic region (4.7 cM) shared between the Spanish family 13 (13_XP04ES.eu and 13_XP05ES.eu) and all individuals from Goiás that carry the intron 6 mutation (Fig. 6C).

Based on the genetic lengths shared between the individuals from Araras (excluding full siblings), we estimated the age of intron 6 mutation in that community. The entrance date at Araras is approximately 7.7 (95 % CI: 2.9–21.4) generations, corresponding to roughly 200 (95 % CI: 75–525) years ago, assuming 25 years per generation. Indeed, according to their history and to their previously reconstituted genealogy [2] (Appendix, Fig. A2), European (Portuguese) arrived at that location at the beginning of XVIII century, when most probably the founder haplotype comprising the intron 6 mutation at *POLH* gene (indicated in red in Figs. 5 and A2) was introduced.

4. Discussion

The present work aimed to study the genetic resemblance from 22 XP-V patients harboring the same mutation at *POLH* gene. The population relationship study revealed a high consanguineous rate among the cohort; 13 out of 22 were considered as inbred by FSuite. For the Spanish patients, we identified some inconsistencies with the FSuite consanguinity estimation, probably due to statistical issues. For individual 13_XP05ES.eu, the inbreeding coefficient (0.024) is probably not valid (only 35 % of the valid submaps were considered for statistical analysis). The two sisters, 15_XP02ES.eu and 15_XP03ES.eu, are probably consanguineous (significant test were detected for 100 out of 100 and 96 out of 98 submaps, respectively). The estimated coefficient for 15_XP06ES.eu was valid only for two submaps, out of 99, and it is probably not reliable.

The global ancestry study with ADMIXTURE and PCA unravel that the Brazilians patients presented an essential proportion of the geographical-associated components of European, Middle East and African ancestry. Differently, and expected, the Europeans patients did not present the African substructure. The European and Middle East components compose the Brazilian and Europeans XP-V ancestry structure, as well as, the populations from South Europe. This distribution may reflect the fact that Middle East played a major role in modern human expansions out of Africa, and for later migrations into and out of Europe [42]. In fact, the European patients from Basque Country and Kosovo present more than 30 % of this ME substructure, which indicates that this component could represent the European genetic background.

Indeed, we detected a shared IBD segment among the Brazilian and European patients. Taking into account the history from Brazil and what was reported by the individuals from the community, together with the LAI, we showed that the genomic segment comprising *POLH* intron 6 mutation at Araras is associated with Iberian ancestry. The data indicates that this mutation was probably introduced in Araras close to 200 years ago, which coincides with the reported arrival of Portuguese descendants. Moreover, the results clearly indicate the Iberian legacy of this XP mutation and, interestingly, the closer resemblance of haplotype segments from patients in Goiás and from one of the families in the Basque Country.

5. Conclusions

The outcomes from global ancestry from the Brazilian XP-V patients corroborate the three-hybrid profile from the general Brazilian genetic structure. It includes the background from Native Amerindians, Europeans and Africans. The Natives are represented in a minority proportion, due to their holocaust that almost depleted their genetic contribution. The European ancestry, mainly from the Iberian Peninsula, had a stronger effect at the Brazilian genetics context due to their colonization since the XVI century. The African ancestry,

following 300 years of slavery period, varies across the different Brazilian geopolitical regions. Therefore, in spite of phenotypic diversity, Brazilians genetic ancestry is more uniform than expected and is strongly determined by European and African components [32,39,43,44].

Herein, we demonstrated the Iberian genetic legacy into an isolated community in central Brazil, where a mutation at the *POLH* gene is responsible for several patients affected by the XP syndrome. Interestingly, the haplotype segment from family 13 of the Basque Country presents a higher resemblance with segments observed in Araras, than with segments from the other two families in Spain. Therefore, the data indicate that approximately at least 200 years ago, a rare genetic mutation at *POLH* crossed the Atlantic Ocean and arrived on the Brazilian coast. This is an interesting example of the impact of human migrations and demographic events (such as founder effects and inbreeding) that happened during the historical colonization of Brazil on the incidence of genetic disorders.

Funding

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP),

Appendix A

Table A1

Inbreeding detected by Fsuite. For each individual, Fsuite estimated the probability to be the descendant offspring, classified as first cousin (1C), second cousin (2C), avuncular (AV), or unrelated (OUT). A likelihood ratio test (pLRT) was performed to assign each individual as inbred (1) or not (0) (Inbreed column). We highlighted the most probably inbreeding level of offspring classification from each individual.

Group	Individual ID	Submaps	Quality	F_Median	A_Median	pLRT_MEDIAN	Inbreed	pLRT_<0.05	1C	2C	AV	OUT
1	1_XP08GO.br	100/100	100.00	0.037	0.074	7.7E-14	1	100	0.6591	0.3395	0.0014	0
	1_XP33GO.br	100/100	100.00	0.062	0.059	1.0E-29	1	100	0.9669	0.0033	0.0299	0.0001
	4_XP110GO.br	100/100	100.00	0.008	0.098	1.3E+05	1	100	0.0012	0.9988	0	0
	3_XP78GO.br	100/100	100.00	0.103	0.046	1.2E-39	1	100	0.4336	0	0.5664	0
	6_XP100GO.br	100/100	100.00	0.000	0.019	0.97487	0	0	0	0.0660	0	0.9340
2	8_XP06GO.br	100/100	100.00	0.029	0.106	2.2E-11	1	100	0.4232	0.5764	0.0005	0.0001
	7_XP11GO.br	100/100	100.00	0.012	0.140	3.3E-05	1	100	0.0072	0.9928	0	0
	2_XP25GO.br	100/100	100.00	0.071	0.063	4.4E-33	1	100	0.8956	0.0003	0.1041	0
	2_XP52GO.br	100/100	100.00	0.092	0.065	6.6E-55	1	100	0.3941	0	0.6059	0
	5_XP85GO.br	100/100	100.00	0.111	0.058	1.6E-48	1	100	0.1064	0	0.8936	0
3	12_XP03GO.br	100/100	100.00	0.000	0.016	0.97492	0	0	0	0.0425	0	0.9575
	12_XP04GO.br	100/100	100.00	0.000	0.015	0.97485	0	0	0	0.0381	0	0.9619
	12_XP39GO.br	100/100	100.00	0.000	0.015	0.97489	0	0	0	0.0380	0	0.9620
	9_XP56GO.br	100/100	100.00	0.000	0.015	0.97487	0	0	0	0.0385	0	0.9615
	10_XP88GO.br	100/100	100.00	0.011	0.286	0.00045	1	97	0.0074	0.9871	0	0.0055
4	11_XP01KO.eu	100/100	100.00	0.000	0.026	0.97491	0	4	0.0001	0.1500	0	0.8499
5	14_XP01ES.eu	100/100	100.00	0.000	0.018	0.97487	0	0	0	0.0530	0	0.9470
	15_XP02ES.eu	100/100	100.00	0.011	0.361	0.05161	1	100	0.0035	0.9964	0	0.0001
	15_XP03ES.eu	98/100	98.00	0.011	0.485	0.00041	1	96	0.0018	0.9921	0	0.0061
	15_XP06ES.eu	99/100	99.00	0.000	0.024	0.97494	0	2	0	0.0985	0	0.9015
	13_XP04ES.eu	98/100	98.00	0.000	0.029	0.97495	0	3	0.0002	0.2228	0	0.7770
	13_XP05ES.eu	35/100	35.00	0.024	0.638	0.07749	1	20	0.0075	0.9727	0	0.0198

Table A2

Results from ADMIXTURE analysis representing the associated proportion with the geographical inferred ancestry cluster (K).

	K = 3 (CV 0.587)			K = 5 (CV 0.550)					K = 6 (CV 0.547)						
	EUR	AFR	NAT	EUR	AFR	NAT	JAP	ME	POP	EUR	AFR	NAT_BR	NAT_PER	JAP	ME
POP	0.42	0.46	0.12	0.28	0.46	0.11	0.01	0.15	Group 1	0.28	0.46	0.06	0.04	0.01	0.15
Group 1	0.71	0.23	0.06	0.46	0.22	0.05	0.01	0.27	Group 2	0.46	0.22	0.04	0.01	0.01	0.27
Group 2	0.79	0.15	0.06	0.51	0.14	0.04	0.01	0.30	Group 3	0.51	0.14	0.03	0.01	0.01	0.30
Group 3	0.97	0.02	0.01	0.54	0.00	0.00	0.01	0.46	Group 4	0.54	0.00	0.01	0.00	0.01	0.46
Group 4	0.98	0.01	0.01	0.65	0.00	0.00	0.00	0.35	Group 5	0.65	0.00	0.00	0.00	0.00	0.35
Group 5	0.96	0.01	0.03	0.73	0.00	0.01	0.02	0.24	EUR	0.73	0.00	0.01	0.00	0.02	0.24
EUR	0.00	1.00	0.00	0.00	0.98	0.00	0.00	0.02	AFR	0.00	0.98	0.00	0.00	0.00	0.02
AFR	0.31	0.18	0.51	0.19	0.18	0.49	0.03	0.11	NAT_BR	0.29	0.28	0.17	0.06	0.03	0.17
NAT	-	-	-	0.00	0.00	0.01	0.99	0.00	NAT_PER	0.02	0.00	0.57	0.38	0.02	0.01
JAP	-	-	-	0.16	0.10	0.01	0.01	0.72	JAP	0.00	0.00	0.01	0.00	0.99	0.00
ME	-	-	-	-	-	-	-	-	ME	0.16	0.10	0.01	0.01	0.01	0.71
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Abbreviations: POP = Population, CV = Cross-validation error, EUR = Europeans, AFR = Africans, NAT = Natives Americans, JAP = Japanese, ME = Middle East, NAT_BR = Admixed Native Americans, NAT_PER = Native Americans Peruvians.

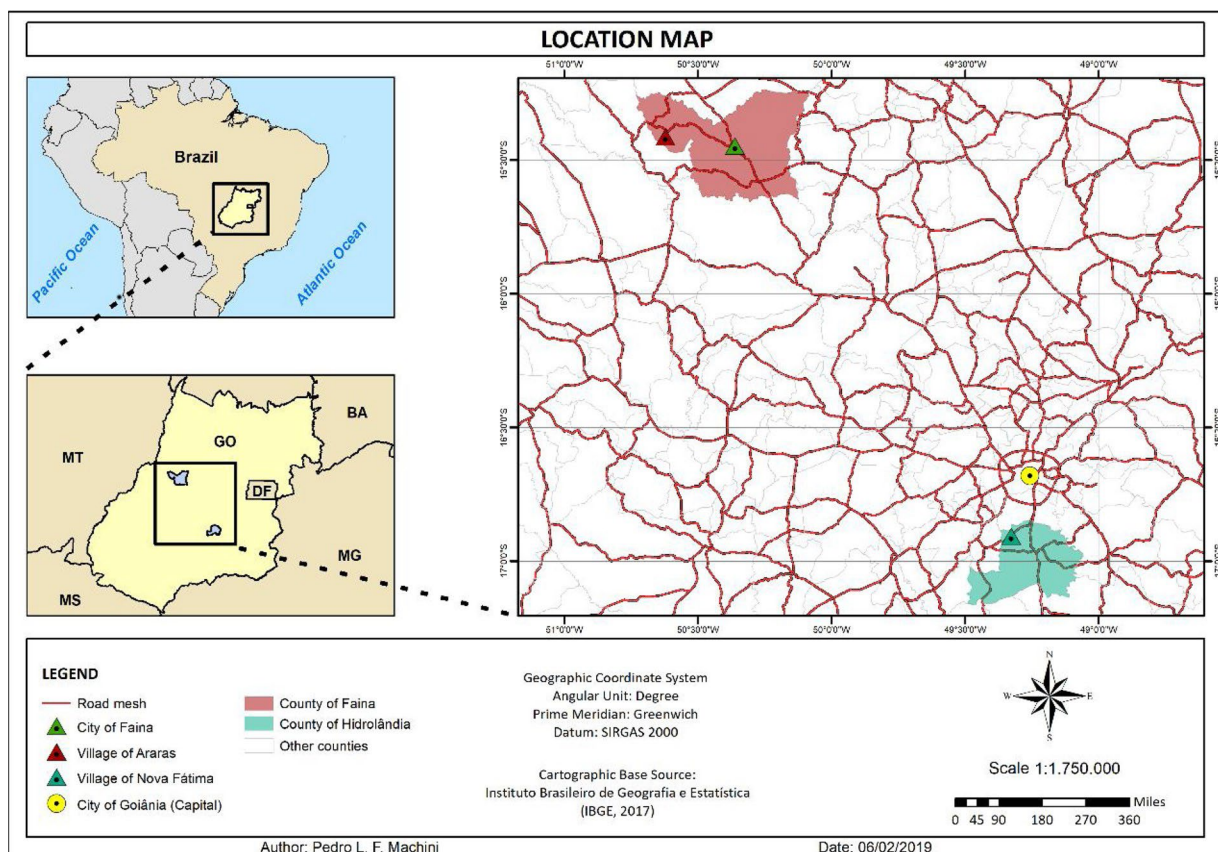


Fig. A1. Location map from Goiás state (BR) pointing Araras and Nova Fátima villages, the city of Faina and its municipality. Araras village and Faina city are in the municipality of Faina. Goiânia (yellow) is the capital from Goiás state, with 1.3 million inhabitants (Brazilian Institute of Geography and Statistics - IBGE 2010). Abbreviations: GO, Goiás; MT, Mato Grosso; MS, Mato Grosso do Sul; MG, Minas Gerais; BA, Bahia; DF, Federal District.

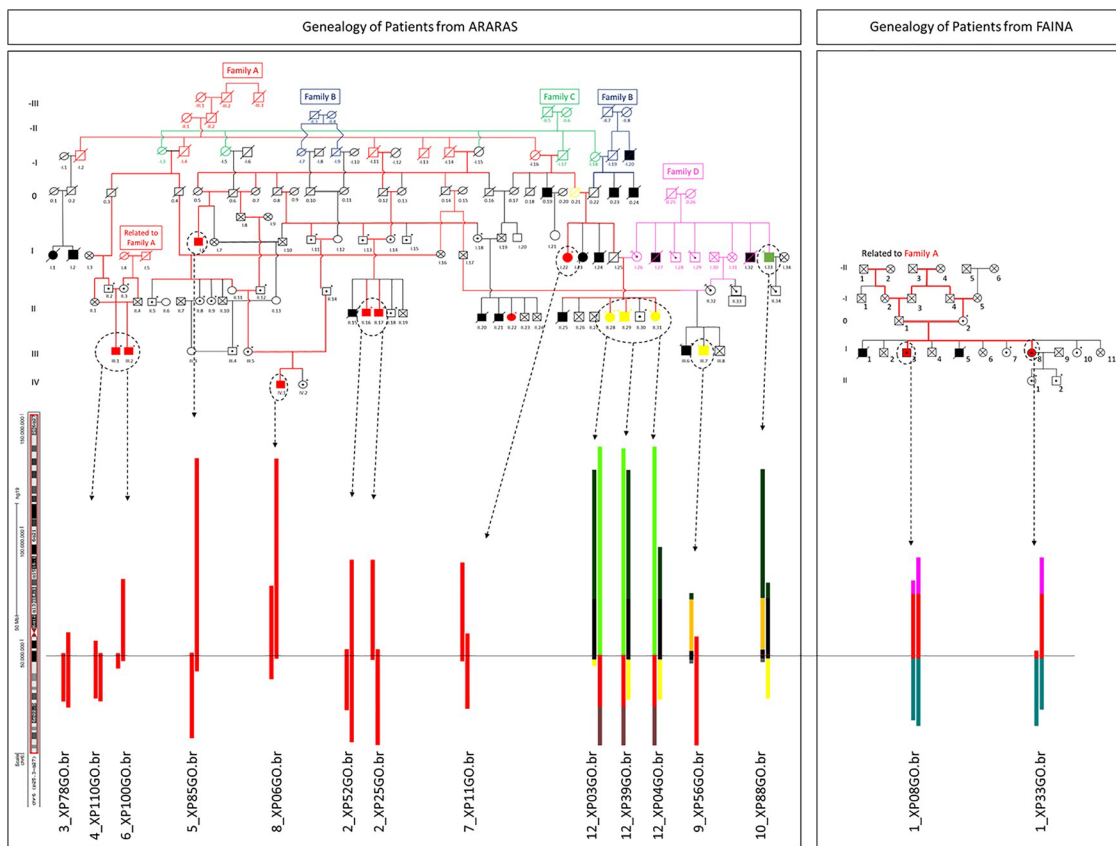


Fig. A2. Graphical representation of the haplotypes from each individual and their link with the founder family (Family A) represented in this study as the Araras/Faina’s haplotype in red. From the genealogy, individuals in red represent the homozygotes for intron 6 mutation, in yellow the compound heterozygotes and green the homozygote for exon 8 mutation. Family B corresponds basically to people who live in Faina. The horizontal line that cuts the graphic indicates the position of the studied mutations.

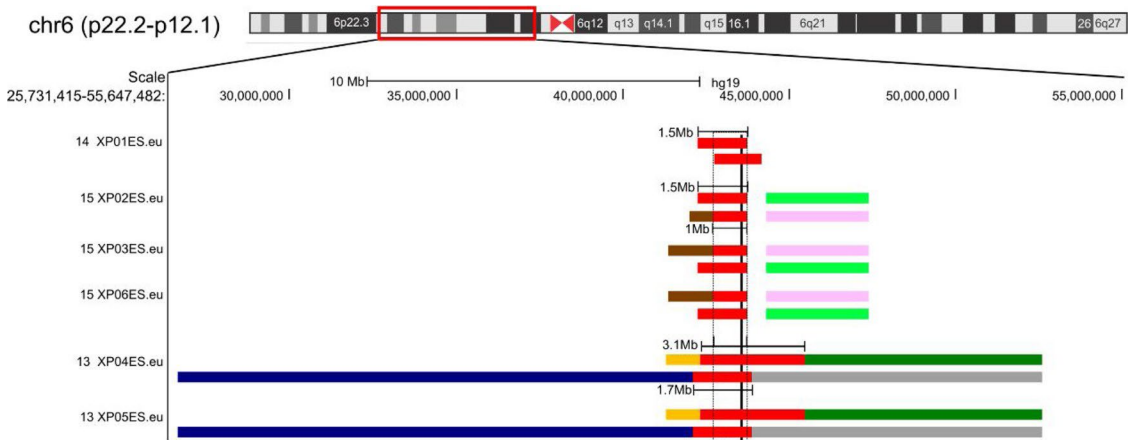


Fig. A3. Zoom in (from 25,731,415 to 55,647,482 genomic location at chr 6, from GRCh37/hg19) from the Genome browser visualization at the Spanish families haplotypes segments. In red, the haplotype segment shared with Araras patients. The black box (dotted) represents the one Mbp that is shared between the three families. The solid black line crossing all haplotypes represents the *POLH* gene location.

References

[1] G.C. Cardoso, M.Z. de Oliveira, V.R. Paixão-Côrtes, E.E. Castilla, L. Schuler-Faccini, Clusters of genetic diseases in Brazil, *J. Community Genet.* 10 (2018) 121–128.

[2] V. Munford, L.P. Castro, R. Souto, L.K. Lerner, J.B. Vilar, C. Quayle, H. Asif, A.P. Schuch, T.A. de Souza, S. Ienne, F.I.A. Alves, L.M.S. Moura, P.A.F. Galante, A.A. Camargo, R. Liboredo, S.D.J. Pena, A. Sarasin, S.C. Chaibub, C.F.M. Menck, A genetic cluster of patients with variant xeroderma pigmentosum with two different founder mutations, *Br. J. Dermatol.* 176 (2017) 1270–1278.

[3] J. Walburn, M. Canfield, S. Norton, K. Sainsbury, V. Araújo-Soares, L. Foster, M. Berneburg, A. Sarasin, N. Morrison-Bowen, F.F. Sniehotta, R. Sarkany, J. Weinman, Psychological correlates of adherence to photoprotection in a rare disease: international survey of people with Xeroderma Pigmentosum, *Br. J. Health Psychol.* 24 (2019) 668–686.

[4] H. Fasshi, M. Sethi, H. Fawcett, J. Wing, N. Chandler, S. Mohammed, E. Craythorne, A.M. Morley, R. Lim, S. Turner, T. Henshaw, I. Garrood, P. Giunti, T. Hedderly, A. Abiona, H. Naik, G. Harrop, D. McGibbon, N.G. Jaspers, E. Botta,

- T. Nardo, M. Stefanini, A.R. Young, R.P. Sarkany, A.R. Lehmann, Deep phenotyping of 89 xeroderma pigmentosum patients reveals unexpected heterogeneity dependent on the precise molecular defect, *Proc. Natl. Acad. Sci.* 113 (2016) e1236–e1245.
- [5] K. Opletalova, A. Bourillon, W. Yang, C. Pouvelle, J. Armier, E. Despras, M. Ludovic, C. Mateus, C. Robert, P. Kannouche, N. Soufir, A. Sarasin, Correlation of phenotype/genotype in a cohort of 23 xeroderma pigmentosum-variant patients reveals 12 new disease-causing POLH mutations, *Hum. Mutat.* 35 (2014) 117–128.
- [6] W.J. Kleijer, V. Laugel, M. Bernsburg, T. Nardo, H. Fawcett, A. Gratchev, N.G. Jaspers, A. Sarasin, M. Stefanini, A.R. Lehmann, Incidence of DNA repair deficiency disorders in western Europe: Xeroderma pigmentosum, Cockayne syndrome and trichothiodystrophy, *DNA Repair (Amst)* 7 (2008) 744–750.
- [7] Y. Hirai, Y. Kodama, S. Moriwaki, A. Noda, H.M. Cullings, D.G. Macphee, K. Kodama, K. Mabuchi, K.H. Kraemer, C.E. Land, N. Nakamura, Heterozygous individuals bearing a founder mutation in the XPA DNA repair gene comprise nearly 1% of the Japanese population, *Mutat. Res.* 601 (2006) 171–178.
- [8] M. Jerbi, M. Ben Rekaya, C. Naouali, M. Jones, O. Messaoud, H. Tounsi, M. Nagara, M. Chargui, R. Kefi, H. Boussen, M. Mokni, R. Mrad, M.S. Boubaker, S. Abdelhak, A. Khaled, M. Zghal, H. Yacoub-Youssef, Clinical, genealogical and molecular investigation of the xeroderma pigmentosum type C complementation group in Tunisia, *Br. J. Dermatol.* 174 (2016) 439–443.
- [9] M. Ben Rekaya, N. Laroussi, O. Messaoud, M. Jones, M. Jerbi, C. Naouali, Y. Bouyacoub, M. Chargui, R. Kefi, B. Fazaa, M.S. Boubaker, H. Boussen, M. Mokni, S. Abdelhak, M. Zghal, A. Khaled, H. Yacoub-Youssef, A founder large deletion mutation in Xeroderma pigmentosum-Variant form in Tunisia: implication for molecular diagnosis and therapy, *Biomed Res. Int.* 2014 (2014) 256245.
- [10] M. Kgekolo, F. Morice-Picard, H.R. Rezvani, F. Austerlitz, F. Cartault, A. Sarasin, M. Sathekge, A. Taieb, C. Ged, Xeroderma pigmentosum in South Africa: evidence for a prevalent founder effect, *Br. J. Dermatol.* 181 (2019) 1070–1072.
- [11] F. Cartault, C. Nava, A.C. Malbrunot, P. Munier, J.C. Hebert, P. N'guyen, N. Djeridi, P. Pariaud, J. Pariaud, A. Dupuy, F. Austerlitz, A. Sarasin, A new XPC gene splicing mutation has lead to the highest worldwide prevalence of xeroderma pigmentosum in black Mahori patients, *DNA Repair (Amst)* 10 (2011) 577–585.
- [12] O. Messaoud, M. Ben Rekaya, W. Cherif, F. Talmoudi, H. Boussen, I. Mokhtar, S. Boubaker, A. Amouri, S. Abdelhak, M. Zghal, Genetic homogeneity of mutational spectrum of group-A xeroderma pigmentosum in Tunisian patients, *Int. J. Dermatol.* 49 (2010) 544–548.
- [13] N. Soufir, C. Ged, A. Bourillon, F. Austerlitz, C. Chemin, A. Stary, J. Armier, D. Pham, K. Khadir, J. Roume, S. Hadj-Rabia, B. Bouadjar, A. Taieb, H. de Verneuil, H. Benchiki, B. Grandchamp, A. Sarasin, A prevalent mutation with founder effect in xeroderma pigmentosum group C from north Africa, *J. Invest. Dermatol.* 130(2010) 1537–1542.
- [14] J.E. Cleaver, L. Feeney, J.Y. Tang, P. Tuttle, Xeroderma pigmentosum group C in an isolated region of Guatemala, *J. Invest. Dermatol.* 127 (2007) 493–496.
- [15] A. Ijaz, S. Basit, A. Gul, L. Batoor, A. Hussain, S. Afzal, K. Ramzan, J. Ahmad, A. Wali, XPC gene mutations in families with xeroderma pigmentosum from Pakistan: prevalent founder effect, *Congenit. Anom. (Kyoto)* 59 (2019) 18–21.
- [16] T.C. Falik-Zaccari, R. Erel-Segal, L. Horev, O. Bitterman-Deutsch, S. Koka, S. Chaim, Z. Keren, L. Kalfon, B. Gross, Z. Segal, S. Orgal, Y. Shoval, H. Slor, G. Spivak, P.C. Hanawalt, A novel XPD mutation in a compound heterozygote; the mutation in the second allele is present in three homozygous patients with mild sun sensitivity, *Environ. Mol. Mutagen.* 53 (2012) 505–514.
- [17] A. Sarasin, P. Munier, F. Cartault, How history and geography may explain the distribution in the Comorian archipelago of a novel mutation in DNA repair-deficient xeroderma pigmentosum patients, *Genet. Mol. Biol.* 43 (2020) e20190046.
- [18] L. Alwatban, Y. Binamer, Xeroderma pigmentosum at a tertiary care center in Saudi Arabia, *Ann. Saudi Med.* 37 (2017) 240–244.
- [19] N. Calmels, G. Greff, C. Obringer, N. Kempf, C. Gasnier, J. Tarabeux, M. Miguet, G. Baujat, D. Bessis, P. Bretones, A. Cavau, B. Digeon, M. Doco-Fenzy, B. Doray, F. Feillet, J. Gardeazabal, B. Gener, S. Julia, I. Llano-Rivas, A. Mazur, C. Michot, F. Renaldo-Robin, M. Rossi, P. Sabouraud, B. Keren, C. Depienne, J. Muller, J.L. Mandel, V. Laugel, Uncommon nucleotide excision repair phenotypes revealed by targeted high-throughput sequencing, *Orphanet J. Rare Dis.* 11 (2016) 26.
- [20] G. Machado, Nas Asas Da Esperança: A História De Dor E Resistência Da Comunidade De Araras, second edition ed., Kelps, Goiania - GO, 2011.
- [21] J.L. Pedrosa, P. Braga-Neto, J. Radvany, O.G. Barsottini, Machado-Joseph disease in Brazil: from the first descriptions to the emergence as the most common spinocerebellar ataxia, *Arq. Neuropsiquiatr.* 70 (2012) 630–632.
- [22] F.M. Costa-Motta, F. Bender, A. Acosta, K. Abé-Sandes, T. Machado, T. Bomfim, T. Boa Sorte, D. da Silva, A. Bittles, R. Giugliani, S. Leistner-Segal, A community-based study of mucopolysaccharidosis type VI in Brazil: the influence of founder effect, endogamy and consanguinity, *Hum. Hered.* 77 (2014) 189–196.
- [23] G. Coutinho, M. Mitui, C. Campbell, B.T. Costa Carvalho, S. Nahas, X. Sun, Y. Huo, C.H. Lai, Y. Thorstenson, R. Tanouye, S. Raskin, C.A. Kim, J. Llerena, R.A. Gatti, Five haplotypes account for fifty-five percent of ATM mutations in Brazilian patients with ataxia telangiectasia: seven new mutations, *Am. J. Med. Genet. A* 126A(2004) 33–40.
- [24] C.V. Dillenburg, I.C. Bandeira, T.V. Tubino, L.G. Rossato, E.S. Dias, A.C. Bittelbrunn, S. Leistner-Segal, Prevalence of 185delAG and 5382insC mutations in BRCA1, and 6174delT in BRCA2 in women of Ashkenazi Jewish origin insouthern Brazil, *Genet. Mol. Biol.* 35 (2012) 599–602.
- [25] A. Manichaikul, J.C. Mychaleckyj, S.S. Rich, K. Daly, M. Sale, W.M. Chen, Robustrelationship inference in genome-wide association studies, *Bioinformatics* 26 (2010) 2867–2873.
- [26] S. Gazal, M. Sahbatou, M.C. Babron, E. Génin, A.L. Leutenegger, FSuite: exploiting inbreeding in dense SNP chip and exome data, *Bioinformatics* 30 (2014) 1940–1941.
- [27] A.L. Leutenegger, B. Prum, E. Génin, C. Verna, A. Lemainque, F. Clerget-Darpoux, W.F. Bodmer, B. Bonne-Tamir, Estimation of the inbreeding coefficient through use of genomic data, *Am. J. Hum. Genet.* 73 (2003) 516–523.
- [28] D.H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals, *Genome Res.* 19 (2009) 1655–1664.
- [29] A.L. Price, J.P. Nick, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies, *Nat. Genet.* 38 (2006) 904–909.
- [30] H.M. Cann, C. de Toma, L. Cazes, M.F. LeGrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Caracci, L. Contu, R. Du, L. Excoffier, G.B. Ferrara, J.S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R.J. Herrera, X. Huang, J. Kidd, K.K. Kidd, A. Langaney, A.A. Lin, S.Q. Mehdi, P. Parham, A. Piazza, M.P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J.L. Weber, H.T. Greely, M.W. Feldman, G. Thomas, J. Dausset, L.L. Cavalli-Sforza, A human genome diversity cell line panel, *Science* 296 (2002) 261–262.
- [31] G.R. Abecasis, A. Auton, L.D. Brooks, M.A. DePristo, R.M. Durbin, R.E. Handsaker, H.M. Kang, G.T. Marth, G.A. McVean, G.P. Consortium, An integrated map of genetic variation from 1,092 human genomes, *Nature* 491 (2012) 56–65.
- [32] F.S. Kehdy, M.H. Gouveia, M. Machado, W.C. Magalhães, A.R. Horimoto, B.L. Horta, R.G. Moreira, T.P. Leal, M.O. Scliar, G.B. Soares-Souza, F. Rodrigues-Soares, G.S. Araújo, R. Zamudio, H.P. Sant Anna, H.C. Santos, N.E. Duarte, R.L. Fiaccone, C.A. Figueiredo, T.M. Silva, G.N. Costa, S. Beleza, D.E. Berg, L. Cabrera, G. Debortoli, D. Duarte, S. Ghirrotto, R.H. Gilman, V.F. Gonçalves, A.R. Marrero, Y.C. Muniz, H. Weissensteiner, M. Yeager, L.C. Rodrigues, M.L. Barreto, M.F. Lima-Costa, A.C. Pereira, M.R. Rodrigues, E. Tarazona-Santos, B.E.P. Consortium, Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations, *Proc. Natl. Acad. Sci. U. S. A.* 112 (2015) 8696–8701.
- [33] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly, P.C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses, *Am. J. Hum. Genet.* 81 (2007) 559–575.
- [34] 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 (2013) 278–288.
- [35] A.A. de Farias, K. Nunes, R.B. Lemes, R. Moura, G.R. Fernandes, U.S. Melo, M. Zatz, F. Kok, S. Santos, Origin and age of the causative mutations in KLC2, IMPA1, MED25 and WNT7A unravelled through Brazilian admixed populations, *Sci. Rep.* 8 (2018) 16552.
- [36] O. Delaneau, B. Howie, A.J. Cox, J.F. Zagury, J. Marchini, Haplotype estimation using sequencing reads, *Am. J. Hum. Genet.* 93 (2013) 687–696.
- [37] L.C. Gandolfo, M. Bahlo, T.P. Speed, Dating rare mutations from small samples with dense marker data, *Genetics* 197 (2014) 1315–1327.
- [38] B.L. Browning, S.R. Browning, Improving the accuracy and efficiency of identity-by-descent detection in population data, *Genetics* 194 (2013) 459–471.
- [39] S.D. Pena, G. Di Pietro, M. Fuchshuber-Moraes, J.P. Genro, M.H. Hutz, Fe.S. Kehdy, F. Kohlrausch, L.A. Magno, R.C. Montenegro, M.O. Moraes, M.E. de Moraes, M.R. de Moraes, E.B. Jopji, J.A. Perini, C. Racciopi, A.K. Ribeiro-Dos-Santos, F. Rios-Santos, M.A. Romano-Silva, V.A. Sortica, G. Suarez-Kurtz, The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected, *PLoS One* 6 (2011) e17063.
- [40] G.D. Valle-Silva, F.D.N. Souza, L. Marcorin, A.I.E. Pereira, T.M.T. Carratto, G. Debortoli, M.L.G. Oliveira, N.C.A. Fracasso, E.S. Andrade, E.A. Donadi, H.L. Norton, E.J. Parra, A.L. Simões, E.C. Castelli, C.T. Mendes-Junior, Applicability of the SNPforID 52-plex panel for human identification and ancestry evaluation in a Brazilian population sample by next-generation sequencing, *Forensic Sci. Int. Genet.* 40 (2019) 201–209.
- [41] A.L. Severson, S. Carmi, N.A. Rosenberg, The effect of consanguinity on between-individual identity-by-descent sharing, *Genetics* 212 (2019) 305–316.
- [42] D.A. Badro, et al., Y-chromosome and mtDNA genetics reveal significant contrasts in affinities of modern middle eastern populations with european and african populations, *PLoS One* 8 (2013) e54616.
- [43] M.R. Passos-Bueno, D. Bertola, D.D. Horovitz, V.E. de Faria Ferraz, L.A. Brito, Genetics and genomics in Brazil: a promising future, *Mol. Genet. Genomic Med.* 2 (2014) 280–291.
- [44] F.M. Salzano, M.C. Bortolini, The Evolution and Genetics of Latin American Populations, Cambridge University Press, 2002.

Origins, Admixture Dynamics, and Homogenization of the African Gene Pool in the Americas

Mateus H. Gouveia,^{†,1,2,3} Victor Borda,^{†,1} Thiago P. Leal,^{†,1,4} Rennan G. Moreira,^{1,5} Andrew W. Bergen,⁶ Fernanda S.G. Kehdy,^{1,7} Isabela Alvim,¹ Marla M. Aquino,¹ Gilderlanio S. Araujo,^{1,8} Nathalia M. Araujo,¹ Vinicius Furlan,^{1,9} Raquel Liboredo,¹ Moara Machado,^{1,10} Wagner C.S. Magalhaes,^{1,11} Lucas A. Michelin,¹ Maíra R. Rodrigues,^{1,12} Fernanda Rodrigues-Soares,^{1,13} Hanaisa P. Sant Anna,^{1,14} Meddly L. Santolalla,¹ Marília O. Scliar,^{1,15} Giordano Soares-Souza,¹ Roxana Zamudio,¹ Camila Zolini,^{1,16,17} Maria Catira Bortolini,¹⁸ Michael Dean,¹⁹ Robert H. Gilman,^{20,21} Heinner Guio,²² Jorge Rocha,^{23,24} Alexandre C. Pereira,²⁵ Mauricio L. Barreto,^{26,27} Bernardo L. Horta,²⁸ Maria F. Lima-Costa,² Sam M. Mbulaiteye,⁶ Stephen J. Chanock,⁶ Sarah A. Tishkoff,²⁹ Meredith Yeager,^{‡,19} and Eduardo Tarazona-Santos^{*,‡,1,17,21,30}

¹Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

²Instituto de Pesquisa Rene Rachou, Fundação Oswaldo Cruz, Belo Horizonte, MG, Brazil

³Center for Research on Genomics and Global Health, National Human Genome Research Institute, Bethesda, MD

⁴Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

⁵Laboratório de Genômica, Centro de Laboratórios Multiusuário (CELAM), ICB, UFMG, Belo Horizonte, MG, Brazil

⁶Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, MD

⁷Laboratório de Hanseníase, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, RJ, Brazil

⁸Laboratório de Genética Humana e Médica, Instituto de Ciências Biológicas, Universidade Federal do Pará – Campus Guamã, Belém, PA, Brazil

⁹Instituto de Ciências Exatas e Tecnológicas, Universidade Federal de Viçosa, Campus UFV-Florestal, Florestal, MG, Brazil

¹⁰Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD

¹¹Núcleo de Ensino e Pesquisas do Instituto Mário Penna – NEP-IMP, Bairro Luxemburgo, Belo Horizonte, MG, Brazil

¹²Department of Genetics and Evolutionary Biology, Biosciences Institute, University of Sao Paulo, Sao Paulo, SP, Brazil

¹³Departamento de Patologia, Genética e Evolução, Instituto de Ciências Biológicas e Naturais, Universidade Federal do Triângulo Mineiro, Uberaba, MG, Brazil

¹⁴Melbourne Integrative Genomics, The University of Melbourne, Melbourne, VIC, Australia

¹⁵Human Genome and Stem Cell Research Center, Biosciences Institute, University of Sao Paulo, Sao Paulo, SP, Brazil

¹⁶Beagle, Belo Horizonte, MG, Brazil

¹⁷Mosaico Translational Genomics Initiative, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

¹⁸Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil ¹⁹Cancer Genomics Research Laboratory, Frederick National Laboratory for Cancer Research, Frederick, MD

²⁰Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD

²¹Universidad Peruana Cayetano Heredia, Lima, Peru

²²Instituto Nacional de Salud, Lima, Peru

²³Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal

²⁴CIBIO/InBIO: Research Center in Biodiversity and Genetic Resources, Vairao, Portugal

²⁵Instituto do Coração, Universidade de Sao Paulo, Sao Paulo, SP, Brazil

²⁶Instituto de Saúde Coletiva, Universidade Federal da Bahia, Salvador, BA, Brazil

²⁷Center of Data and Knowledge Integration for Health (CIDACS), Fundação Oswaldo Cruz (FIOCRUZ), Salvador, Brazil

²⁸Programa de Pós-Graduação em Epidemiologia, Universidade Federal de Pelotas, Pelotas, RS, Brazil

²⁹Department of Genetics and Department of Biology, University of Pennsylvania, Philadelphia, PA

³⁰Instituto de Estudos Avançados Transdisciplinares, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

[†]These authors contributed equally as First Authors.

[‡]These authors contributed equally as Senior Authors.

*Corresponding author: E-mail: edutars@icb.ufmg.br.

Associate editor: Rasmus Nielsen

Abstract

The Transatlantic Slave Trade transported more than 9 million Africans to the Americas between the early 16th and the mid-19th centuries. We performed a genome-wide analysis using 6,267 individuals from 25 populations to infer how different African groups contributed to North-, South-American, and Caribbean populations, in the context of geographic and geopolitical factors, and compared genetic data with demographic history records of the Transatlantic Slave Trade. We observed that West-Central Africa and Western Africa-associated ancestry clusters are more prevalent in northern latitudes of the Americas, whereas the South/East Africa-associated ancestry cluster is more prevalent in southern latitudes of the Americas. This pattern results from geographic and geopolitical factors leading to population differentiation. However, there is a substantial decrease in the between-population differentiation of the African gene pool within the Americas, when compared with the regions of origin from Africa, underscoring the importance of historical factors favoring admixture between individuals with different African origins in the New World. This between-population homogenization in the Americas is consistent with the excess of West-Central Africa ancestry (the most prevalent in the Americas) in the United States and Southeast-Brazil, with respect to historical-demography expectations. We also inferred that in most of the Americas, intercontinental admixture intensification occurred between 1750 and 1850, which correlates strongly with the peak of arrivals from Africa. This study contributes with a population genetics perspective to the ongoing social, cultural, and political debate regarding ancestry, admixture, and the *mestizaje* process in the Americas.

Key words: African diaspora, Transatlantic Slave Trade, admixture dynamics, *mestizaje*.

Introduction

The Transatlantic Slave Trade was an international enterprise involving Brazilian, British, Danish, Dutch, French, German, Portuguese, Spanish, and Swedish traders. They brought over 9 million Africans to the Americas between the early 16th and the mid-19th centuries. African regions of origin included far away locations as Senegambia is from Tanzania. Destiny ports in the Americas were also distant as Boston is from Buenos Aires (Thomas 1999; Eltis 2008; Gomes 2019). The Transatlantic Slave Trade shaped the genetic structure of American continent populations (Alves-Silva et al. 2000; Carvalho-Silva et al. 2001; Salzano and Bortolini 2001; Tishkoff et al. 2009; Bryc et al. 2010; Moreno-Estrada et al. 2013; Campbell et al. 2014; Kehdy et al. 2015; Baharian et al. 2016; Mathias et al. 2016; Rotimi et al. 2016; Ongaro et al., 2019). Although most genetic studies have estimated the overall African ancestry in the Americas, a finer genomic and geographic analysis is needed to infer how different African groups contributed to North-, Central-, South-American, and Caribbean populations and to estimate these contributions. The geopolitical factors that permeated the African Diaspora have been seldom discussed at a continental scale, despite its potential influence on the genetic structure of populations.

Formal integration of genetic and demographic data has historical and solid root of more than 50 years in human population genetics (Cavalli-Sforza et al. 2013), but this kind of analysis has become rare in the era of human population genomics. In particular, a formal comparison of information from demographic history records of the Transatlantic Slave Trade with inferences based on genomic diversity of current populations from Africa and the Americas has not been performed. Here, we perform a joint systematic analysis of genetic data and historical records of the Transatlantic Slave Trade to address the following questions: 1) Is there a

correspondence between the geographic origin of specific African populations of the Diaspora and specific destinations in the Americas?; 2) Was intercontinental admixture dynamics in the Americas associated with the dynamics of arrivals of African slaves?; 3) Considering the geographic extension and the massive demographic magnitude of the African Diaspora, as well as the level of between-populations genetic differentiation in the African regions of origin of slaves, did the Transatlantic Slave Trade lead to a higher, similar or lower level of between-population differentiation of the African gene pool in the Americas?

Results and Discussion

We combined genome-wide data from 25 populations: 9 admixed from the Americas, 11 Africans, 2 Europeans, and

3 Native Americans and created a data set of 6,267 unrelated individuals with >10% of African ancestry (fig. 1A and B, supplementary fig. S2, table S1, and sections S1 and S2, Supplementary Material online). Using ADMIXTURE (Alexander et al. 2009), we identified two continental

(European and Native American) and four African-specific ancestry clusters, named based on their association with geographic regions (supplementary table S1, Supplementary Material online, represented by different colors in fig. 1):

1) West-Central African (blue), 2) Western African (purple), and 3) South/East African (yellow), which are prevalent in the Americas, as well as 4) Northern Ugandan (cyan), which accounts for a very low proportion of African ancestry in the Americas. Hereafter, whereas in African individuals, the proportions of ADMIXTURE ancestry clusters are relative to their whole genome ancestry (fig. 1A, supplementary table S1, Supplementary Material online), in American continent individuals, these proportions are relative to the sum of the four African ancestry clusters (fig. 1B). We also estimated haplotype-

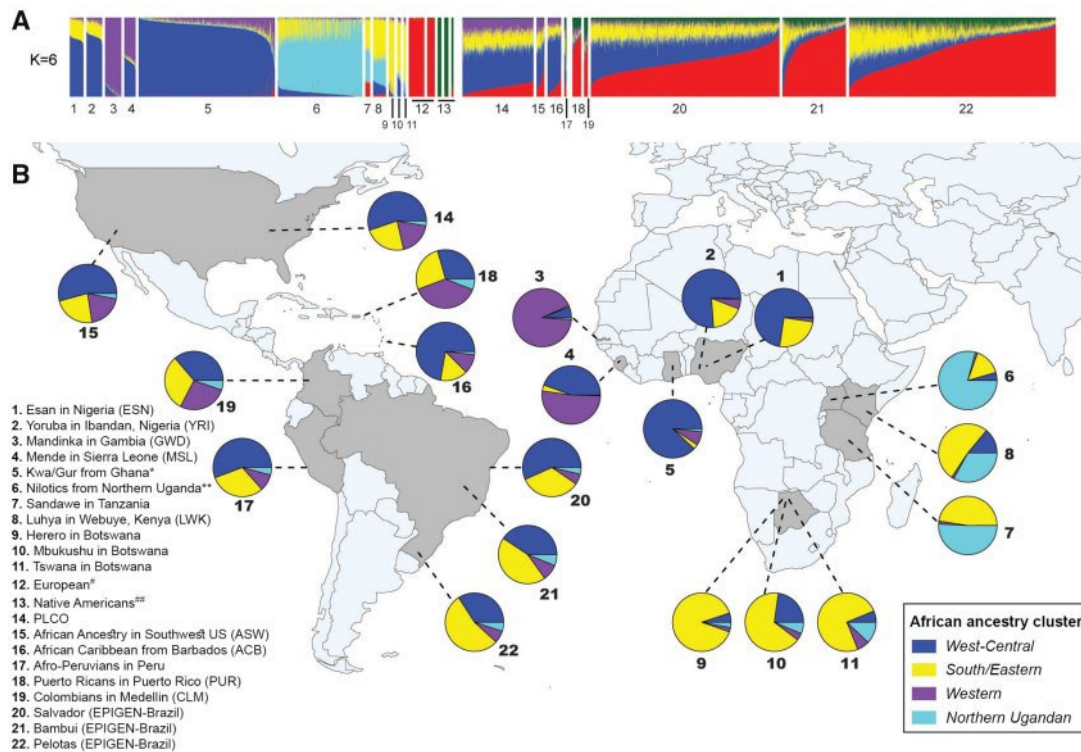


FIG. 1. Ancestry analysis of African and admixed populations of the Americas inferred using ADMIXTURE ($K = 6$). (A) Vertical bar plot showing the total African, European, and Native American proportions of the ancestry clusters (supplementary fig. S1 and section S2.6.1, Supplementary Material online). (B) Percentages of subcontinental African ancestry clusters. For admixed populations of the American continent these percentages are relative to the total African ancestry (i.e., the sum of the four African-associated clusters: West-Central, Western, Southern/Eastern, Northern Ugandan). The arrows on the map represent the regions from where the samples were collected. *The Kwa/Gur data set includes approximately 35 ethno-linguistic groups, predominantly from the Kwa and Gur Niger-Congo linguistic group (Gouveia et al. 2019). **The Nilotic data set includes predominantly three ethno-linguistic groups in Northern Uganda (Langi, Acholi, and Lugbara) from the Nilotic linguistic group (Gouveia et al. 2019); #the Europeans are: Iberian Population in Spain (IBS) and Utah residents with Northern and Western European ancestry (CEU), in this order in the ADMIXTURE bar plot; ###The Native Americans are: Shimaa, Ashaninka, and Aymara, respectively from Borda V et al. (2019); the PLCO (Prostate, Lung, Colorectal, and Ovarian Cancer Screening) data comprised African-Americans from East United States.

proportions from different African regions (Lawson et al. 2012; Hellenthal et al. 2014) relative to the total contribution of African populations (fig. 2A and B, supplementary fig. S3, tables S3 and S4, Supplementary Material online).

Supplementary Material online).

Ancestry Correspondence between African and Admixed American Continent Populations, and the Influence of Geography and Geopolitics

The West-Central Africa-associated ancestry cluster is the most prevalent African cluster in the Americas, including African-Caribbean from Barbados (72% of the total African ancestry), Northeastern Brazilians (57%), Afro-Peruvians (56%), and US African-Americans (54–55%) (blue in fig. 1B, supplementary table S1 and section S2.1, Supplementary Material online). Moreover, haplotype-based analysis (Lawson et al. 2012; Hellenthal et al. 2014) reveals a higher contribution in the Americas from Yoruba-like and Esan-like populations (from Nigeria, mean: 38%) than from Kwa/Gur-like populations (from Ghana, mean: 18%) (fig. 2A and B, supplementary tables S3 and S4,

The Western Africa-associated ancestry cluster has its highest proportions in Puerto Ricans (38% of the total African ancestry), Colombians (27%), and US African-Americans (19–20%, purple in [fig. 1B](#), [supplementary table S1](#), [Supplementary Material](#) online), whereas Brazilians have the lowest proportion (<9%), limited to a Mandinka-like (Gambia) contribution and with no Mende-like (Sierra Leone) contribution ([fig. 2A and B](#), [supplementary tables S3 and S4](#), [Supplementary Material](#) online).

The South/East Africa-associated ancestry cluster, in contrast, shows its highest proportion in South and Southeast Brazil (44% and 54% of total African ancestry, respectively) (yellow in [fig. 1B](#), [supplementary table S1](#), [Supplementary Material](#) online). Haplotype-based methods ([Lawson et al. 2012](#); [Hellenthal et al. 2014](#)) identified two different sources of gene flow associated with the South/Eastern Africa ancestry cluster: one from Mbukushu-like populations (Botswana, Western Bantu speakers from Southern Africa, 20–24% to South/Southeast Brazil) and one from Luhya-like populations (Kenya, Eastern Bantu speakers from Eastern Africa, 17–20% to South/Southeast Brazil, [fig. 2A and B](#), [supplementary tables S3 and S4](#), [Supplementary Material](#) online). Western- and Eastern-Bantu speakers historically correspond to the two

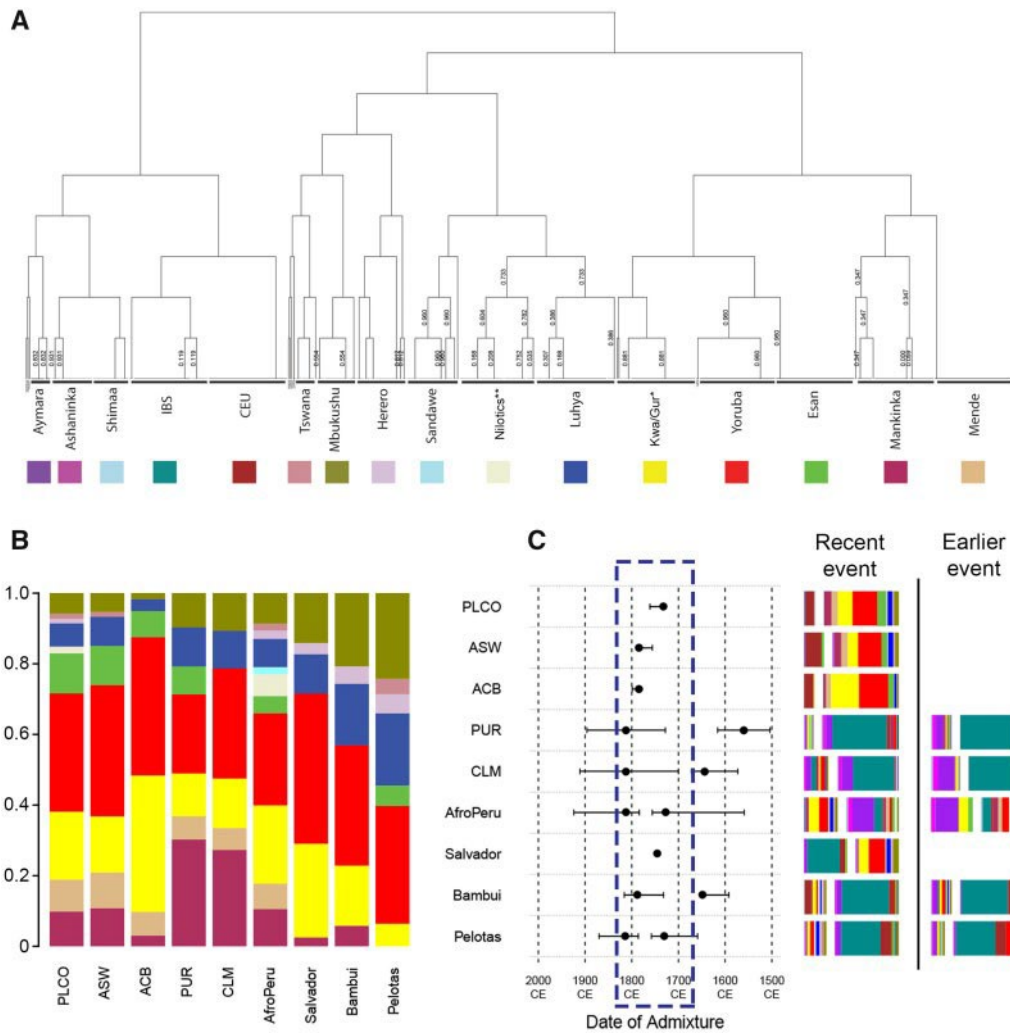


FIG. 2. Haplotype-based clustering of parental individuals and admixture inferences for admixed American continent populations. (A) fineSTRUCTURE tree of parental individuals. *The Kwa/Gur data set includes approximately 35 ethno-linguistic groups, predominantly from the Kwa and Gur linguistic group (Gouveia et al. 2019). **The Nilotics data set includes predominantly three ethno-linguistic groups from Northern Uganda (Langi, Acholi, and Lugbara) of the Nilotic linguistic group (Gouveia et al. 2019). (B) Subcontinental contributions relative to the total African ancestry in admixed populations inferred by the MIXTURE MODEL (supplementary section S2.3, Supplementary Material online). (C) GLOBETROTTER inference of admixture events for each admixed population. Inferred date(s) and 95% confidence intervals are represented by dots and horizontal lines in the graph. Dashed rectangle in the admixture dates plots highlights the most dynamic period for admixture. Beside the dating graph, we represented the inferred admixing sources (bars) for recent and earlier events. Bar size represents the genetic contribution of the source. Each color corresponds to the proportion of each parental population contribution. 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.

streams of the Bantu migrations in the last 4,000–2,500 years (Tishkoff et al. 2009; Busby et al. 2016; Patin et al. 2017).

Pioneering mitochondrial DNA studies of the first decade of this century showed how African Bantu-associated haplotypes were more frequent in South America than in Central- and North-America (Salas et al. 2004; Hünemeier et al. 2007; Gonçalves et al. 2008). However, our approach, based on a genome-wide data set and a larger number of studied individuals, allows for finer geographic inferences and estimates of genome-wide admixture proportions from the different African regions, adding new layers of knowledge to our understanding of the African Diaspora.

1650

This emerging portrait of the African ancestry in the Americas

suggests an influence of geography and geopolitics. Geographical factors include: 1) the latitudinal proximity between Western Africa and Caribe-Central/North America, as well as between South/East Africa and Southern Brazil, 2) the winds and ocean currents, that shaped two navigation systems: the North-Atlantic, with voyages mostly to North America, and the South-Atlantic, with voyages predominantly to Brazil (Domingues da Silva 2008). Indeed, West-Central Africa- and Western Africa-associated ancestry clusters are more commonly observed in northern latitudes, whereas the South/East Africa-associated ancestry cluster is more evident in southern latitudes.

Differently, the Portugal possessions in the Americas (Brazil) and its influence in South and East African coasts (current Angola and Mozambique) (Klein 1987) exemplify the geopolitical factors that affected, in particular, the distribution of the South/East Africa-associated ancestry cluster. Although the Portuguese Crown had earlier privileged relations with the kingdoms of Benin in nowadays Nigeria, it later extended its influence to Bantu-speaking areas such as Congo/Angola and Mozambique (Coelho et al. 2009). Indeed, Portuguese–Brazilian slave trade routes departed from Luanda and Cabinda (Angola) and from Zanzibar (Tanzania) and Inhambane (Mozambique) during 18th and 19th centuries (Eltis 2008). The abolition of slavery by the British in 1807, who controlled the North Atlantic route, also led Portuguese traders to prefer routes in the South Atlantic (Versiani 2008). Therefore, geography (intercontinental distances and climatic factors affecting transatlantic navigation) and geopolitics (European colonial influences and possessions) influenced the geographic and linguistic diversity of African emigrants as well as favored the regional differentiation of African ancestry in the Americas.

African-Specific Genetic Distance (ASGD), see Materials and Methods section, fig. 3, supplementary section S4, and fig. S5, Supplementary Material online), is observed between African populations (mean: 0.057, mean

The Dynamics of African Admixture in the Americas with Europeans and Native Americans Accompanied the Dynamics of Arrivals of African Slaves

Remarkably, linkage-disequilibrium-based inference (Hellenthal et al. 2014) shows that all the studied admixed populations of the Americas exhibit the signature of an intensification of intercontinental admixture in the interval from 1750 to 1850 (fig. 2C, supplementary table S5 and section S3, Supplementary Material online), revealing a continental trend. This trend is consistent with results by Baharian et al. (2016), focused in the United States and by Fortes-Lima et al. (2017) focused on French Guiana and Suriname isolated populations and on Colombia and Rio de Janeiro. Importantly, this time interval matches or is immediately subsequent to regional peaks of number of slaves arriving from Africa to United States, Barbados, Puerto Rico, and Brazil (supplementary fig. S4, Supplementary Material online). Thus, we reveal that in most of the Americas, the arrival of the largest contingent of Africans between 1700 and 1850 (supplementary fig. S4, Supplementary Material online) was almost synchronic with intensive intercontinental admixture, a process that was also characterized by positive ancestry-based assortative mating (Kehdy et al. 2015).

The African Gene Pool Is More Homogenous Between-Populations in the Americas Than in Africa

Figure 1B suggests that African ancestry clusters are more homogeneously distributed between admixed American continent populations than between the African populations that contributed to the Transatlantic Slave Trade. Considering only the African gene pool, the largest differentiation, measured by the

excluding populations with marginal contribution to the Americas [Nilotics and Sandawe: 0.53]), followed by differentiation between African versus America's populations (mean: 0.043) and between populations of the Americas (mean:

0.018, 32% of the *ASGD* between African populations) (Wilcoxon test, $P < 10^{-6}$ for the three pairwise comparisons, [fig. 3A](#)).

Corroborating these results, our approach based on local ancestry (see Materials and Methods and supplementary section S4.2, [Supplementary Material](#) online) showed

that: 1) the *between-populations* differentiation of the African gene pool in the American continent populations (single-nucleotide polymorphisms, SNPs mean $F_{ST} = 0.02$) is two-thirds of the value observed between the African pop-

ulations that contributed to the African Diaspora (i.e., excluding the Nilotics, SNPs mean $F_{ST} = 0.03$, $P < 10^{-16}$ for comparison between the distributions, [fig. 3C](#)); and 2) the within-population genetic diversity (i.e., mean heterozygosity across SNPs) is not lower in the African segments of American

continent population than in African populations (Kruskal-Wallis $P = 0.53$, [fig. 3C](#)). Thus, on average, chromosomal fragments of African origin from different populations are more similar in the Americas than in Africa, despite the very similar within-population African genetic diversity in the Americas and Africa ([fig. 3C](#)).

To better understand this pattern of *between-populations* homogenization of the African gene pool in the Americas we compared: 1) proportions of West-Central Africa-, Western Africa-, and South-East Africa-associated ADMIXTURE ancestry clusters ([fig. 1](#)) with 2) expected proportion of these ancestry clusters, estimated considering both the proportions of arrivals from different African locations ([fig. 4](#), [supplementary tables S6 and S7](#), section S5, [Supplementary Material](#) online) and the ADMIXTURE ancestry clusters composition of those African locations of the origin of the Diaspora. We performed these comparisons for the geographic regions represented in our data set for which there are also historical demography records of origin and destination of Africans ([Eltis 2008](#)).

Although we recognize that an accurate estimation of the expected proportions of ancestry based on the number of disembarks from different regions from Africa is a complex task, here we formally attempt to integrate demographic and genetics data from the African populations of origin of the Diaspora to obtain such estimation. The assumptions of our

approach are shared by several population genetics methods:

1) that the current ancestry compositions of African populations are good proxies of real sources of the African Diaspora located in the same geographic areas, and 2), that the migration from Africa to the Americas occurred in a unique migration event.

Overall, for New World admixed populations, the proportions of South-Eastern African and Western African ancestry clusters are highly correlated with the expected ancestry based on the numbers of arrivals to Americas ports and departures from African ports (Spearman $\rho = 0.89$, $P = 0.02$). However, for the West-Central African ancestry cluster the correlation does not reach significance ([fig. 4](#)). For the entire American continent, we observe an excess of the observed individual proportions of West-Central Africa

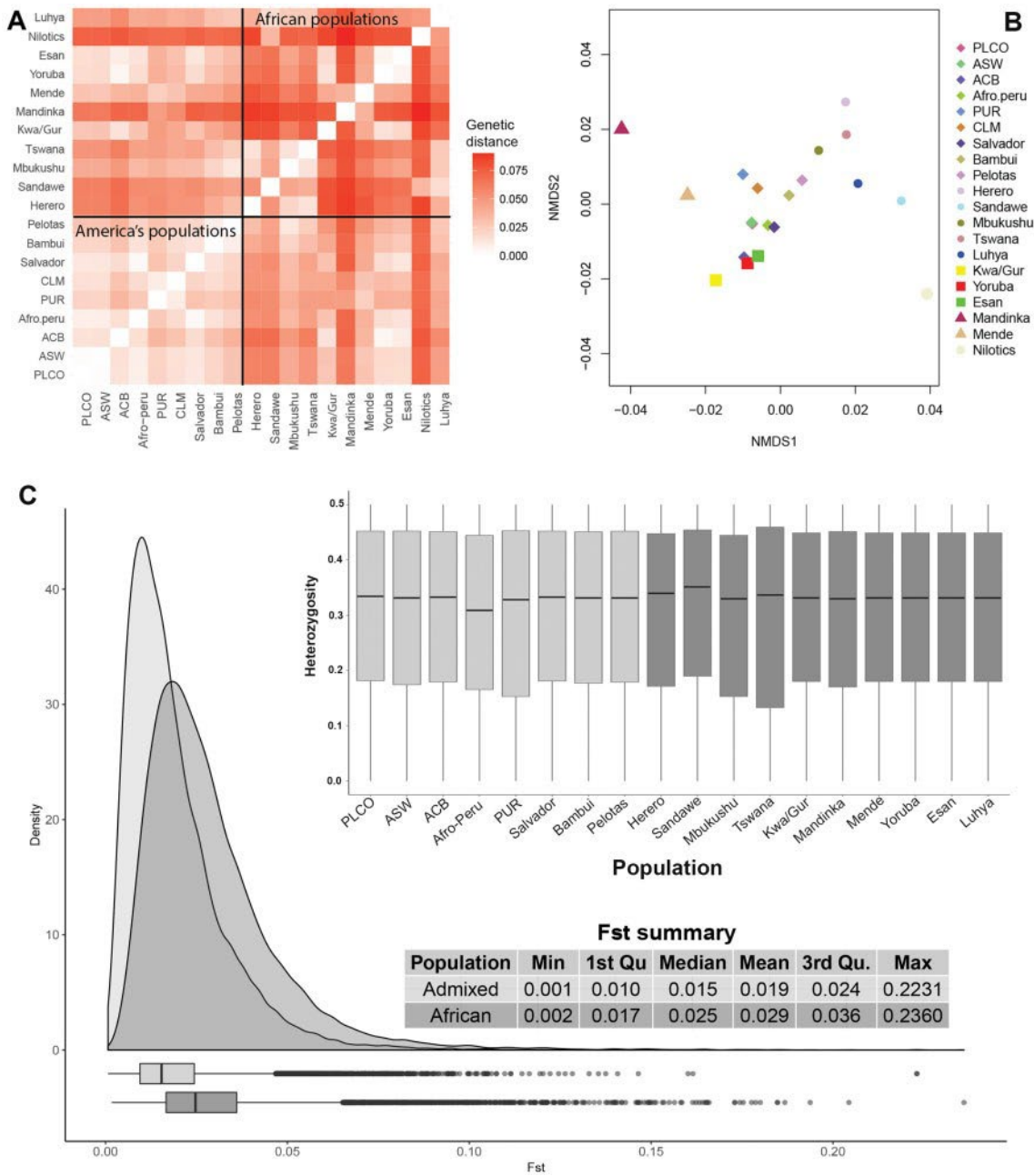


FIG. 3. Pairwise genetic distances of the African gene pool between populations of the American continent and Africa. (A) Heatmap Matrix and (B) multidimensional scaling of the African gene pool genetic distances. We used solid squares, triangles, and circles to represent populations associated with WCA, West-Central Africa; SEA, South/East Africa; WA, Western Africa ancestry clusters. CLM, Colombians from Medellin; PUR, Puerto Ricans from Puerto Rico; ACB, African Caribbeans in Barbados; ASW, Americans of African ancestry in South western United States; PLCO, African-Americans from Eastern United States. (C) SNPs F_{ST} distributions between: 1) African populations that contributed to the African Diaspora (dark gray) and 2) American continent populations (gray), considering only chromosome fragments of African origin; and the within-population African genetic heterozygosity in the Americas and Africa. The CLM population was not included in this analysis because it did not have enough SNPs inferred as being of African origin.

ancestry cluster (47.7% observed vs. 40% expected, being this a conservative estimation of the difference, supplementary section S5, [Supplementary Material online](#), $P < 2.2 \times 10^{-16}$), mainly determined by Southeastern Brazil and the US populations ([supplementary table S8](#), [Supplementary Material online](#)). The poorest concordance between observed and expected ancestries is observed in

Southeastern Brazil, that presents more of the West-Central African ancestry cluster (37%) than expected (20%) ($P < 2.2 \times 10^{-16}$) and complementarily, less of the South/East African ancestry cluster than expected based on arrivals (55% observed vs. 76% expected, $P < 2.2 \times 10^{-16}$). The US population also shows an excess of the West-Central African ancestry cluster (54.7% observed vs. 43.1% expected, $P < 3.33 \times 10^{-16}$),

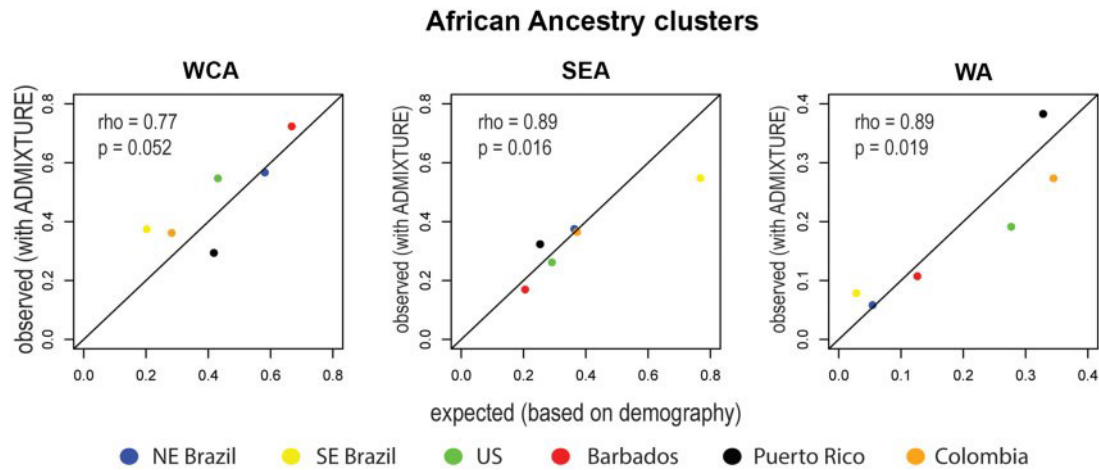


FIG. 4. Observed and expected proportions of genomic African ancestry clusters in the Americas. We compared 1) the observed proportions of genomic African ancestry clusters (inferred using ADMIXTURE [Alexander et al. 2009]) in the vertical axis, with 2) expected proportions of genomic African ancestry clusters, estimated based on demographic historical records from the African Voyages Database1, in the horizontal axis (see supplementary table S3, Supplementary Material online). ρ , Spearman's coefficients of correlation; p , p value significance. The significance was evaluated using randomization tests of 10,000 replications. WCA, West-Central Africa; SEA, South/East Africa; WA, Western Africa.

compensated by a deficit of the Western African ancestry cluster (19.1% observed vs. 27.7% expected, $P < 3.33 \cdot 10^{-16}$). Therefore, the *between-population* homogenization of the African gene pool in the Americas is partly explained by the excess of the West-Central Africa ancestry cluster in Southeast Brazil and in the United States (fig. 4). The higher *between-population* homogeneity of the African gene pool in the Americas by reducing population stratification, contributes to a more statistical power of genetic association studies involving individuals with African ancestry from different populations of the Americas.

In general, the limitations of our study derive from: 1) the smaller sample sizes of the non-Brazilian samples with respect to Brazilians, except for the United States, from where we included a fair number of Afro-Americans ($n = 524$) from the PLCO cohort; 2) the lack of a Central America sample; 3) the use of SNP-array data that contain an ascertainment bias. Even considering these limitations, our results are based on observed general patterns and not on results based on a specific population and our *within-* and *between-populations* estimates of genetic diversity are consistently calculated from genomic fragments of African origin, and therefore, the possible effect of the ascertainment bias is the same in African and American continent populations.

In conclusion, genetic data trace the African genetic roots of admixed individuals of the Americas to a broad geographic extension (from Western Africa to East Africa), associated with a high linguistic diversity (Niger Kordofanian non-Bantu and Western- and Eastern-Bantu language speakers). Considering the level of *between-populations* genetic differentiation in the African regions of origin of slaves, historical facts that homogenized the *between-populations* component of genetic diversity in the Americas have predominated over facts that tend to maintain or increase it. This latter group of facts includes geographic (i.e., intercontinental distances and maritime winds/currents) and geopolitical

factors (i.e., specific European colonial influences and possessions and the

abolition of the slavery by British in 1807), that shaped an association of Western African ancestry with northern latitudes and South/East African ancestry with southern latitudes. Contrastingly, the following combination of facts, that occurred in Africa and the Americas, associated with the African Diaspora, have contributed to gene flow between individuals with different African ancestries and therefore, to the *between-populations* homogenization of the African gene pool in the Americas: 1) the heterogeneous contribution via the Transatlantic Slave Trade of the different African regions to the Americas, 2) despite their specific European origins, traders/vessels transported slaves, frequently illegally, to different American continent ports (Klein 1987, 2010; Eltis 2008); 3) *forced amalgamation*, which is the preference of slave owners for slaves from different geographic and linguistic origins, so that they could not understand each other and thus, reducing the risk of riots (Olcott 1838); and 4) the role of islands in the Americas such as Jamaica and Barbados, which centralized parts of arrivals of African slaves and redistributed them to different parts of the Americas (Thomas 1999) and also ports/islands in Africa with similar roles. Other factors that contributed to the between-population homogenization of the African gene pool may be related to more general demographic trends of admixed populations of the Americas, and are not necessarily and specifically related to the African Diaspora. Importantly, by combining genetic and demographic data, we show that the *between-population* homogenization of the African gene pool in the Americas is partly explained by the excess of the West-Central Africa ancestry cluster (the most prevalent in the Americas) in the United States and Southeast Brazil with respect to demographic expectations, which suggests a spread of this ancestry in the American continent. Interestingly, in most of the Americas, the arrival of the largest contingent of Africans between 1700 and 1850 was almost synchronic with the intensification of intercontinental admixture, which implies that this time interval was critical to shape the structure of

the African gene pool in the New World. This study, by dissecting and estimating the African ancestry proportions in different populations of the Americas, and inferring the dynamics of biological admixture, contributes with a population genetics perspective to the ongoing social, cultural and political debate regarding ancestry, admixture, and *mestizaje* and the different perceptions of *race* in the Americas (Clinton 2001; Wade et al. 2014).

Materials and Methods

Database and Population Structure Analyses

We analyzed a final data set of 6,267 unrelated individuals from Africa and the Americas (with more than 10% of African ancestry) for 533,242 SNPs (supplementary table S1, Supplementary Material online). We inferred population structure and admixture using ADMIXTURE (Alexander et al. 2009) and Principal Component Analysis (Price et al. 2006) for unlinked SNPs. Presented results are based on ADMIXTURE runs with $K = 6$ because it corresponds to the lower cross-validation error. We inferred haplotypes using the SHAPEIT2 software (Delaneau et al. 2012). Haplotype-based analyses were performed using ChromoPainter and fineSTRUCTURE (Lawson et al. 2012). The admixture contributions from the different African regions were inferred using GLOBETROTTER (Hellenthal et al. 2014). Demographic information of embarked and disembarked African slaves was obtained from the African Voyages database (<https://www.slavevoyages.org/>; last accessed February 20, 2020).

African-Ancestry Genetic Distance

The genetic differentiation between populations considering only the African gene pool was estimated using two strategies. First, we conceived the *African-ancestry genetic distance* (AAGD, supplementary section S4.1, Supplementary Material online), based on: 1) the mean proportions of the subcontinental African ancestry clusters from each population based on ADMIXTURE results ($K = 6$) (supplementary table S1, Supplementary Material online). In the case of the population of the Americas, these proportions were with respect to the total African ancestry. 2) The F_{ST} between the African ancestry clusters estimated by the ADMIXTURE software (Alexander et al. 2009) (supplementary table S2, Supplementary Material online). AAGD between two populations is given by the sum of the Euclidean genetic distances between each pair of subcontinental ancestries weighted by the F_{ST} (in sensu ADMIXTURE [Alexander et al. 2009]) between the ancestry clusters. Specifically, considering two populations (A and B) with C ancestry clusters (c_1, c_2 ; and c_3), the African-ancestry genetic distance is calculated as:

AAGD_{A;B}

$$\frac{1}{4} \sum_{x/y; x,y < C} F_{ST_{x,y}} \left(\delta A_x - B_x b^2 \right)^2 + \left(\delta A_y - B_y b^2 \right)^2$$

Our second strategy (supplementary section S4.1, Supplementary Material online) to measure the genetic differentiation between populations considering only the African gene pool (from chromosome fragments of African origins), consisted in comparing the distribution of continental SNPs- F_{ST} , estimated as (Wright 1943, 1949) (supplementary section S4.1, Supplementary Material online):

$$F_{st} = \frac{\text{var}(\delta p)}{p\delta 1 - p^2}$$

where A_c and B_c are the ancestry proportions of the ADMIXTURE cluster c in the populations A and B, with respect to the total African ancestry.

where p is the minor-allele frequency of the SNP i in the population j , p is a vector with allele frequencies of an SNP for all the considered populations, $\text{var}(p)$ denotes the between-population variance of p_i and \bar{p} denotes the mean of p_i across populations. Allele frequencies in the African populations were those that, on the basis of our results, contributed to the African Diaspora (i.e., conservatively excluding Nilotics). For the American continent populations, allele frequencies of the African gene pool were estimated from a minimum of 20 chromosome fragments of African origin, as inferred using RFMix (Maples et al. 2013). Analogously, within-population diversity for the African gene pool was estimated by the i -SNP-heterozygosity for the j -population, as:

$$h_{ij} = 2p_{ij}(1-p_{ij})$$

Estimating the Expected Proportions of African Ancestry Clusters Based on Demography

We used data available in the African Voyages database (Eltis 2008) to estimate the expected ancestry in a specific destiny proxy of the Americas, by considering the proportion of

individuals from each embarkation major region (representing the ancestry origin proxies), that arrived in specific ports of disembarkation in the Americas (supplementary table S7, Supplementary Material online). To avoid the unrealistic assumption that the individuals from the African embarkation major regions have a homogeneous ancestry, we calculated the weighted expected ancestry. This was estimated by taking into account the proportion of WCA, WA, SEA genomic ancestry clusters estimated in selected current African populations from the embarkation major regions that contributed to the African genomic pool in the Americas (supplementary table S7, Supplementary Material online and figs. 1A and 2B): 1) Kwa/Gur and Yoruba populations for the WCA proportion; 2) Mandinka (GWD) and Mende (MSL) for the WA proportion; and 3) Mbukushu and Luhya (LWK) for the SEA proportion. Exception were the Brazilian populations, in which we used only GWD population to obtain the WA proportion, since the Mende (MSL) did show contribution to the Brazilian populations (fig. 2A and B, supplementary tables S3 and S4, Supplementary Material online).

Thus, we calculated the weighted expected proportion of the i -ancestry cluster (W_i) for each proxy-destination as:

$$W_i = \frac{1}{n} \sum_{p=1}^n e_p \times o_{p,i};$$

where n is the number of African population proxies of origin (p); e_p is the expected ancestry of the population proxies of origin p based on the proportion of individuals arrived from each of the n African populations, with respect to the total of individuals disembarked in the population; $o_{p,i}$ is the observed proportion of the ADMIXTURE ancestry cluster i of population proxies of origin p . To assess the correlation between observed and expected ancestries, we applied the Spearman correlation test, implemented in the *R* and the significance was evaluated using 10,000 randomization tests (supplementary section S5, [Supplementary Material](#) online).

Flowcharts of the performed analyses are available in the EPIGEN Scientific Workflow ([Magalhães et al. 2018](#)) website (<http://ldgh.com.br/scientificworkflow>; last accessed February 20, 2020). Masterscripts are available under request from the authors for academic purposes. Details of Materials and Methods section are in the [Supplementary information](#).

Data Availability

EPIGEN-Brazil data are deposited at the European Nucleotide Archive (PRJEB9080 [ERP010139]), accession number EGAS00001001245, under EPIGEN Committee Controlled Access mode. The Nilotics and Kwa/Gur data sets are deposited in dbGaP at phs001705.v1.p1 and phs000838.v1.p1, respectively. The Botswana and Tanzania data sets from Sarah Tishkoff Lab are available at dbGaP accession number phs001396.v1.p1 and SRA BioProject PRJNA392485.

Supplementary Material

[Supplementary data](#) are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Sergio D. Pena, Marcia Beltrame, Rosangela Loschi, Eduardo F. Paiva, Fabricio Santos, Renan Souza, Claudio Struchiner, Ricardo Santos, and Garrett Hellenthal for advice, discussions and criticisms. This work was supported by the Brazilian Ministry of Health (Department of Science and Technology from the Secretaria de Ciência, Tecnologia e Insumos Estratégicos) through Financiadora de Estudos e Projetos (FINEP) to the EPIGEN-Brazil Initiative. The EPIGEN-Brazil investigators were also supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior of the Brazilian Ministry of Education (CAPES Agency). E.T.S., M.H.G., V.B., T.P.L., and M.F.L.C. were supported by Brazilian National Research Council (CNPq), Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG), and Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais. M.H.G. performed part of this study as CAPES-PDSE fellow, V.B. was a CAPES-PEC-PG fellow. M.L.S. was a TWAS-CNPq PhD fellow. MHG is supported by the Intramural Research Program of the National Institutes of Health in the

Digestive and Kidney Diseases, the Center for Information Technology, and the Office of the Director at the Center for Research on Genomics and Global Health (CRGGH). The CRGGH is supported by the National Human Genome Research Institute, the National Institute of Diabetes and

Institutes of Health (1ZIAHG200362). Tishkoff Laboratory is funded by the National Institutes of Health (1R01DK104339-0 and 1R01GM113657-01). EMBLEM is funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI) (HHSN261201100063C and HHSN261201100007I) and, in part, by the Intramural Research Program, National Institute of Allergy, and Infectious Diseases (SJR), National Institutes of Health, Department of Health and Human Services. Bioinformatics support was provided by the Sagarana HPC cluster, CPAD- ICB-UFMG, Brazil.

- C, Tishkoff SA, et al. 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A*. 107(2):786–791.
- Busby GB, Band G, Si Le Q, Jallow M, Bougama E, Mangano VD, Amenga- Etego LN, Enimil A, Apinjoh T, Ndila CM, et al. 2016. Admixture into and within sub-Saharan Africa. *Elife* 5:pii:e15266. [Internet]
- Campbell MC, Hirbo JB, Townsend JP, Tishkoff SA. 2014. The peopling of the African continent and the diaspora into the new world. *Curr Opin Genet Dev*. 29:120–132.
- Carvalho-Silva DR, Santos FR, Rocha J, Pena SD. 2001. The phylogeography of Brazilian Y-chromosome lineages. *Am J Hum Genet*. 68(1):281–286.
- Cavalli-Sforza LL, Moroni A, Zei G. 2013. Consanguinity, inbreeding, and genetic drift in Italy (MPB-39). Oxford, UK: Princeton University Press. Available from: <http://dx.doi.org/10.1515/9781400847273>

1655 146

Author Contributions

The project was conceived by M.H.G. and E.T.S. M.H.G. assembled data sets. M.H.G., V.B., T.P.L., R.G.M., M.M.A., G.S.A., N.M.A., F.S.G.K., M.M., W.C.S.M., L.A.M., M.R.R., F.R.-S., H.P.S.A., M.L.S., M.O.S., G.S.S., C.Z. analyzed genetic data. R.L. and R.Z. performed laboratory experiments. E.T.S. supervised bioinformatic and statistical analyses. M.D., R.H.G., H.G., A.C.P., M.F.L.C., M.L.B., B.L.H., S.M.M., S.J.C., S.A.T., and M.Y. contributed with data. M.H.G., M.C.B., V.B., A.W.B., M.Y., S.A.T. contributed to data interpretation. M.H.G., V.B., and E.T.S. wrote the manuscript. All authors read the manuscripts and contributed with suggestions.

References

- Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 19(9):1655–1664.
- Alves-Silva J, da Silva Santos M, Guimaraes PE, Ferreira AC, Bandelt HJ, Pena SD, Prado VF. 2000. The ancestry of Brazilian mtDNA lineages. *Am J Hum Genet*. 67(2):444–461.
- Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, Bustamante CD, Kenny EE, Williams SM, Aldrich MC, et al. 2016. The Great Migration and African-American genomic diversity. *PLoS Genet*. 12(5):e1006059.
- Borda V, Alvim I, Aquino MM, Silva C, Soares-Souza GB, Leal TP, Scliar MO, Zamudio R, Zolini C, Padilla C, et al. 2020. The genetic structure and adaptation of Andean highlanders and Amazonian dwellers is influenced by the interplay between geography and culture. *bioRxiv* [Internet]:2020.01.30.916270. Available from: <https://www.biorxiv.org/content/10.1101/2020.01.30.916270v2>, last accessed February 20, 2020.
- Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo J-M, Wambebe

- Clinton WJ. 2001. Erasing America's color lines. The New York Times [Internet]. [cited 2019 Feb 19]; Section 4, Page 17. Available from: <https://www.nytimes.com/2001/01/14/opinion/erasing-america-s-color-lines.html>. Accessed February 20, 2020.
- Coelho M, Sequeira F, Luiselli D, Beleza S, Rocha J. 2009. On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. *BMC Evol Biol*. 9:80.
- Delaneau O, Marchini J, Zagury J-F. 2012. A linear complexity phasing method for thousands of genomes. *Nat Methods* 9(2):179–181.
- Domingues da Silva DB. 2008. The Atlantic slave trade to Maranhão, 1680–1846: volume, routes and organisation. *Slavery Abol*. 29(4):477–501.
- Eltis D. 2008. A brief overview of the trans-Atlantic slave trade. Voyages: the trans-Atlantic slave trade database: <http://www.slavevoyages.org> [Internet]. [cited 2019 Feb 19]; 1:1-11. Available from: <http://www.redemaosdadas.org/wp-content/uploads/2014/02/HIST211-1.3.3-TransAtlanticSlaveTrade.pdf>. Accessed February 20, 2020.
- Fortes-Lima C, Gessain A, Ruiz-Linares A, Bortolini M-C, Migot-Nabias F, Bellis G, Moreno-Mayar JV, Restrepo BN, Rojas W, Avendaño-Tamayo E, et al. 2017. Genome-wide ancestry and demographic history of African-descendant Maroon communities from French Guiana and Suriname. *Am J Hum Genet*. 101(5):725–736.
- Gomes L. 2019. Escravidão—Vol. 1: do primeiro leilão de cativos em Portugal até a morte de Zumbi dos Palmares. Rio de Janeiro: *Globo Livros*.
- Gonçalves VF, Carvalho CMB, Bortolini MC, Bydlowski SP, Pena S. 2008. The phylogeography of African Brazilians. *Hum Hered*. 65(1):23–32. Gouveia MH, Bergen AW, Borda V, Nunes K, Leal TP, Ogwang MD, Yeboah ED, Mensah JE, Kinyera T, Otim I, et al. 2019. Genetic signatures of gene flow and malaria-driven natural selection in sub-Saharan populations of the “endemic Burkitt Lymphoma belt”. *PLoS Genet*. 15(3):e1008027.
- Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. *Science* 343(6172):747–751.
- Hünemeier T, Carvalho C, Marrero AR, Salzano FM, Pena SDJ, Bortolini MC. 2007. Niger-Congo speaking populations and the formation of the Brazilian gene pool: mtDNA and Y-chromosome data. *Am J Phys Anthropol*. 133(2):854–867.
- Kehdy FSG, Gouveia MH, Machado M, Magalhães WCS, Horimoto AR, Horta BL, Moreira RG, Leal TP, Scliar MO, Soares-Souza GB, et al. 2015. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proc Natl Acad Sci U S A*. 112(28):8696–8701.
- Klein HS. 1987. A demografia do tráfico atlântico de escravos para o Brasil. *Estud Econ*. 17:129–149.
- Klein HS. 2010. The Atlantic slave trade. Cambridge: Cambridge University Press.
- Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. *PLoS Genet*. 8(1):e1002453.
- Magalhães WCS, Araujo NM, Leal TP, Araujo GS, Viriato PJS, Kehdy FS, Costa GN, Barreto ML, Horta BL, Lima-Costa MF, et al. 2018. EPIGEN- Brazil initiative resources: a Latin American imputation panel and the scientific workflow. *Genome Res*. 28(7):1090–1095.
- Maples B, Gravel S, Kenny E, Bustamante C. 2013. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am J Hum Genet*. 93(2):278–288.
- Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, Vergara C, Torgerson DG, Pino-Yanes M, Shringarpure SS, et al. 2016. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. *Nat Commun*. 7:12522.
- Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Ortiz-Tello PA, Martínez RJ, Hedges DJ, Morris RW, et al. 2013. Reconstructing the population genetic history of the Caribbean. *PLoS Genet*. 9(11):e1003925.
- Olcott C. 1838. Two lectures on the subjects of slavery and abolition. Massillon: Massillon, Ohio.
- Ongaro L, Scliar MO, Flores R, Raveane A, Marnetto D, Sarno S, Gnecci-Ruscione GA, Alarcón-Riquelme ME, Patin E, Wangkumhang P, et al. 2019. The Genomic Impact of European Colonization of the Americas. *Curr Biol*. 29(23):3974–3986.e4.
- Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, Laval G, Perry GH, Barreiro LB, Froment A, et al. 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* 356(6337):543–546.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 38(8):904–909.
- Rotimi CN, Tekola-Ayele F, Baker JL, Shriener D. 2016. The African diaspora: history, adaptation and health. *Curr Opin Genet Dev*. 41:77–84. Salas A, Richards M, Lareu M-V, Scozzari R, Coppa A, Torroni A, Macaulay V, Carracedo A. 2004. The African diaspora: mitochondrial DNA and the Atlantic slave trade. *Am J Hum Genet*. 74(3):454–465.
- Salzano FM, Bortolini MC. 2001. The evolution and genetics of Latin American populations by Francisco M. Salzano. Cambridge: Cambridge University Press.
- Thomas H. 1999. The slave trade: the story of the Atlantic slave trade: 1440–1870. New York: Simon and Schuster Paperbacks.
- Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, et al. 2009. The genetic structure and history of Africans and African Americans. *Science* 324(5930):1035–1044.
- Versiani FR. 2008. D. João VI e a (não) abolição do tráfico de escravos para o Brasil. In: Committee of IX BRASA, editors. Brasa IX Proceedings. New Orleans: Brazilian Studies Association. p. 27–29.
- Wade P, Lopez-Beltran C, Restrepo E, Ventura-Santos R. 2014. Mestizo genomics. North Carolina: Duke University Press.
- Wright S. 1943. Isolation by distance. *Genetics* 28:114–138.
- Wright S. 1949. The genetical structure of populations. *Ann Eugen*. 15(1):323–354.

165
6

147



Short Communication
COVID-19 – Special Issue

Human-SARS-CoV-2 interactome and human genetic diversity: *TMPRSS2*-rs2070788, associated with severe influenza, and its population genetics caveats in Native Americans

Fernanda S.G. Kehdy¹, Murilo Pita-Oliveira², Mariana M. Scudeler², Sabrina Torres-Loureiro², Camila Zolini^{3,4}, Rennan Moreira³, Lucas A. Michelin³, Isabela Alvim³, Carolina Silva-Carvalho³, Vinicius C. Furlan³, Marla M. Aquino³, Meddly L. Santolalla⁵, Victor Borda⁶, Giordano B. Soares-Souza³, Luis Jaramillo-Valverde⁷, Andres Vasquez-Dominguez⁷, Cesar Sanchez Neira⁸, Renato S. Aguiar³, Ricardo A. Verdugo^{9,10}, Timothy D. O'Connor^{11,12,13}, Heinner Guio^{8,14} □, Eduardo Tarazona-Santos³, Thiago P. Leal^{3,*} and Fernanda Rodrigues-Soares^{2,*} □

¹Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Laboratório de Hanseníase, Rio de Janeiro, RJ, Brazil.

²Universidade Federal do Triângulo Mineiro, Instituto de Ciências Biológicas e Naturais, Departamento de Patologia, Genética e Evolução, Uberaba, MG, Brazil.

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

⁴Mosaico Translational Genomics Initiative, Belo Horizonte, MG, Brazil.

⁵Universidad Peruana Cayetano Heredia, School of Public Health and Administration, Emerging Diseases and Climate Change Research Unit, Lima, Peru.

⁶Laboratório Nacional de Computação Científica (LNCC), Laboratório de Bioinformática, Petrópolis, RJ, Brazil.

⁷INBIOMEDIC Research and Technological Center, Lima, Peru.

⁸Instituto Nacional de Salud, Lima, Peru.

⁹Universidad de Chile, Facultad de Medicina, Instituto de Ciencias Biomédicas, Programa de Genética Humana, Santiago, Chile.

¹⁰Universidad de Chile, Facultad de Medicina, Departamento de Oncología Básico Clínica, Santiago, Chile.

¹¹University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, United States.

¹²University of Maryland School of Medicine, Program in Personalized and Genomic Medicine, Baltimore, United States.

¹³University of Maryland School of Medicine, Department of Medicine, Baltimore, United States.

¹⁴Universidad de Huánuco, Huanuco, Peru.

Abstract

For human/SARS-CoV-2 interactome genes *ACE2*, *TMPRSS2* and *BSG*, there is a convincing evidence of association in Asians with influenza-induced SARS for *TMPRSS2*-rs2070788, tag-SNP of the eQTL rs383510. This case illustrates the importance of population genetics and of sequencing data in the design of genetic association studies in different human populations: the high linkage disequilibrium (LD) between rs2070788 and rs383510 is Asian-specific. Leveraging on a combination of genotyping and sequencing data for Native Americans (neglected in genetic studies), we show that while their frequencies of the Asian tag-SNP rs2070788 is, surprisingly, the highest worldwide, it is not in LD with the eQTL rs383510, that therefore, should be directly genotyped in genetic association studies of SARS in populations with Native American ancestry.

Keywords: *TMPRSS2*, *ACE2*, Native Americans, SARS-CoV-2, population genomics.

Received: January 06, 2021; Accepted: June 23, 2021.

Send correspondence to Fernanda Rodrigues-Soares. Universidade Federal do Triângulo Mineiro, Instituto de Ciências Biológicas e Naturais, Rua Vigário Carlos, 100, sala 314, Nossa Senhora da Abadia, 38025-350, Uberaba, MG, Brazil. E-mail: fernanda.soares@uftm.edu.br.

* These authors contributed equally to the article.

In the context of a global interest in host genetic determinants of COVID-19 susceptibility (Casanova and Su, 2020) we established a three-step protocol to gain evidence about human genetic susceptibility to the SARS-CoV-2, the causative agent of the COVID-19 disease:

(i) a systematic review of the literature about genes *ACE2* (angiotensin converting enzyme 2, Xp22.2), *TMPRSS2* (transmembrane serine protease 2, 21q22.3) and *BSG* (basigin, 19p13.3), which codify important proteins for severe acute

respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. SARS-CoV-2 spike S protein contains subunits S1 and S2, which bind the ACE2 cellular receptor, leading to an endosome formation around the virus. After this binding, TMPRSS2 host's transmembrane serine protease cleaves S1/S2 subunits and induces a conformational change in S2, facilitating the endosome formation and allowing the entrance of virus cellular into the cytoplasm. CD147 (also called basigin - BSG) is a transmembrane glycoprotein, encoded by the *BSG* gene, discovered as a new SARS-CoV-2 cellular entry route (Wang *et al.* 2020). We performed a systematic review under the terms “[gene name] genetics infection]”, covering articles published until June 4th, 2020 in PubMed and in bioRxiv during 2020 (Figure 1A). For the ACE2 and BSG viral receptors, there was no solid and direct evidence of association between genetic polymorphisms and any respiratory viral infections.

(ii) we annotated SNVs in *ACE2*, *TMPRSS2*, and *BSG* mining and integrating information from 24 biological and biomedical databases, using our bioinformatics tool (MASSA) [Multi-Agent System for SNP Annotation (Soares-Souza, 2014)], to identify functionally relevant variants (Table S1-A). MASSA integrates data with clinical findings from NCBI Databases like ClinVar and ClinGen. MASSA also includes approaches to distinguish between functional alleles, underlying clinical phenotypes and benign variants, cross-checking the data with multiple different databases. To ensure that collected variants are relevant for our analysis, MASSA performs some secondary filters, taking into account the frequency of alleles and SIFT and Polyphen predictions. The tool, in addition to performing the filters described above, searches for variants that have been cited in PubMed and also compares them to the OMIM database. From that, we've found 26 putatively functional variants for *ACE2*, 5 for *TMPRSS2* and 17 for *BSG* gene, resulting in a total of 48 genetic variants.

(iii) we performed a population genetics analysis of the 48 functionally relevant variants in the *ACE2*, *TMPRSS2* and *BSG* genes in human populations to detect particular patterns of between-population genetic differentiation and independently of evidence of genetic association between *ACE2*, *TMPRSS2* and *BSG* variants and infectious diseases, using published and unpublished data from different worldwide populations (Table S1-B), enriched for Latin Americans, who are mainly the product of admixture of Native Americans, Europeans and Africans. Unpublished data include the Peruvian Native Americans from the *Laboratório de Diversidade Genética Humana (UFMG)* and the whole genome sequenced Native Americans and admixed Peruvian populations from the Peruvian Genome Project. Detailed methodology is available on Text S1.

ACE2 and *BSG* allele frequencies and their regression analyses between population genomic ancestry (Native American, African, European and East Asian) and frequencies of functionally relevant SNPs are presented in Table S2 (A and B) and Table S3 (A and B), respectively. We did not observe a particular pattern of inter-population genetic diversity for most of our 48 analyzed SNPs. Our most illustrative result regards *TMPRSS2* (Table S4). In our systematic review, the only genotype/infection association was reported by Cheng *et al.* (2015), between rs2070788-G, a tag-SNP (i.e. in high linkage disequilibrium, $r^2 > 0.80$) of the regulatory e-QTL

rs383510. Both SNPs are located in intronic regions and were associated in Asiatic populations with severe pulmonary damage caused by influenza A(H7N9) in 2014 (OR 1.70 [1.13-2.55]) and rs2070788 was associated with severe pulmonary damage caused by the influenza A(H1N1) in 2009 (OR 1.54 [1.14-2.06]). The authors validated their finding by an *in-vitro* polymerase assay, showing that rs383510 maps on a region that regulates *TMPRSS2* expression (rs383510-T promotes a higher expression of *TMPRSS2* than rs383510-C), and therefore is a functionally relevant SNP tagged by rs2070788-G. This result and the role of *TMPRSS2* in SARS-CoV-2 infection suggest that there are shared elements in the pathogenesis of SARS caused by different viral infections.

As in Cheng *et al.* (2015), the tag-SNP rs2070788 (<https://www.ncbi.nlm.nih.gov/snp/rs2070788>) is more commonly studied than the functional SNP rs383510 (<https://www.ncbi.nlm.nih.gov/snp/rs383510>), because the former is present in more SNP genome-wide arrays and has a TaqMan (Thermo Fisher, US) probe, while rs383510 does not. Irham *et al.* (2020) by analyzing variants that modify *TMPRSS2* expression, have observed that rs2070788-G and rs383510-T were associated with the increase of protein expression in lung tissue. For this reason, there is a possibility of association to a higher susceptibility to COVID-19 development. Moreover, Latini *et al.* (2020), using complete exome sequencing, have evidenced that *TMPRSS2*-rs75603675 and rs12329760 were associated with COVID-19 protection. We examined our unpublished dataset of Native American and of admixed Latin Americans for the putative tag-SNP rs2070788 (genotyped with the Illumina Omni2.5 array) but not for rs383510 because there is no large dataset available for it. We realized that, interestingly, frequencies of the putative tag-SNP rs2070788-G are strongly correlated with population Native American ancestry (Figure 1B, Table S4), and its highest frequency worldwide are in Native Americans. Non-admixed Native American populations have frequencies between 76% and 94%, compared to around 50% in Europeans, 30-40% in Asians and 18-33% in Africans. Furthermore, the putative tag-SNP rs2070788-G is among the 5% most differentiated SNPs in Native Americans respect to Asians (the genetically closest continental group, Figure 1C). This result led us to hypothesize that Native Americans may have the highest frequencies of SARS-CoV-2 susceptibility alleles in *TMPRSS2* and to test this hypothesis we designed a further association study between rs2070788 and COVID-19 in Peru (a country inhabited by populations with predominant Native American ancestry).

Mills and Rahal (2020) described that in 2020, 81,5% and 11,2% of the genome-wide association studies (GWAS) have analyzed, respectively, Europeans and Asians; in contrast, 0,38% have investigated Latin Americans. Recently, Ellinghaus *et al.* (2020) have published a GWAS (n=3,815 Europeans) and found a 3p21.31 gene cluster as a susceptibility locus in COVID-19 with respiratory failure and a possible contribution of the ABO blood-group system. However, none of recent COVID-19 GWAS have analyzed Native American populations.

Because Harris *et al.* (2018) have published whole genome sequencing data for 150 Peruvian individuals with high Native American ancestry, we used those data to test the linkage disequilibrium between the putative tag-SNP rs2070788 and the functional SNP rs383510. Surprisingly

for us, in these Native Americans, the continental group that, on average, shows the highest linkage disequilibrium in the human genome (Bosch *et al.* 2009), there is no linkage disequilibrium between rs2070788 and rs383510 ($r^2=0.05$, $D'=0.61$, Figure 1D). We verified that rs2070788 and rs383510 are in linkage disequilibrium only in Asian populations (Figure 1D) and therefore, the former is a tag-SNP of the latter functional SNP only in Asians. Thus, based on our current knowledge, there is no evidence that Native Americans have the highest frequency worldwide of *TMPRSS2* SARS susceptibility variants, as a superficial analysis would suggest, which was not the case of this study. In this context, as a previous example of distinct patterns of LD, Hünemeier *et al.* (2015) have demonstrated that two-SNP haplotypes, earlier suggested as proxies for 5-HTTLPR by Vinkhuyzen *et al.* (2011) in European descendants, could not be used in such way for Native Americans due to their absence of linkage disequilibrium at this locus. An association study in Native Americans should focus on the causative variant rs383510, to test its involvement in SARS induced by viral infection.

In summary, this case illustrates that, to properly design genetic association studies, it is compelling to: (i) consider the complexities of population genetics concepts such as differences not only in frequencies but also in linkage disequilibrium among different human populations, (ii) to have access to whole genome sequencing data for the broadest array of human populations, as we have in this case for Peruvians Native Americans, (iii) to perform genetic studies including neglected populations, such as Native American, aiming to create specific genetic knowledge for these populations. Moreover, if for any reason, including socioeconomic vulnerability, COVID-19 is more common in individuals with high Native American ancestries, the test of association between the rs383510 and COVID-19 phenotypes should be controlled for ancestry. Without considering differences in linkage disequilibrium (also for imputation in GWAS) and sequencing data, as well as ancestry, this is an example of how association studies may reach misleading conclusions in times when a search for susceptibility variants for SARS-CoV-2 is intense.

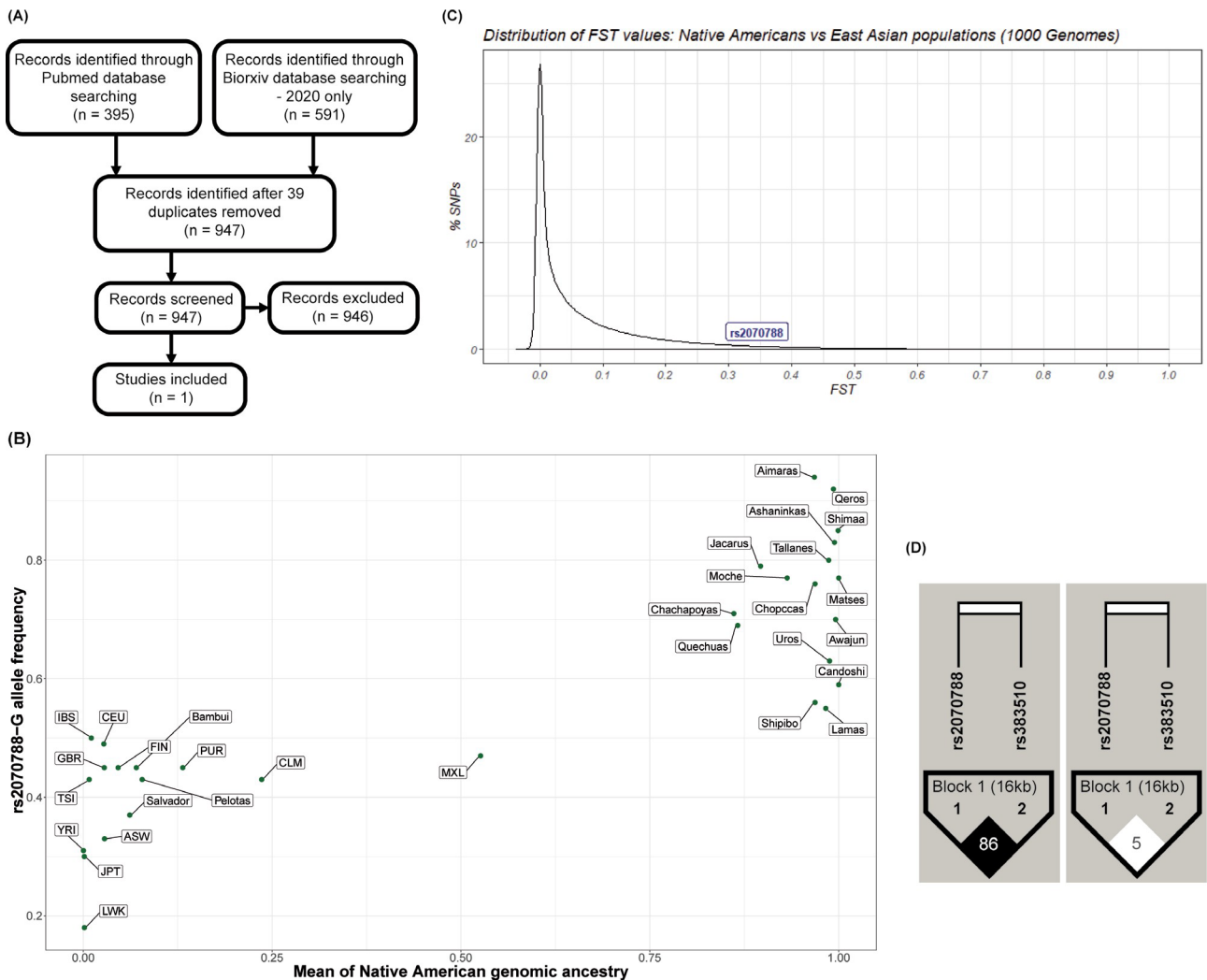


Figure 1 – (A) PRISMA flowchart of the systematic review; (B) Frequencies of the rs2070788 SNP and Native American ancestry in different populations (Populations from 1000 Genomes Project: ASW, Americans of African Ancestry in SW USA; CEU, Utah Residents (CEPH) with Northern and Western European Ancestry; CLM, Colombians from Medellin, Colombia; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian Population in Spain; JPT, Japanese in Tokyo, Japan; LWK, Luhya in Webuye, Kenya; PUR, Puerto Ricans from Puerto Rico; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeria); (C) Fst values distribution of Native Americans vs East Asian populations for 71 SNPs of *TMPRSS2* gene; (D) Linkage disequilibrium between rs2070788 and rs383510 in East Asian and Native American populations.

Acknowledgements

This work was funded by CNPq, CAPES, Department of Science and Technology of the Brazilian Ministry of Health (DECIT/MS), the Peruvian Genome Project from the Peruvian National Institute of Health and grants FONDEF D10I1007, D10E1007 and FONDEQUIPEQM140157 (CONICYT, Chile).

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Study design: FSGK, HG, RSA, ET-S, TPL and FR-S; Contribution of new reagents or analytical tools: LJ-V, AV-D, CSN, RAV, HG, TO'C, ET-S; Data analysis: FSGK, MP-O, MMS, ST-L, CZ, RM, LAM, IA, CS-C, VCF, MMA, MLS, VB, GBS-S; Manuscript preparation: FSGK, ET-S, TPL, FR-S. All authors have revised and approved the final version of the manuscript.

References

- Bosch E, Laayouni H, Morcillo-Suarez C, Casals F, Moreno-Estrada A, Ferrer-Admetlla A, Gardner M, Rosa A, Navarro A, Comas D *et al.* (2009) Decay of linkage disequilibrium within genes across HGDP-CEPH human samples: Most population isolates do not show increased LD. *BMC Genomics* 10:338.
- Casanova J-L, Su HC and COVID Human Genetic Effort (2020) A global effort to define the human genetics of protective immunity to SARS-CoV-2 infection. *Cell* 181:1194-1199.
- Cheng Z, Zhou J, To KK-W, Chu H, Li C, Wang D, Yang D, Zheng S, Hao K, Bossé Y *et al.* (2015) Identification of TMPRSS2 as a susceptibility gene for severe 2009 pandemic A(H1N1) influenza and A(H7N9) influenza. *J Infect Dis* 212:1214-1221.
- Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, Invernizzi P, Fernández J, Prati D, Baselli G, Asselta R *et al.* (2020) Genomewide association study of severe Covid-19 with respiratory failure. *N Engl J Med* 383:1522-1534.
- Harris DN, Song W, Shetty AC, Levano KS, 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 U S A* 115: E6526-E6535.
- Hünemeier T, Bisso-Machado R, Salzano FM and Bortolini MC (2015) Native American ancestry leads to complexity in 5-HTTLPR polymorphism association studies. *Mol Psychiatry* 20:659-660.

- Irham LM, Chou W-H, Calkins MJ, Adikusuma W, Hsieh S-L and Chang W-C (2020) Genetic variants that influence SARS-CoV-2 receptor TMPRSS2 expression among population cohorts from multiple continents. *Biochem Biophys Res Commun* 529:263-269.
- Latini A, Agolini E, Novelli A, Borgiani P, Giannini R, Gravina P, Smarrazzo A, Dauri M, Andreoni M, Rogliani P *et al.* (2020) COVID-19 and genetic variants of protein involved in the SARS-CoV-2 entry into the host cells. *Genes (Basel)* 11:1-8.
- Mills MC and Rahal C (2020) The GWAS diversity monitor tracks diversity by disease in real time. *Nat Genet* 52:242-243.
- Soares-Souza GB (2014) New approaches for database integration and development of bioinformatics tools for population genetics studies. D. Sc. Thesis, Federal University of Minas Gerais, 213 p.
- Vinkhuyzen AAE, Dumenil T, Ryan L, Gordon SD, Henders AK, Madden PAF, Heath AC, Montgomery GW, Martin NG and Wray NR (2011) Identification of tag haplotypes for 5HTTLPR for different genome-wide SNP platforms. *Mol Psychiatry* 16:1073-1075.
- Wang K, Chen W, Zhou Y-S, Lian J-Q, Zhang Z, Du P, Gong L, Zhang Y, Cui H-Y, Geng J-J *et al.* (2020) SARS-CoV-2 invades host cells via a novel route: CD147-spike protein. [bioRxiv:2020.03.14.988345](https://doi.org/10.1101/2020.03.14.988345).

Supplementary material

The following online material is available for this article:

- Text S1 – Detailed Methods.
 Table S1-A – Twenty six biological and biomedical databases integrated by MASSA.
 Table S1-B – Sample datasets descriptions.
 Table S2-A – *ACE2* allele frequencies.
 Table S2-B – *ACE2* regression values.
 Table S3-A – *BSG* allele frequencies.
 Table S3-B – *BSG* regression values.
 Table S4-A – *TMPRSS2* allele frequencies.
 Table S4-B – *TMPRSS2* regression values.

Associate Editor: Diogo Meyer

License information: This is an open-access article distributed under the terms of the Creative Commons Attribution License (type CC-BY), which permits unrestricted use, distribution and reproduction in any medium, provided the original article is properly cited.



Tracing the Distribution of European Lactase Persistence Genotypes Along the Americas

OPEN ACCESS

Edited by:

Edward Hollox,
University of Leicester,
United Kingdom

Reviewed by:

Andres Moreno-Estrada,
National Polytechnic Institute of
Mexico (CINVESTAV), Mexico
Dallas Swallow,
University College London,
United Kingdom

*Correspondence:

Marcia Holsbach Beltrame
marcia.hbeltrame@gmail.com
Victor Borda
VBorda@som.umaryland.edu

†These authors have contributed
equally to this work and share last
authorship

Specialty section:

This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: 23 February 2021

Accepted: 23 June 2021

Published: 22 September 2021

Citation:

Guimarães Alves AC, Sukow NM,
Adelman Cipolla G, Mendes M,
Leal TP, Petzl-Erler ML,
Lehtonen Rodrigues Souza R,
Rainha de Souza I, Sanchez C,
Santolalla M, Loesch D, Dean M,
Machado M, Moon J-Y, Kaplan R,
North KE, Weiss S, Barreto ML, Lima-
Costa MF, Guio H, Cáceres O,
Padilla C, Tarazona-Santos E, Mata IF,
Dieguez E, Raggio V, Lescano A, Tumas
V, Borges V, Ferraz HB, Rieder CR,
Schumacher-Schuh A, Santos-Lobato
BL, Chana-Cuevas P, Fernandez W,
Arboleda G,
Arboleda H, Arboleda-Bustos CE,
O'Connor TD, Beltrame MH and
Borda V (2021) Tracing
the Distribution of European Lactase
Persistence Genotypes Along
the Americas.
Front. Genet. 12:671079.
doi: 10.3389/fgene.2021.671079

Ana Cecília Guimarães Alves^{1,2}, Natalie Mary Sukow¹, Gabriel Adelman Cipolla¹, Marla Mendes³, Thiago P. Leal³, Maria Luiza Petzl-Erler^{1,2}, Ricardo Lehtonen Rodrigues Souza^{2,4}, Iliada Rainha de Souza^{1,5}, Cesar Sanchez⁶, Meddy Santolalla⁷, Douglas Loesch⁸, Michael Dean⁹, Moara Machado³, Jee-Young Moon¹⁰, Robert Kaplan^{10,11}, Kari E. North¹², Scott Weiss¹³, Mauricio L. Barreto^{14,15}, M. Fernanda Lima-Costa^{16,17}, Heinner Guio^{6,18}, Omar Cáceres^{6,19}, Carlos Padilla⁶, Eduardo Tarazona-Santos³, Ignacio F. Mata^{20,21,22}, Elena Dieguez²³, Víctor Raggio²⁴, Andres Lescano²³, Vitor Tumas²⁵, Vanderci Borges²⁶, Henrique B. Ferraz²⁶, Carlos R. Rieder²⁷, Artur Schumacher-Schuh^{28,29}, Bruno L. Santos-Lobato³⁰, Pedro Chana-Cuevas³¹, William Fernandez³², Gonzalo Arboleda³², Humberto Arboleda³², Carlos E. Arboleda-Bustos³², Timothy D. O'Connor^{8,33,34}, Marcia Holsbach Beltrame^{1,2,*†} and Victor Borda^{8,*†}

¹ Laboratório de Genética Molecular Humana, Departamento de Genética, Setor de Ciências Biológicas, Universidade Federal do Paraná, Curitiba, Brazil, ² Programa de Pós-Graduação em Genética, Departamento de Genética, Setor de Ciências Biológicas, Universidade Federal do Paraná, Curitiba, Brazil, ³ Laboratório de Diversidade Genética Humana, Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, ⁴ Laboratório de Polimorfismos e Ligação, Departamento de Genética, Setor de Ciências Biológicas, Universidade Federal do Paraná, Curitiba, Brazil, ⁵ Laboratório de Polimorfismos Genéticos, Departamento de Biologia Celular, Embriologia e Genética, Centro de Ciências Biológicas, Universidade Federal de Santa Catarina, Florianópolis, Brazil, ⁶ Laboratorio de Biotecnología y Biología Molecular, Instituto Nacional de Salud, Lima, Peru, ⁷ Emerging Diseases and Climate Change Research Unit, School of Public Health and Administration, Universidad Peruana Cayetano Heredia, Lima, Peru, ⁸ Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, United States, ⁹ Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, MD, United States, ¹⁰ Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, United States, ¹¹ Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, United States, ¹² Department of Epidemiology, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, ¹³ Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States, ¹⁴ Universidade Federal da Bahia, Instituto de Saúde Coletiva, Salvador, Brazil, ¹⁵ Fundação Oswaldo Cruz, Centro de Integração de Dados e Conhecimentos para Saúde (Cidacs), Salvador, Brazil, ¹⁶ Fundação Oswaldo Cruz, Instituto René Rachou, Belo Horizonte, Brazil, ¹⁷ Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Saúde Pública, Belo Horizonte, Brazil, ¹⁸ Facultad de Ciencias de la Salud, Universidad de Huánuco, Huánuco, Peru, ¹⁹ Carrera de Medicina Humana, Facultad de Ciencias de la Salud, Universidad Científica del Sur, Lima, Peru, ²⁰ Veterans Affairs Puget Sound Health Care System, Seattle, WA, United States, ²¹ Department of Neurology, University of Washington, Seattle, WA, United States, ²² Lerner Research Institute, Genomic Medicine, Cleveland Clinic, Cleveland, OH, United States, ²³ Neurology Institute, Universidad de la República, Montevideo, Uruguay, ²⁴ Department of Genetics, Facultad de Medicina, Universidad de la República, Montevideo, Uruguay, ²⁵ Ribeirão Preto Medical School, Universidade de São Paulo, Ribeirão Preto, Brazil, ²⁶ Movement Disorders Unit, Department of Neurology and Neurosurgery, Universidade Federal de São Paulo, São Paulo, Brazil, ²⁷ Departamento de Neurologia, Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre, Brazil, ²⁸ Serviço de Neurologia, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil, ²⁹ Departamento de Farmacologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, ³⁰ Instituto de Ciências da Saúde, Universidade Federal do Pará, Belém, Brazil, ³¹ CETRAM, Facultad de Ciencias Médicas, Universidad de Santiago de Chile, Santiago, Chile, ³² Neuroscience and Cell Death Research Groups, Medical School and Genetic Institute, Universidad Nacional de Colombia, Bogotá, Colombia, ³³ Program for Personalized and Genomic Medicine, School of Medicine, University of Maryland, Baltimore, MD, United States, ³⁴ Department of Medicine, School of Medicine, University of Maryland, Baltimore, MD, United States

In adulthood, the ability to digest lactose, the main sugar present in milk of mammals, is a phenotype (lactase persistence) observed in historically herder populations, mainly Northern Europeans, Eastern Africans, and Middle Eastern nomads. As the -13910^*T

allele in the *MCM6* gene is the most well-characterized allele responsible for the lactase persistence phenotype, the $-13910C > T$ (rs4988235) polymorphism is commonly evaluated in lactase persistence studies. Lactase non-persistent adults may develop symptoms of lactose intolerance when consuming dairy products. In the Americas, there is no evidence of the consumption of these products until the arrival of Europeans. However, several American countries' dietary guidelines recommend consuming dairy for adequate human nutrition and health promotion. Considering the extensive use of dairy and the complex ancestry of Pan-American admixed populations, we studied the distribution of $-13910C > T$ lactase persistence genotypes and its flanking haplotypes of European origin in 7,428 individuals from several Pan-American admixed populations. We found that the -13910^*T allele frequency in Pan-American admixed populations is directly correlated with allele frequency of the European sources. Moreover, we did not observe any overrepresentation of European haplotypes in the $-13910C > T$ flanking region, suggesting no selective pressure after admixture in the Americas. Finally, considering the dominant effect of the -13910^*T allele, our results indicate that Pan-American admixed populations are likely to have higher frequency of lactose intolerance, suggesting that general dietary guidelines deserve further evaluation across the continent.

Keywords: $-13910C > T$, *MCM6* gene, lactose intolerance, dairy consumption, nutrition policies, Latin America, population genetics

INTRODUCTION

Lactase phlorizin hydrolase, popularly known as lactase, is an enzyme expressed by enterocytes from the small intestine. Lactase is responsible for the hydrolysis of lactose, a non-absorbable sugar present in milk of mammals, into glucose and galactose, which in turn are simple sugars absorbable by the intestinal mucosa (Mantei et al., 1988). In mammals, lactase is highly expressed by newborns; after weaning, however, its expression is naturally downregulated (Wang et al., 1994), leading to a phenotype known as lactase non-persistence (LNP). This trait, also known as adult-type hypolactasia, is related to lactose indigestion and malabsorption. It may progress to lactose intolerance associated with indigestion, bloating, abdominal pain, nausea, vomiting, flatulence, and diarrhea (Auricchio et al., 1963; Boll et al., 1991). In contrast, a high prevalence of lactase persistence (LP) occurs in populations with a long history of pastoralism and dairy consumption, mainly in Northern Europe and among nomads in Africa and the Middle East (Ingram et al., 2009). Single-nucleotide polymorphisms (SNPs) in the *MCM6* gene, located 14 kb upstream of the *LCT* gene, in its enhancer region, are believed to be responsible for the LP phenotype. The -13910^*T allele ($-13910C > T$, rs4988235, intron 13, *MCM6* gene) is widely distributed in the European population; the allele frequencies vary from 8.9% in Tuscany (Italy) to 72.0% among British (England and Scotland), with an average of 50.8% in Europe (1000 Genomes Project Consortium, Auton et al., 2015). The *T* allele enhances *LCT* expression at the mRNA level, causing LP in adulthood, even among heterozygotes, thus determining a dominant trait (*TT* and *TC* genotypes) (Enattah et al., 2002). In Asia, despite the occurrence of pastoralist populations, the

LP phenotype frequency is lower than that reported in Europe and Africa (Ségurel and Bon, 2017), and the -13910^*T allele frequencies rarely exceed 10% (Itan et al., 2010). Other SNPs in African and Middle Eastern populations are observed in the same *MCM6* gene region (intron 13) and are also associated with the LP phenotype. Among these, the most well-known are $-14010G > C$ (rs145946881), $-14009T > G$ (rs869051967), $-13915T > G$ (rs41380347), and $-13907C > G$ (rs41525747) (Tishkoff et al., 2007; Ranciaro et al., 2014; Liebert et al., 2017).

In the Americas, archaeozoological evidence supports the occurrence of mammal domestication around 7,000–6,000 years ago and included the ancestors of modern llamas, alpacas, vicuñas, and guanacos (Wheeler, 1995; Kadwell et al., 2001). Despite breeding these animals, there is no archeological or cultural evidence of dairy consumption by Native American populations until Europeans arrived in the late fifteenth century (Gade, 1999) – cattle were not introduced in the Americas until 1493 (Primo, 1992). Not surprisingly, studies regarding living Native Americans from Brazil (Friedrich et al., 2012a), Chile (Fernández et al., 2016), Ecuador (Paz-Y-Miño et al., 2016), Mexico (Ojeda-Granados et al., 2016), Peru (Figueroa et al., 1971), and the United States (Duncan and Scott, 1972; Casey, 2005) identified lower LP phenotype frequencies in these populations (20% on average), suggesting that most people cannot digest lactose naturally. Countries such as Peru, Mexico, and Chile have a high proportion of Native American ancestry – 80, 57.5, and 49.3%, respectively (Ruiz-Linares et al., 2014; Adhikari et al., 2016; Harris et al., 2018) – which could explain the low frequency of the LP phenotype. Although there are no -13910^*T allele frequencies reported in ancient Native American samples, the low frequencies observed in Asian peoples and

present-day Native Americans suggest that the -13910^*T allele was introduced in the Americas after the arrival of Europeans. Moreover, African and Middle Eastern SNPs have been reported at very low frequencies in the Americas. The frequency of

-14010^*C ($-14010G > C$, rs145946881) in Afro-Brazilians from the South is 0.27% (Friedrich et al., 2012b); the frequencies of -14011^*T ($-14011C > T$, rs4988233) in admixed Brazilians from the North and the Northeast are 0.25 and 0.58%, respectively (Friedrich et al., 2012b); and the frequencies of -13913^*C ($-13913T > C$, rs4145614) and -13915^*G ($-13915T > G$, rs41380347) in Mestizos from Ecuador are 0.20 and 0.50%,

respectively (Paz-Y-Miño et al., 2016). Therefore, the phenotypic variation observed is mainly due to the $-13910C > T$ (rs4988235) substitution of European origin. Thus, the admixture events that led to the formation of current Latin American populations contributed to introducing LP-associated alleles in these groups (Friedrich et al., 2012a; Mendoza Torres et al., 2012; Latorre et al., 2014; Fernández et al., 2016; Ojeda-Granados et al., 2016; Paz-Y-Miño et al., 2016; Valencia et al., 2017; Montalva et al., 2019). Consequently, higher frequencies of the LP phenotype are observed in admixed populations from Brazil (Friedrich et al., 2012b), and Mestizos from Chile (Fernández et al., 2016), Colombia (Mendoza Torres et al., 2012), Ecuador (Paz-Y-Miño et al., 2016), and Mexico (Ojeda-Granados et al., 2016).

Because of the Latin American population formation process, genetic analyses have revealed the heterogeneous pattern of ancestries across the continent (Kehdy et al., 2015; Chacón-Duque et al., 2018; Harris et al., 2018; Soares-Souza et al., 2018). The peopling of the Americas occurred about 20,000 years ago, when an East Asian-derived group accessed the continent through the Bering Strait (Goebel et al., 2008; Dillehay, 2009; Reich et al., 2012; Mendes et al., 2020). Sequential population divisions and little gene flow from other continents after divergence gave rise to distinct Native American populations, with highly differentiated population groups, such as the Mesoamericans, Andeans, and Amazonians, distributed across the continent (Greenberg et al., 1986; Salzano, 2011; Reich et al., 2012; Mendes et al., 2020). Later on, during the Colonial period, Europeans brought about 9 million Africans through the Transatlantic Slave Trade to the Americas (Wehling et al., 1994; John, 1997; de Mello e Souza, 2007). Although there are no official historical records of the subcontinental origin of African peoples, previous genetic studies confirmed that West- Central African ancestry is the most prevalent in the Americas (Gouveia et al., 2020). Further on, between the nineteenth and twentieth centuries, European immigration was intensified mostly in South America. The resulting admixture drastically modified the genetic makeup of the continent.

The complex demographic processes involved in the formation of the Americas gave rise to mosaic populations in which ancestry proportions vary among and within the countries, affecting the distribution of Mendelian and complex phenotypes (Pena et al., 2011, 2020; Ruiz-Linares et al., 2014; Adhikari et al., 2016). Considering this, in the present study, we (i) addressed the geographical distribution of the $-13910C > T$ SNP across the Americas, as well as (ii) analyzed this locus for evidence of positive selection in Pan-American admixed populations

(American continent populations, i.e., North, Central, and South Americas). Moreover, we discussed our results in light of the use of dairy products in public health policies for nourishment in Latin American countries.

MATERIALS AND METHODS

Publicly Available Datasets

We analyzed genomic information of 7,428 unrelated individuals from North, Central, and South Americas (**Supplementary Table 1**). This includes whole-genome sequencing (WGS) data from admixed individuals of the 1000 Genomes Project (1000 Genomes Project Consortium, Auton et al., 2015) [African Americans from Southwest United States (ASW); African Caribbeans from Barbados (ACB); individuals of Mexican ancestry from Los Angeles, United States (MXL); Puerto Ricans from Puerto Rico (PUR); and Colombians from Medellín (CLM)], as well as WGS data from individuals of the TOPMED Project (Taliun et al., 2021) (individuals from the Dominican Republic and Cuba from HCHS/SOL cohort, and individuals from Costa Rica from CRA cohort) and the Peruvian Genome Project (Harris et al., 2018). We also included genotype array data from the LARGE-PD Project (Loesch et al., 2020) (individuals from Brazil, Colombia, Chile, and Uruguay), two Brazilian EPIGEN population-based cohorts (Kehdy et al., 2015) (individuals from Salvador and Bambuí), and from admixed and Native American individuals of the Peruvian Genome Project (Harris et al., 2018; Borda et al., 2020). All these datasets were generated from different sources (**Supplementary Table 1**) but include the $-13910C > T$ SNP (rs4988235). In order to keep a higher SNP density for haplotype-based analyses, we organized the genomic information in two datasets: (i) LARGE-PD only and (ii) a merged dataset (including all other datasets). During the quality control of these two datasets, we excluded SNPs with significant missing data (>10%), loci with 100% of heterozygosity, non-chromosomal information, A/T– C/G genotypes, and SNPs with minor allele count = 1 using PLINK 1.9 (Chang et al., 2015). For the merging process, we used the flags `-bmerge` and `-flip` when necessary. Finally, we kept biallelic SNPs and removed singletons and monomorphic positions as they are not informative for population genetic analyses. We ended with 1,010,078 and 1,528,206 SNPs for the LARGE-PD Project and the merged dataset, respectively.

We also considered allele frequency information of ancient Native Americans from Posth et al. (2018) and Nakatsuka et al. (2020) and of European populations [Utah residents (CEPH) with Northern and Western European ancestry (CEU), British from England and Scotland (GBR), Finnish from Finland (FIN), Iberian from Spain (IBS), and individuals from Toscani in Italy (TSI)] from the 1000 Genomes Project (**Supplementary Table 1**).

Newly Generated Datasets

Moreover, we generated sequencing data for the *LCT* enhancer (*MCM6* gene, intron 13) for 259 Afro-Brazilians individuals from the south region of Brazil. The sample included 241 individuals from Curitiba and its metropolitan region (Paraná state) and

18 individuals from the quilombola community of Sertão do Valongo (Santa Catarina state) (**Supplementary Figure 1**). For Curitiba individuals, first-degree relatives were excluded based on self-declared information. Blood samples had been previously collected as authorized under the Brazilian CONEP (Comissão Nacional de Ética em Pesquisa) registry numbers 180/2001 and 2.970.200 (CAAE: 02727412.4.0000.0096). These participants were classified as Afro-Brazilians. For the quilombola community, blood samples were collected according to the ethical guidelines in effect at that time, and individuals were classified as Afro-Brazilians considering the settlement history and isolation of the community described elsewhere (Souza and Culpí, 1992, 2005). All participants gave their written informed consent. DNA was extracted either by the phenol–chloroform–isoamyl alcohol method (Sambrook et al., 1989) or by the salting-out method (Lahiri and Nurnberger, 1991).

Sequencing

The Afro-Brazilians sample ($n = 259$) was sequenced for a fragment of 594 bp in the *MCM6* gene, including the $-13910C > T$ SNP (rs4988235). The fragment was amplified by polymerase chain reaction (PCR) on a Mastercycler EP Gradient S[®] (Eppendorf, Germany). The forward and reverse primers used were as follows: 5'-GGCAGGGGTTGGAACCTTC-3' and 5'-CTGTTGAATGCTCATACGACCA-3', respectively. Other reagents and the protocol used are described in **Supplementary Tables 2, 3**, respectively.

The PCR products were purified using exonuclease I (Fermentas, United States) and alkaline phosphatase (Thermo Fisher Scientific, United States) on a Mastercycler EP Gradient S[®] (Eppendorf) at 37°C for 1 h and 80°C for 15 min. The sequencing was performed using the same primers used in the PCR and BigDye[®] Terminator Cycle Sequencing Standard v3.1 (Life Technologies, United States), according to the instructions of the manufacturer. Sequencing reactions consisted of a first step at 95°C for 1 min, followed by 25 cycles of 95°C for 10 s, 50°C for 5 s, and 60°C for 4 min. The sequencing products were purified using ethanol (Merck, Germany), resuspended in Hi-Di Formamide (Life Technologies), and, finally, submitted to capillary electrophoresis in a 3500xl Genetic Analyzer Sequencer (Life Technologies).

We analyzed the obtained sequences using the Mutation Surveyor[®] 3.30 software (SoftGenetics, United States), which

aligns the forward and reverse sequences to a human genome reference sequence (GRCh38) available in NCBI (National Center for Biotechnology Information) resources, enabling the evaluation of amplified sequences and the genotype annotation for the $-13910C > T$ SNP (**Supplementary Table 4**). The sequences were deposited in the GenBank at NCBI website under the accession numbers MZ362598 to MZ362856.

Allelic and Genotypic Frequencies for the -13910^*T Variant

We calculated allelic frequencies for the -13910^*T allele (rs4988235) in each population using the `-extract` and `-freq` flags of PLINK 1.9. For genotype frequencies, we used the `-hardy`

flag, which also calculates whether the population is in Hardy–Weinberg equilibrium for the given locus.

Ancestry Analyses

Global and local ancestry inferences were performed for each dataset. For both analyses, we merged each dataset with a reference panel of 848 individuals from four continental ancestries: 206 European (IBS and CEU), 207 African (LWK and YRI), and 207 East Asian (CDX and JPT) individuals from the 1000 Genomes Project, and 228 unadmixed Native American individuals from the Peruvian Genome Project (Borda et al., 2020). As the genetic information available for Afro-Brazilians from Curitiba and Sertão do Valongo corresponds only to the *MCM6* region, it was not possible to perform global or local ancestry analyses for these samples. Instead, we extracted the global ancestry information for Sertão do Valongo individuals from Luizon (2007), who performed this inference using eight ancestry-informative markers in the same population. No data that could enable this analysis were available for Afro-Brazilians from Curitiba.

For global ancestry analysis, we removed linked variants ($r^2 > 0.1$) using the PLINK flag `-indep-pairwise` with the

following parameters: 50 10 0.1, and applied a minor allele frequency (MAF) filter of 1%. We inferred global ancestry proportions using ADMIXTURE (Alexander et al., 2009) on supervised mode for four ancestry clusters for each population merged with the references. After ADMIXTURE runs, we calculated the average proportion of European ancestry for each population. We used R to estimate the Spearman correlation (ρ) between the average proportion of European ancestry and allelic and genotypic frequencies for all populations.

To infer the ancestry of the genomic region flanking the $-13910C > T$ SNP, we analyzed the complete chromosome 2, which includes the *MCM6* gene, for each dataset (LARGE-PD Project dataset and the merged dataset) without removing linked variants. First, for each dataset, all chromosomes were phased with shapeit4 (Delaneau et al., 2019) using the GRCh38 genetic map and the MCMC parameters: `-mcmc-iterations 10b,1p,1b,1p,1b,1p,1b,1p,10m`, which perform 10 burn-in iterations, followed by four paired runs of pruning and burn-in, and, finally, 10 main iterations of sampling. Then, we ran RFMix ver2 in order to infer the local ancestry (Maples et al., 2013) using the phased dataset with two expectation–

maximization runs. After RFMix completion, we used the RFMix msp output, which includes the most likely assignment of ancestry per conditional random field point to obtain the size of the haplotypes per ancestry. The size was determined by considering the uninterrupted length (in base pairs) of the genomic region flanking the $-13910C > T$ SNP until a switch in ancestry along the haplotype. We used in-house Perl and R scripts to determine the ancestry of the flanking region of -13910^*T allele and the length distribution of European haplotypes of the flanking region of the *MCM6* gene.¹

¹ All Perl and R scripts used for this study are freely available at: <https://github.com/vicbp1/Lactase-persistence-in-the-Americas.git>.

Natural Selection Analysis

In order to identify evidence of natural selection acting over the *MCM6* region, we applied two approaches. First, to identify a potential overrepresentation of European haplotypes due to recent, post-admixture natural selection acting over the *MCM6* region, we compared the average global European ancestry with the proportion of European haplotypes inferred by RFMix for each population. Moreover, we analyzed the length distribution of uninterrupted European haplotypes around the $-13910C > T$ SNP to explore the genomic region dynamics. Second, we inferred the pattern of extended haplotype homozygosity (EHH) and the length distribution of derived haplotypes around the $-13910C > T$ SNP in each population in order to identify whether the selection signal observed in the populations of European source is also observed in the admixed Pan-American populations. For this purpose, we used the rehh package (Gautier et al., 2017). Then, we estimated the integrated haplotype score (iHS), which indicates if a locus is under recent positive selection. This score compares the levels of linkage disequilibrium surrounding a positively selected allele with the ancestral allele background at the same position (Voight et al., 2006). For such iHS inference, we used the selscan software (Szpiech and Hernandez, 2014) with default parameters for phased information for each population. Finally, we calculated the genome-wide iHS Z-scores value by a normalization using the norm package, provided with selscan, in derived allele frequency bins with the $-bin$ option set as 20. We calculated a two-tailed p -value of the SNPs based on the normalized iHS Z-scores by dividing the proportional rank of the statistic by the total number of values in the distribution.

RESULTS

The -13910^*T Allele Distribution Is Correlated With European Ancestry in the Americas

We explored the geographical distribution of the -13910^*T allele in 25 Pan-American populations of 12 countries. The highest frequency of the -13910^*T allele in the Americas occurs in the Uruguayan population (35%, **Figure 1** and **Supplementary Table 5**). Conversely, the lowest frequencies were observed for Peruvian and African descendant populations. A positive correlation was observed between -13910^*T allele frequencies and the average proportion of European ancestry ($\rho = 0.843$, $p = 4.414e-7$). Consequently, we did not observe an overrepresentation of the -13910^*T allele compared with the average proportion of European ancestry in any population (**Figure 1**). Moreover, in Afro-Brazilians from Curitiba, we identified the -14011^*T ($-14011C > T$, rs145946881) and -13915^*G ($-13915T > G$, rs41380347) LP-associated alleles in low frequencies (0.2%).

Considering that the LP phenotype is a dominant trait, we explored the geographical distribution of -13910^*T allele carriers (*TT* and *TC* genotypes) as a proxy for the proportion of individuals with this phenotype. The observed genotype

frequencies did not differ from those expected under Hardy–Weinberg equilibrium in all Pan-American admixed populations (**Supplementary Table 5**). Higher proportions of *T* allele carriers were observed in Uruguay (61%), Southeast–South Brazil (54%, including Bambuí and Porto Alegre), and Cuba (53%), which also have higher proportions of European ancestry (**Figure 2** and **Supplementary Table 5**), suggesting that more than half of the individuals of these samples are LP. Conversely, this inferred phenotype is absent in ancient and some present-day Peruvian Native Americans (**Figure 2** and **Supplementary Table 5**). Furthermore, admixed Peruvians and African descendants (ASW, ACB, Afro-Peruvians, and Afro-Brazilians) have lower LP frequencies. Especially, Afro-Brazilians from Sertão do Valongo did not carry the -13910^*T allele nor other LP-associated alleles.

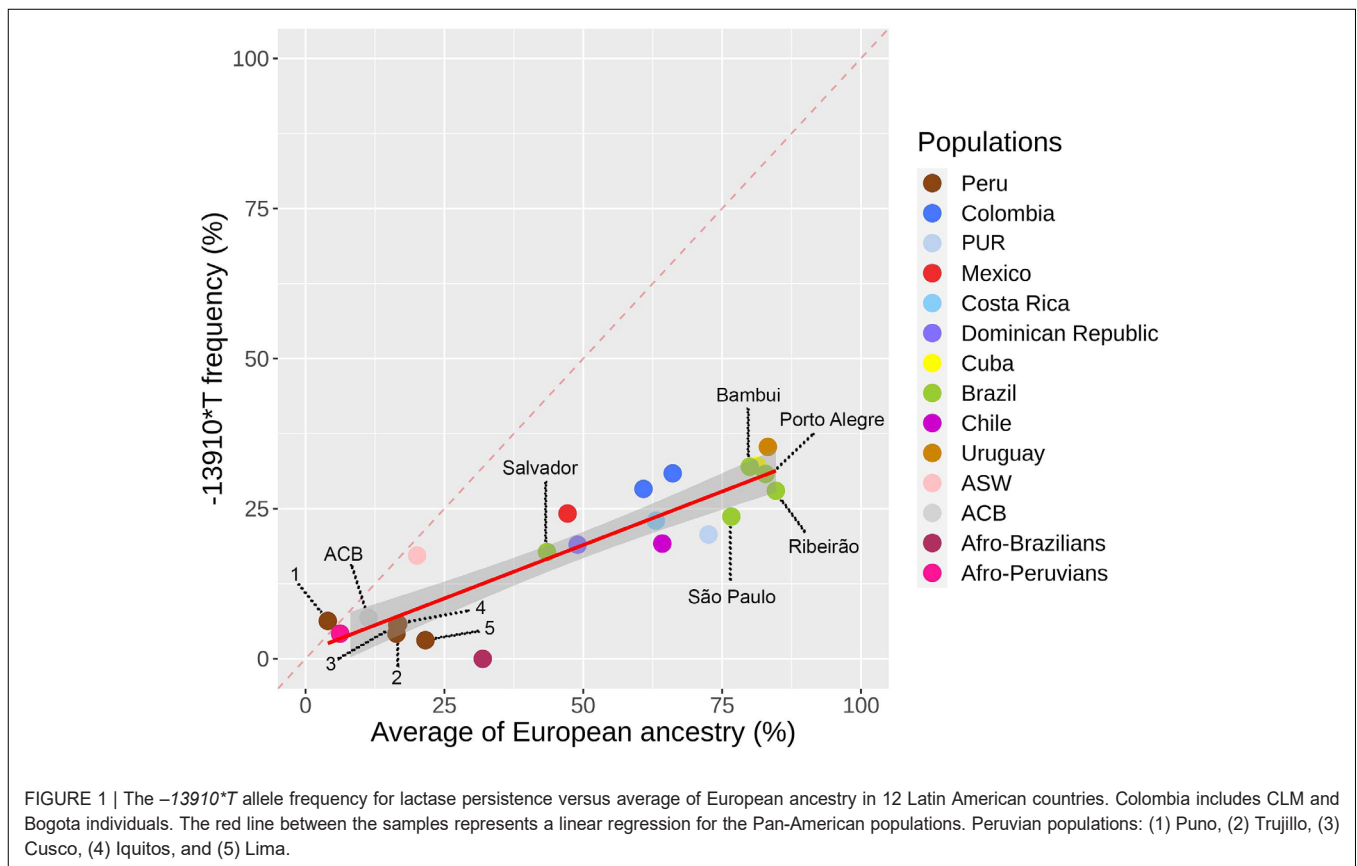
The Haplotype Distribution of the *MCM6* Gene Reflects the History Along the Americas

To infer which forces are playing a significant role in the evolution of the *MCM6* gene in the Americas, we explored the distribution of counts and lengths of European haplotypes that include this gene. To address this goal, we performed a local ancestry inference for this locus and confirmed that the -13910^*T allele is observed only in European haplotypes in all Pan-American admixed populations. To determine if there is an overrepresentation of European haplotypes (with or without the -13910^*T allele) in the *MCM6* region, we estimated the proportion of uninterrupted European haplotypes and compared it to the genome-wide proportion of European ancestry for each population. Our results (**Figure 3B**) revealed a high correlation between European haplotype frequencies and the percentage of genome-wide European ancestry ($\rho = 0.9435$, $p = 4.042e-6$). This pattern

suggests the possibility of neutral evolution of this genomic region after the admixture processes that gave rise to these Pan-American admixed populations.

We explored the admixture dynamics of this region by analyzing the length distribution of uninterrupted European haplotypes that include the *MCM6* gene. We observed a high density of small European segments in populations with a lower proportion of European ancestry (**Figure 3C**), such as Peruvians and African descendants, suggesting an early gene flow from Europeans. Furthermore, populations with higher European ancestry have a wide length distribution of European segments, indicating a continuous or multistage gene flow from Europeans. Both patterns are in agreement with historical records and population genetics studies.

We performed EHH analysis for each population to evaluate the haplotypes and to compare their backgrounds for the -13910^*T allele with the ancestral -13910^*C allele. Most of the assessed populations presented a decay pattern of haplotype homozygosity around the -13910^*T allele (**Supplementary Figure 2**). Peruvians from Cusco and the populations of Bogota, Chile, and Porto Alegre have a linkage disequilibrium (LD) block of 1 Mb with 100% EHH that includes the derived allele (**Figure 4** and **Supplementary Figure 2**). Observing



the length of the EHH blocks (Figure 4 and Supplementary Figure 3), it is remarkable that African American and Porto Alegre populations have more extended and more homogeneous haplotypes for the derived allele in opposition to the haplotypes containing the -13910^*C allele (Figure 4). In both populations, all derived haplotypes are longer than 1 Mb. Specifically, in African Americans, some haplotypes reach 5 Mb. Moreover, several Pan-American populations presented iHS values greater than 2 (Supplementary Table 5), whereas

African Americans presented the highest value ($iHS = 4.2$, Figure 5).

Interestingly, when we calculated the derived haplotype frequency, which includes the -13910^*T allele, in European haplotypes for each Pan-American population, the African Americans from Southwest United States (ASW) had the highest value (80%) (Supplementary Table 5). This frequency is quite similar to the frequency of the derived haplotype in Northern European populations (Supplementary Figure 4). Moreover, we observed that populations that were Spaniard or Portuguese colonies (Peru, Colombia, Mexico, Chile,

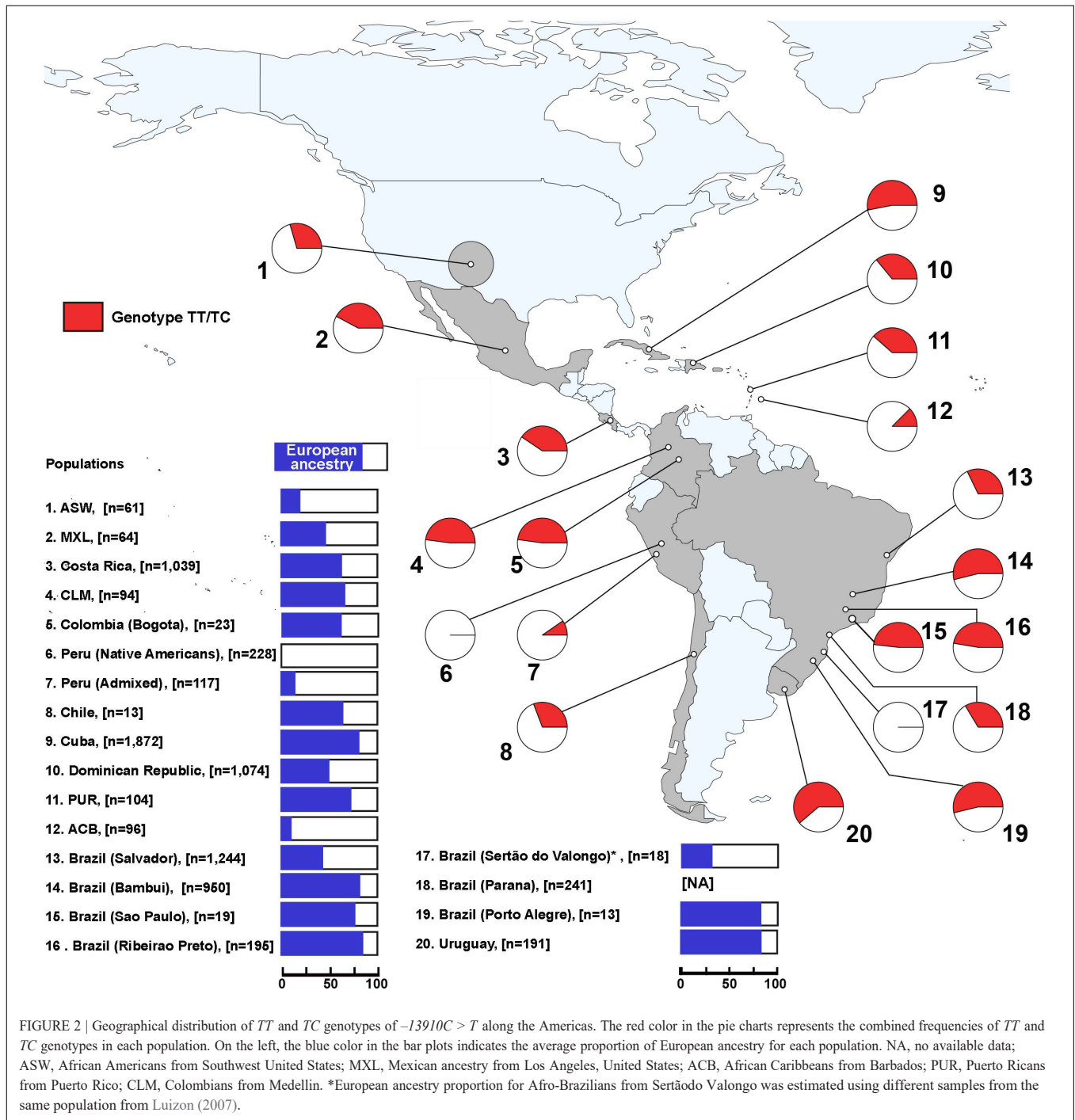
Uruguay, and Brazil) had derived haplotype frequencies more similar to those of Iberians from the 1000 Genomes Project (Supplementary Figure 4). All these observations suggest that in Pan-American admixed populations, for the genomic segment that includes the *MCM6* gene, demography plays a more significant role than natural selection after the admixture processes.

DISCUSSION

LP is among the most strongly selected phenotypes in human populations; for LP, positive selection occurred during the last 5,000–10,000 years (Bersaglieri et al., 2004). In the Americas, there is no evidence of dairy consumption by native groups until the arrival of Europeans (Gade, 1999). Moreover, the digestion test performed in Native American people evidenced the highest incidence (>80%) of lactose intolerance in the Americas (Duncan and Scott, 1972; Caskey et al., 1977). This is consistent with

the absence of the -13910^*T allele in ancient and present-day unadmixed Native Americans and with the hypothesis that this allele was probably introduced in the Americas by Europeans since the Colonial period. Furthermore, the identification of the -14011^*T and -13915^*G alleles in low frequencies in Afro-Brazilians from Curitiba agrees with the frequencies previously reported (Friedrich et al., 2012b; Paz-Y-Miño et al., 2016), corroborating the suggestion that the phenotypic variation observed in this study is mainly due to the $-13910^*C > T$ variation of European origin.

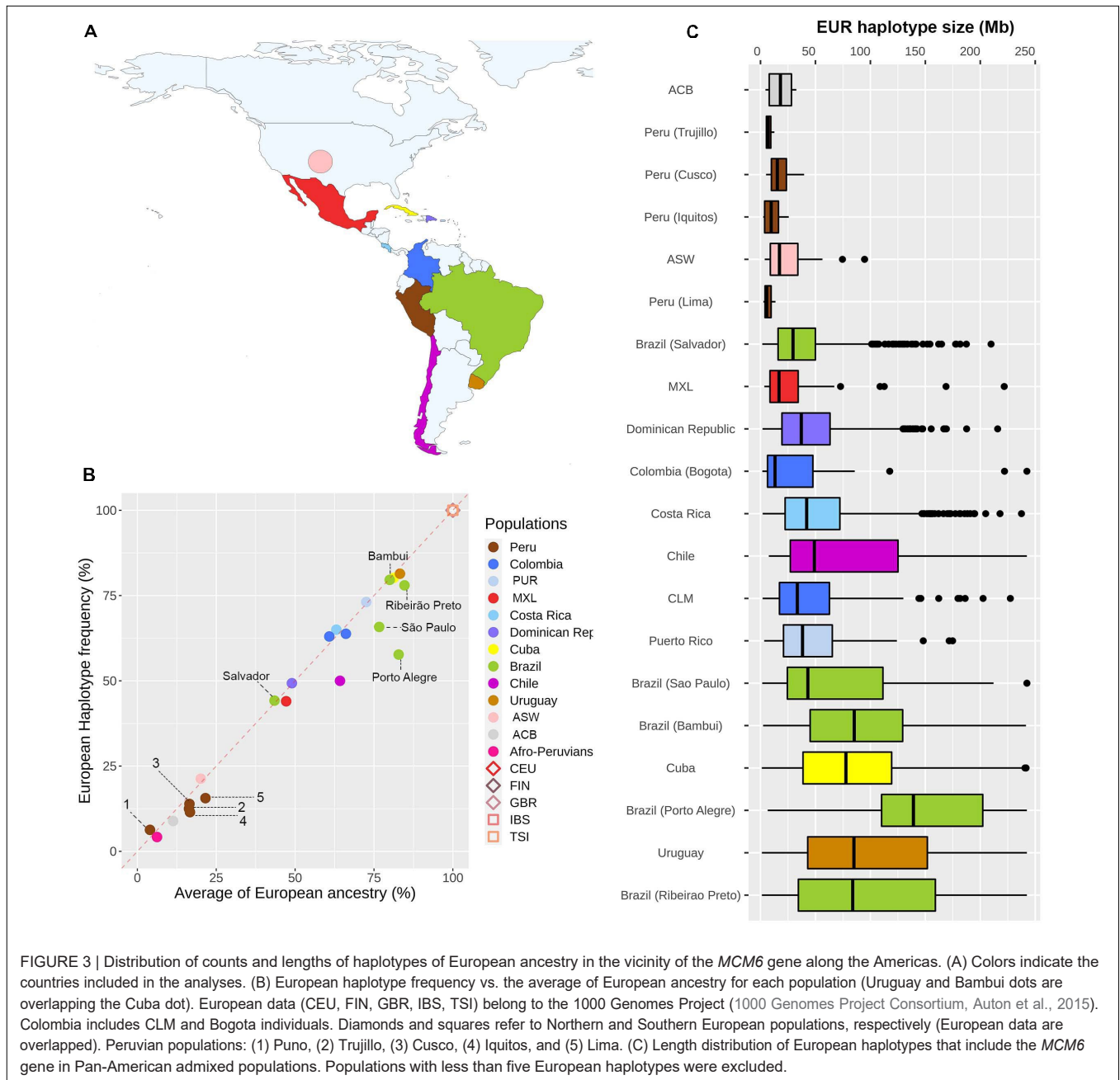
In the present study, we conducted the most geographically extensive analysis of the -13910^*T allele spanning several Pan-American admixed populations (North American African Descendants and Latin Americans), demonstrating high positive correlation between the distribution of the derived allele and European ancestries, with no evidence of post-admixture positive selection. Furthermore, in light of the current Latin American



nourishment policies, our results represent an important aspect to be considered for the understanding of the relationship between the consumption of dairy products and well-being across the continent.

We can address two limitations of our study. The first was the uneven representation of the population sampled within the countries: sample sizes varied from 13 individuals (Chile) to more than 1,000 individuals (Costa Rica, $n = 1,039$). However, if we restrict our analyses to countries with more

than 30 sampled individuals, one may still observe the patterns described. Second, to maintain a balance between individuals and SNP density, we used two separate datasets for the analyses (LARGE-PD and a merged dataset). However, as the study was focused on the $-13910C > T$ SNP and its flanking haplotypes of European origin, the genomic ancestry would still be the same if inferred with a different set of SNPs, as long as the inference is performed with the same SNP density.



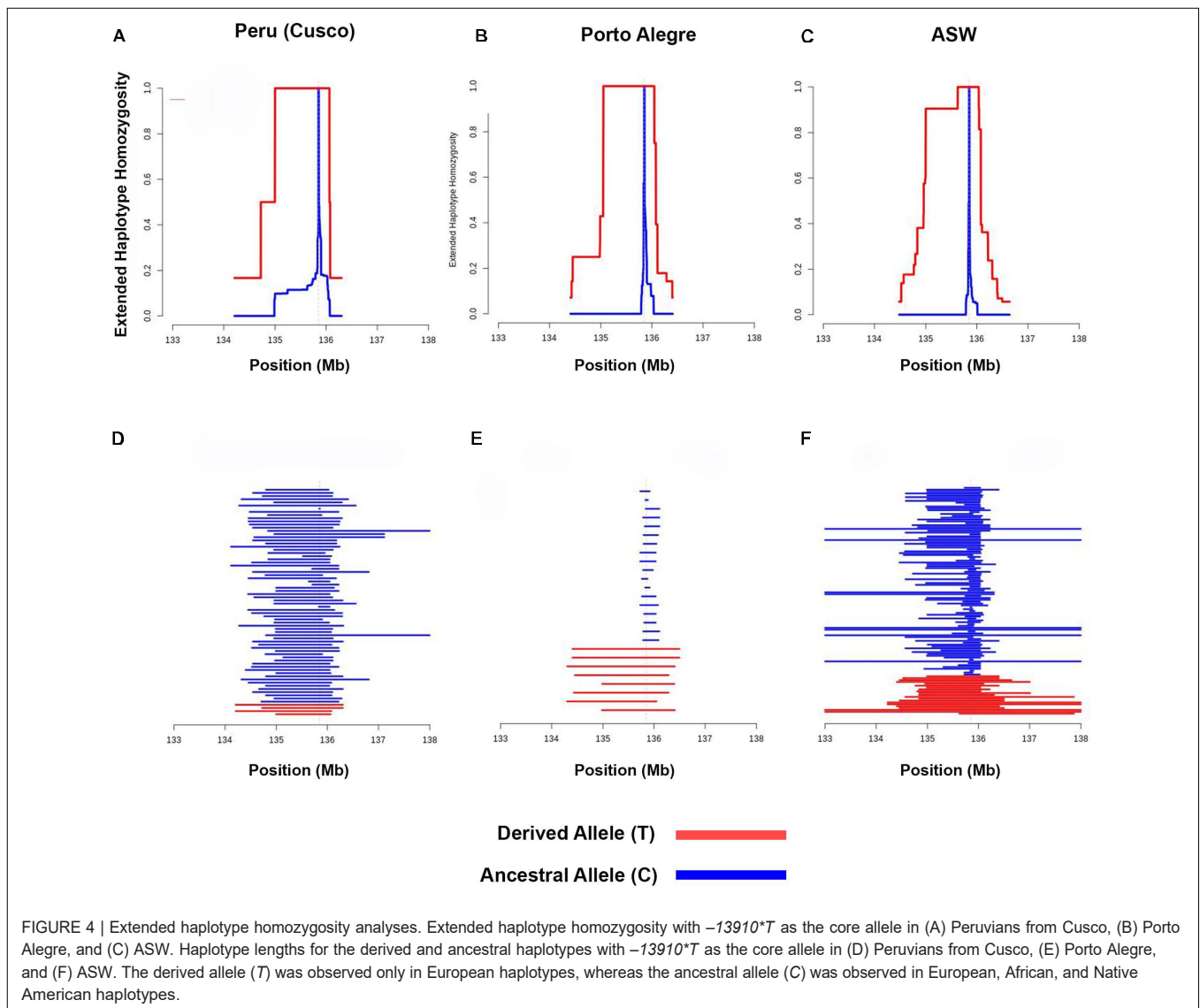
Impact of European Colonization in the Distribution of LP Alleles

European immigration had varied scales and influences across Central and South American regions. Despite occurring since the fifteenth century, European migration to the Americas was intensified during the late nineteenth century. Between the 1870s and the 1930s, approximately 13 million Europeans arrived in Latin America, of which more than 90% migrated to Argentina, Brazil, Cuba, and Uruguay (Sánchez-Alonso, 2019). In agreement, higher proportions of European ancestry have been observed in Brazilian populations (except for Salvador), Cubans, and Uruguayans. Also, the length distribution of European

haplotypes is wider in these populations, which is consistent with demographic data of a more continuous European gene flow.

Previous studies and historical records reported earlier European colonization in Northeast Brazil and higher proportions of African ancestry in this region (Instituto Brasileiro de Geografia e Estatística, 2000; Pena et al., 2011; Kehdy et al., 2015). We observed in the Salvador sample (Northeast Brazil) a low percentage of European ancestry correlated with a low

*-13910*T* allele frequency and a low European haplotype frequency. Moreover, a short European haplotype was observed, possibly introduced earlier in the Salvador population than in other Brazilian populations. A similar pattern was observed in



African descendant populations, in which the lowest average of European ancestry was detected. In African Americans from Southwest United States, the European ancestry corresponds to a minor proportion, which explains the low frequency of the -13910^*T allele in this population. Interestingly, the largest part of the European immigrants who came to the United States originated from countries where the frequencies of the -13910^*T allele are typically high (Bersaglieri et al., 2004; Smith et al., 2009; Itan et al., 2010), resulting in the presence of this variant in most of the European haplotypes observed in African Americans. The patterns observed in our work are consistent with one-way admixture events that have been inferred for these populations in previous studies (Harris et al., 2018; Ongaro et al., 2019; Gouveia et al., 2020).

In the admixed Peruvian populations, we also observed a low -13910^*T allele frequency, a low European haplotype frequency, and a short European haplotype – similarly to those

observed in African descendant populations, including Afro-Peruvians. This pattern is concordant with the hypothesis that the admixture between European colonizers (mainly Spanish) and Native Americans occurred mostly after the independence of Peru in 1824, about 300 years after the European settlement, resulting in the low European ancestry reported in Peruvian and Afro-Peruvian populations, which range between 14 and 22% (Harris et al., 2018).

Genetic studies indicated that the admixture between the Caribbeans and Europeans occurred shortly after the arrival of these populations in the American continent (Moreno-Estrada et al., 2013), resulting in the short European haplotypes observed in populations from Costa Rica, Dominican Republic, Barbados, and Puerto Rico. In contrast to these Central American countries, the wider European haplotype observed in Cubans was possibly due to the marked European migration to this country in the early twentieth century (Salzano and Bortolini, 2005).

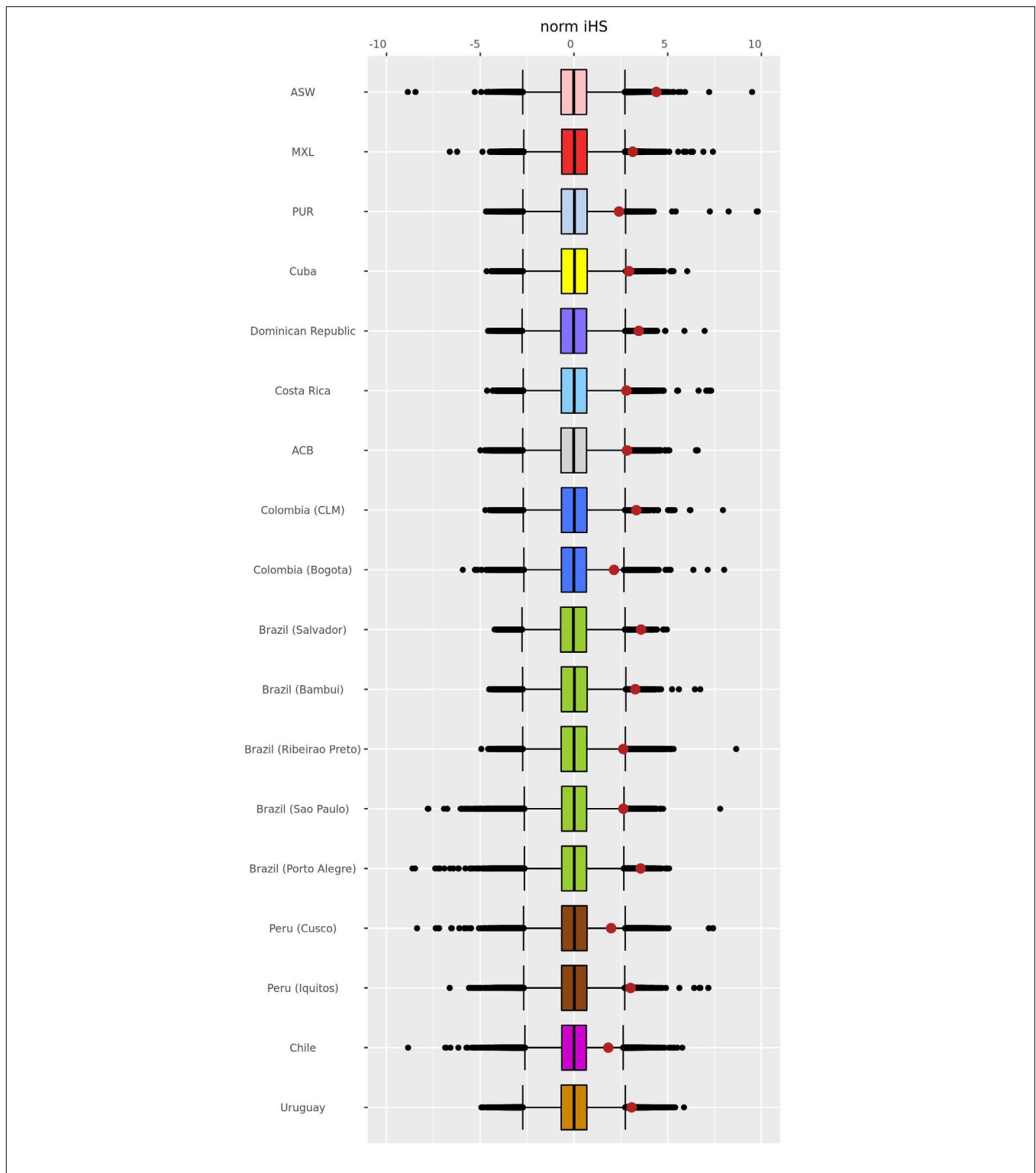


FIGURE 5 | Genome-wide distribution of normalized iHS Z-scores. The scores were estimated using the genome-wide distribution of the unstandardized iHS in Pan-American admixed populations with -13910^*T allelic frequencies $> 5\%$ and more than 10 individuals. After normalization, the mean and median are 0, and the standard deviation is 1. The whisker ends of the box plot indicate a value of ± 2.69 , delimiting 99.3% of the distribution. The middle line inside the box plot corresponds to the median of the normalized iHS Z-score distribution. The red dots refer to normalized iHS Z-scores for the -13910^*T allele. ASW, African Americans from Southwest United States; MXL, Mexican ancestry from Los Angeles, United States; PUR, Puerto Ricans from Puerto Rico; ACB, African Caribbeans from Barbados; CLM, Colombians from Medellin. Colors represent the countries just as the map in Figure 3A. Normalized iHS Z-scores and empirical p -values for the estimated -13910^*T allele are described in Supplementary Table 5.

Altogether, our results show that the distribution of lengths and frequencies related to the -13910^*T allele reflects the demographic history of Pan-American admixed populations.

LP in an Afro-Brazilian QuilomboCommunity

Between the sixteenth and mid-nineteenth centuries, Europeans brought to Brazil around 4 million enslaved Africans (Instituto Brasileiro de Geografia e Estatística, 2000). Africans and their descendants formed communities named Quilombos, a form of resistance to slavery and a stronghold for African culture (Moura, 2001; Raggio et al., 2018). The quilombola community of Sertão do Valongo (Santa Catarina State, Southern Brazil) has a reported African genomic ancestry of 68.1% and a European one of 31.9%, therefore with no reported Native American

ancestry (Luizon, 2007). The absence of the -13910^*T allele in Afro-Brazilians from Sertão do Valongo and, consequently, the absence of the inferred LP phenotype agree with the settlement history of the community marked by the small and early European contribution. Quilombola communities, in general, are not genetically isolated from other populations, although the degree of interaction with neighboring groups may vary (de Almeida, 2010). Hence, as each community has a distinct population structure, settlement history, and cultural practices, the evaluation of other quilombola communities is essential to the understanding of the -13910^*T allele distribution in these groups.

Lack of Postadmixture Selective Pressure for LP in the Pan-AmericanAdmixed Populations

The high correlation between European haplotype proportion and European genome-wide proportion suggests no selective pressure over the *MCM6* gene after admixture. Nonetheless, our EHH analysis showed that several populations have longer LD blocks containing the derived allele. Moreover, the *iHS* values for these populations were greater than 2, which indicates some level of positive selection (Voight et al., 2006). Considering that there is no overrepresentation of European haplotypes in the *MCM6* gene and that the length distribution of these haplotypes reflects demographic history, we hypothesized that the detected *iHS* signals result from positive selection in the source population. Specifically, African Americans, which had the highest *iHS* value, have Northern Europe as the main source of their admixture (Ongaro et al., 2019; Gouveia et al., 2020). Other populations with higher European ancestry levels (i.e., Brazilian populations) could have moderate *iHS* values due to the highly diverse European sources introduced into Brazil during the last century.

In contrast, by evaluating goat herders from Coquimbo (Chile), Montalva et al. (2019) reported a significant enrichment for European ancestry in the *LCT* gene when compared to genome-wide European ancestry, suggesting recent positive selection after admixture. However, the enrichment for European ancestry in the *LCT* gene was absent in urban non-pastoralist Latin American populations evaluated [e.g., MXL, CLM, and PEL (Peruvians in Lima, Peru) from the 1000 Genomes Project].

Thus, the lack of post-admixture selective pressure observed in our study – which datasets included MXL, CLM (1000 Genomes Project Consortium, Auton et al., 2015), and Peruvians (Harris et al., 2018; Borda et al., 2020) – agrees with Montalva et al. (2019), supporting their hypothesis of specific adaptation to milking agropastoralism in the Coquimbo population.

It should be noted that admixture processes could hinder the detection of selection signals due to changes in the allelic frequencies and linkage disequilibrium patterns (Lohmueller et al., 2011; Hamid et al., 2021). Furthermore, a recent study shows that a strong selective pressure of an adaptive phenotype could lead to genome-wide changes, modifying chromosome regions not directly involved in the phenotype, which can potentially bias the ancestry inference (Lohmueller et al., 2011; Hamid et al., 2021). Therefore, although it is unlikely that the LP phenotype has promoted differential survival in the American continent in the last centuries, we cannot ignore that the aspects early mentioned may have influenced our results.

Impact of Our Results on Public HealthPolicies

Dairy plays an essential role in human nutrition because of its supply of protein, lipids, and micronutrients, such as calcium, magnesium, and vitamins B₅ and B₁₂, among others (Muehlhoff et al., 2013). However, it is necessary to emphasize that, in recent years, the consumption benefits of dairy in adulthood have been questioned. Whereas some studies have linked its consumption to the development of cardiovascular diseases and obesity – possibly because of the fat content of milk (Segall, 1994; Berkey et al., 2005; Muehlhoff et al., 2013) – other studies have suggested that milk and dairy intake could reduce the risk of metabolic syndrome, diabetes, and cancer development (Elwood et al., 2007, 2008).

Dairy consumption is frequently encouraged – and even financed – by governmental initiatives to fight famine in Latin American countries due to these products' unarguable nutrient content. The “Programa de Abasto Social de Leche LICONSA” in Mexico (created in 1944), the “Programa Nacional de Alimentación Complementaria” or PNAC in Chile (created in 1952), the “Programa del Vaso de Leche” or PVL in Peru (created in 1984), and the “Programa de Aquisição de Alimentos” in Brazil (created in 2003) are examples of governmental initiatives (Uauy et al., 2001; Stifel and Alderman, 2006; Hespanhol, 2013; Morales-Ruán Mdel et al., 2013). Moreover, in Cuba, Costa Rica, México, Venezuela, Colombia, and Uruguay, dietary guidelines also recommend the consumption of dairy products (Instituto Nacional de Nutrición, 1990; Porrata-Maury, 2004; Comisión Intersectorial de Guías Alimentarias Para Costa Rica Ministerio de Salud, 2010; Secretaria de Salud de México, 2010; Instituto Colombiano de Bienestar Familiar and FAO Colombia, 2015; Dirección General de la Salud Ministerio de Salud de Uruguay, 2019). These initiatives follow the US program pattern, in which the dietary guidelines include milk as an essential nourishment element since its first publication (Bertron et al., 1999; Jacobs et al., 2020). However, the US programs relied on early nutrition scientific studies developed mostly in Northern European populations (Jacobs et al., 2020), where a high

prevalence of LP is observed (Ingram et al., 2009). Considering the prevalence of the inferred LNP phenotype in the Americas and the complex ancestry of Pan-American populations, dairy intake could contribute to the generation of health issues due to the development of lactose intolerance-associated events.

It is important to highlight that the gut microbiota composition and epigenetics modifications affect the manifestation of lactose intolerance-associated events, attenuating or eliminating them in individuals who typically cannot digest the disaccharide (Zhong et al., 2004; He et al., 2008; Labrie et al., 2016; Anguita-Ruiz et al., 2020). The manifestation of these events is also dependent on the lactose concentration of the food, which tends to be reduced in fermented dairy foods, such as cheese and yogurt (Buttriss, 1997). According to this, cultural and colonic adaptation mechanisms could explain the consumption of milk and dairy products without causing gastrointestinal events in non-carriers of genetic variants associated with LP (Hertzler and Savaiano, 1996; Segurel et al., 2020; Bleasdale et al., 2021). Thus, the absence of the -13910^*T allele alone may not be completely predictive of lactose intolerance, especially considering the existence of other alleles of non-European origin associated with the LP phenotype (Tishkoff et al., 2007; Ranciaro et al., 2014), although the low frequencies of such alleles in Pan-American admixed populations (Friedrich et al., 2012b; Paz-Y-Miño et al., 2016) reinforce the effectiveness of the -13910^*T allele for the genetic diagnosis of LP.

Our results revealed that, in the Americas, only a few populations from Cuba, Brazil (Bambui and Porto Alegre), and Uruguay have an elevated proportion (greater than 50%) of individuals who are likely to be lactase persistent, agreeing with the fact that these countries (except for Cuba) are among the largest consumers of dairy products (excluding butter) per year in the continent: 175.3 and 141.8 kg per capita for Uruguayans and Brazilians, respectively (Food and Agriculture Organization of the United Nations, 2018). A lower consumption in other Latin American countries (Food and Agriculture Organization of the United Nations, 2018) is associated with the prevalence of the LNP phenotype, which is consistent with the recent introduction of cattle and, consequently, with the consumption of dairy products in the continent (Gade, 1999). Therefore, traditional consumption of these products by Native American populations is rare (Kiple and Ornelas, 2001). However, it should be noted that European immigrants (and their descendants) have established such practices in culinary in America, resulting in the recent development of typically milk-based dishes in several countries of this continent (Albala, 2011; Wiley, 2015). The socioeconomic status of Latin American countries also contributes to such mentioned lower levels, as consumption of dairy products is known to be lower in developing countries than in developed ones, despite the changes over the last decades due to increasing dairy consumption (Bermudez and Tucker, 2003; Muehlhoff et al., 2013).

The reduced frequency of the -13910^*T allele may be a hint that dairy products are not the best option for dietary guidelines in populations with lower proportions of European ancestry, such as Peruvian and African descendant populations (African Americans, African Caribbeans, Afro-Brazilians from Curitiba,

and the quilombola community of Sertão do Valongo). According to the data discussion above, European-biased policies that include dairy products should be rediscussed and reconsidered. Furthermore, until the role of dairy intake in the pathogenesis of complex diseases is fully understood, alternative milk products and plant-based dairy substitutes would possibly be a better option for the dietary guidelines of America.

DATA AVAILABILITY STATEMENT

Newly generated sequences were deposited in the GenBank at NCBI website under the accession numbers MZ362598–MZ362856.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Brazilian CONEP (Comissão Nacional de Ética em Pesquisa). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MHB, CS, and ViB designed the study. AG and NS performed the Sanger sequencing and the analysis of Afro-Brazilian populations. MHB, MP-E, RL, and IR contributed with samples and reagents. CS, MD, J-YM, RK, KN, SW, MLB, ML-C, HG, OC, CP, ET-S, IM, ED, VR, AL, VT, VaB, HF, CR, AS-S, BS-L, PC-C, WF, GA, HA, and CA-B generated the datasets and were responsible by them. ViB, MHB, AG, NS, GA, MMe, TL, MS, CS, DL, MMA, and TO'C, analyzed the data. AG, NS, MHB, and ViB drafted the manuscript. GAC, MMe, TL, MP-E, TO'C, OC, MD, and ET-S critically edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by scholarships from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/PROAP—Finance Code 001), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Fundação Araucária, Brazil, provided to AG and NS. The Hispanic Community Health Study/Study of Latinos is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to The University of North Carolina (HHSN268201300001I/N01-HC-65233), University of Miami (HHSN268201300004I/N01-HC-65234), Albert Einstein College of Medicine (HHSN268201300002I/N01-HC-65235), the University of Illinois at Chicago (HHSN268201300003I/N01-HC-65236 Northwestern University), and San Diego State University (HHSN268201300005I/N01-HC-65237). The following

Institutes/Centers/Offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, and NIH Institution-Office of Dietary Supplements. Genome sequencing for “NHLBI TOPMed: Whole Genome sequencing in the Hispanic Community Health Study/Study of Latinos” (phs001395) was performed at the Baylor College of Medicine, Human Genome Sequencing Center (HHSN2682016000331). Core support, including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering, was provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN2682018000021). Core support, including phenotype harmonization, data management, sample-identity QC, and general program coordination was provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN2682018000011). RK was supported by R01-MD011389-01 from the National Institute on Minority Health and Health Disparities. LARGE-PD was supported by a Stanley Fahn Junior Faculty Award and an International Research Grants Program award from the

Parkinson’s Foundation, by a research grant from the American Parkinson’s Disease Association, and with resources and the use of facilities at the Veterans Affairs Puget Sound Health Care System. The Peruvian Genome Project was supported by the Peruvian National Institute of Health. MS was supported by training grant D43 TW007393 awarded by the Fogarty International Center of the US National Institutes of Health support.

ACKNOWLEDGMENTS

We would like to thank all the volunteers who participated in this study; Alessia Ranciaro for providing the sequence of the primers used; Priscila Ienzen dos Santos and Valter Antonio de Baura for technical support; and Kelly Nunes, Liana Alves de Oliveira, Márcia Regina Pincerati, and colleagues for the helpful discussions, comments, and support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.671079/full#supplementary-material>

REFERENCES

- 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., and Kang, H. M. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393
- Adhikari, K., Mendoza-Revilla, J., Chacón-Duque, J. C., Fuentes-Guajardo, M., and Ruiz-Linares, A. (2016). Admixture in Latin America. *Curr. Opin. Genet. Dev.* 41, 106–114. doi: 10.1016/j.gde.2016.09.003
- Albala, K. (2011). *Food Cultures of the World Encyclopedia*. Santa Barbara, CA: ABC-CLIO.
- Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109
- Anguita-Ruiz, A., Aguilera, C. M., and Gil, Á (2020). Genetics of lactose intolerance: an updated review and online interactive world maps of phenotype and genotype frequencies. *Nutrients* 12:2689. doi: 10.3390/nu12092689
- Auricchio, S., Rubino, A., Landolt, M., Semenza, G., and Prader, A. (1963). Isolated intestinal lactase deficiency in the adult. *Lancet* 2, 324–326. doi: 10.1016/s0140-6736(63)92991-x
- Berkey, C. S., Rockett, H. R. H., Willett, W. C., and Colditz, G. A. (2005). Milk, dairy fat, dietary calcium, and weight gain: a longitudinal study of adolescents. *Arch. Pediatr. Adolesc. Med.* 159, 543–550. doi: 10.1001/archpedi.159.6.543
- Bermudez, O. I., and Tucker, K. L. (2003). Trends in dietary patterns of Latin American populations. *Cad. Saude Publica* 19, S87–S99.
- Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., et al. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* 74, 1111–1120. doi: 10.1086/421051
- Bertron, P., Barnard, N. D., and Mills, M. (1999). Racial bias in federal nutrition policy, part I: the public health implications of variations in lactase persistence. *J. Natl. Med. Assoc.* 91, 151–157.
- Bleasdale, M., Richter, K. K., Janzen, A., Brown, S., Scott, A., Zech, J., et al. (2021). Ancient proteins provide evidence of dairy consumption in eastern Africa. *Nat. Commun.* 12:632.
- Boll, W., Wagner, P., and Mantei, N. (1991). Structure of the chromosomal gene and cDNAs coding for lactase-phlorizin hydrolase in humans with adult-type hypolactasia or persistence of lactase. *Am. J. Hum. Genet.* 48, 889–902.
- Borda, V., Alvim, I., Mendes, M., Silva-Carvalho, C., Soares-Souza, G. B., Leal, T. P., 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. doi: 10.1073/pnas.2013773117
- Buttriss, J. (1997). Nutritional properties of fermented milk products. *Int. J. Dairy Technol.* 50, 21–27. doi: 10.1111/j.1471-0307.1997.tb01731.x
- Casey, J. (2005). “Holocene occupations of the forest and savanna,” in *African Archaeology*, ed. A. B. Stahl (Oxford: Blackwell Publishing).
- Caskey, D. A., Payne-Bose, D., Welsh, J. D., Gearhart, H. L., Nance, M. K., and Morrison, R. D. (1977). Effects of age on lactose malabsorption in Oklahoma Native Americans as determined by breath H₂ analysis. *Am. J. Dig. Dis.* 22, 113–116. doi: 10.1007/bf01072952
- Chacón-Duque, J.-C., Adhikari, K., Fuentes-Guajardo, M., Mendoza-Revilla, J., Acuña-Alonso, V., Barquera, R., et al. (2018). Latin Americans show widespread Converso ancestry and imprint of local native ancestry on physical appearance. *Nat. Commun.* 9:5388.
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7.
- Comisión Intersectorial de Guías Alimentarias Para Costa Rica Ministerio de Salud. (2010). *Guías Alimentarias Para Costa Rica*. San José: Comisión Intersectorial de Guías Alimentarias Para Costa Rica.
- de Almeida, A. W. B. (2010). *Cadernos de Debates Nova Cartografia Social: Territórios Quilombolas e Conflitos*. Norwich: UEA.
- de Mello e Souza, M. (2007). *África e Brasil Africano*. Brzil: Editora Ática.
- Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., and Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* 10:5436.
- Dillehay, T. D. (2009). Probing deeper into first American studies. *Proc. Natl. Acad. Sci. U. S. A.* 106, 971–978. doi: 10.1073/pnas.0808424106
- Dirección General de la Salud Ministerio de Salud de Uruguay (2019). *Guía Alimentaria Para la Población Uruguaya: Para una Alimentación Saludable, Compartida y Placentera*. Montevideo: Ministerio de Salud de Uruguay.
- Duncan, I. W., and Scott, E. M. (1972). Lactose intolerance in Alaskan Indians and Eskimos. *Am. J. Clin. Nutr.* 25, 867–868. doi: 10.1093/ajcn/25.9.867

- Elwood, P. C., Givens, D. I., Beswick, A. D., Fehily, A. M., Pickering, J. E., and Gallacher, J. (2008). The survival advantage of milk and dairy consumption: an overview of evidence from cohort studies of vascular diseases, diabetes and cancer. *J. Am. Coll. Nutr.* 27, 723S–734S.
- Elwood, P. C., Pickering, J. E., and Fehily, A. M. (2007). Milk and dairy consumption, diabetes and the metabolic syndrome: the Caerphilly prospective study. *J. Epidemiol. Community Health* 61, 695–698. doi: 10.1136/jech.2006.053157
- Enattah, N. S., Sahi, T., Savilanti, E., Terwilliger, J. D., Peltonen, L., and Järvelä, I. (2002). Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* 30, 233–237. doi: 10.1038/ng826
- Fernández, C. I., Montalva, N., Arias, M., Hevia, M., Moraga, M. L., and Flores, S. V. (2016). Lactase non-persistence and general patterns of dairy intake in indigenous and mestizo Chilean populations. *Am. J. Hum. Biol.* 28, 213–219. doi: 10.1002/ajhb.22775
- Figueroa, R. B., Melgar, E., Jó, N., and García, O. L. (1971). Intestinal lactase deficiency in an apparently normal Peruvian population. *Am. J. Dig. Dis.* 16, 881–889. doi: 10.1007/bf02238168
- Food and Agriculture Organization of the United Nations (2018). *FAOSTAT Database. New Food Balances. FAOSTAT Database*. Available Online at: <http://www.fao.org/faostat/en/#data/FBS> (accessed February 02, 2021)
- Friedrich, D. C., Callegari-Jacques, S. M., Petzl-Erler, M. L., Tsuneto, L., Salzano, F. M., and Hutz, M. H. (2012a). Stability or variation? patterns of lactase gene and its enhancer region distributions in Brazilian Amerindians. *Am. J. Phys. Anthropol.* 147, 427–432. doi: 10.1002/ajpa.22010
- Friedrich, D. C., Santos, S., Ribeiro-dos-Santos, ÁK. C., and Hutz, M. H. (2012b). Several different lactase persistence associated alleles and high diversity of the lactase gene in the admixed Brazilian population. *PLoS One* 7:e46520. doi: 10.1371/journal.pone.0046520
- Gade, D. W. (1999). *Nature and Culture in the Andes*. Madison: University of Wisconsin press.
- Gautier, M., Klassmann, A., and Vitalis, R. (2017). rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol. Ecol. Resour.* 17, 78–90. doi: 10.1111/1755-0998.12634
- Goeblich, D. C., Waters, M. R., and O'Rourke, D. H. (2008). The late Pleistocene dispersal of modern humans in the Americas. *Science* 319, 1497–1502. doi: 10.1126/science.1153569
- Gouveia, M. H., Borda, V., Leal, T. P., Moreira, R. G., Bergen, A. W., Kehdy, F. S. G., et al. (2020). Origins, admixture dynamics, and homogenization of the African gene pool in the Americas. *Mol. Biol. Evol.* 37, 1647–1656. doi: 10.1093/molbev/msaa033
- Greenberg, J. H., Turner, C. G., Zegura, S. L., Campbell, L., Fox, J. A., Laughlin, W. S., et al. (1986). The settlement of the Americas: a comparison of the linguistic, dental, and genetic evidence [and comments and reply]. *Curr. Anthropol.* 27, 477–497. doi: 10.1086/203472
- Hamid, I., Korunes, K. L., Beleza, S., and Goldberg, A. (2021). Rapid adaptation to malaria facilitated by admixture in the human population of Cabo Verde. *Elife* 10:e63177.
- Harris, D. N., Song, W., Shetty, A. C., Levano, K. S., Cáceres, O., Padilla, C., et al. (2018). Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proc. Natl. Acad. Sci. U. S. A.* 115, E6526–E6535.
- He, T., Venema, K., Priebe, M. G., Welling, G. W., Brummer, R.-J. M., and Vonk, R. J. (2008). The role of colonic metabolism in lactose intolerance. *Eur. J. Clin. Invest.* 38, 541–547. doi: 10.1111/j.1365-2362.2008.01966.x
- Hertzler, S. R., and Savaiano, D. A. (1996). Colonic adaptation to daily lactose feeding in lactose maldigesters reduces lactose intolerance. *Am. J. Clin. Nutr.* 64, 232–236. doi: 10.1093/ajcn/64.2.232
- Hespanhol, R. A. M. (2013). Programa de Aquisição de Alimentos: limites e potencialidades de políticas de segurança alimentar para a agricultura familiar. *Soc. Nat. Resour.* 25, 469–483. doi: 10.1590/s1982-45132013000300003
- Ingram, C. J. E., Mulcare, C. A., Itan, Y., Thomas, M. G., and Swallow, D. M. (2009). Lactose digestion and the evolutionary genetics of lactase persistence. *Hum. Genet.* 124, 579–591. doi: 10.1007/s00439-008-0593-6
- Instituto Brasileiro de Geografia e Estatística (2000). *Brasil: 500 Anos de Povoamento*. Rio de Janeiro: IBGE.
- Instituto Colombiano de Bienestar Familiar and FAO Colombia. (2015). *Documento Técnico de las Guías Alimentarias Basadas en Alimentos Para la Población Colombiana Mayor de 2 Años*. Rome: FAO.
- Instituto Nacional de Nutrición (1990). *Guías de Alimentación Para Venezuela*. Mexico: Instituto Nacional de Nutrición.
- Itan, Y., Jones, B. L., Ingram, C. J. E., Swallow, D. M., and Thomas, M. G. (2010). A worldwide correlation of lactase persistence phenotype and genotypes. *BMC Evol. Biol.* 10:36. doi: 10.1186/1471-2148-10-36
- Jacobs, E. T., Foote, J. A., Kohler, L. N., Skiba, M. B., and Thomson, C. A. (2020). Re-examination of dairy as a single commodity in US dietary guidance. *Nutr. Rev.* 78, 225–234. doi: 10.1093/nutrit/nuz093
- John, R. (1997). *Africa: a Biography of the Continent*. London: Penguin Adult.
- Kadwell, M., Fernandez, M., Stanley, H. F., Baldi, R., Wheeler, J. C., Rosadio, R., et al. (2001). Genetic analysis reveals the wild ancestors of the llama and the alpaca. *Proc. Biol. Sci.* 268, 2575–2584. doi: 10.1098/rspb.2001.1774
- Kehdy, F. S. G., Gouveia, M. H., Machado, M., Magalhães, W. C. S., Horimoto, A. R., Horta, B. L., et al. (2015). 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. doi: 10.1073/pnas.1504447112
- Kiple, K. F., and Ornelas, K. C. (2001). *The Cambridge World History of Food*. Cambridge: Cambridge University Press.
- Labrie, V., Buske, O. J., Oh, E., Jeremian, R., Ptak, C., Gasiūnas, G., et al. (2016). Lactase nonpersistence is directed by DNA-variation-dependent epigenetic aging. *Nat. Struct. Mol. Biol.* 23, 566–573. doi: 10.1038/nsmb.3227
- Lahiri, D. K., and Nurnberger, J. I. (1991). A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies. *Nucleic Acids Res.* 19:5444. doi: 10.1093/nar/19.19.5444
- Latorre, G., Besa, P., Parodi, C. G., Ferrer, V., Azocar, L., Quirolo, M., et al. (2014). Prevalence of lactose intolerance in Chile: a double-blind placebo study. *Digestion* 90, 18–26. doi: 10.1159/000363229
- Liebert, A., López, S., Jones, B. L., Montalva, N., Gerbault, P., Lau, W., et al. (2017). World-wide distributions of lactase persistence alleles and the complex effects of recombination and selection. *Hum. Genet.* 136, 1445–1453. doi: 10.1007/s00439-017-1847-y
- Loesch, D., Andrea, R. V., Heilbron, K., Sarihan, E. I., Inca-Martinez, M., and Mason, E. (2020). Characterizing the genetic architecture of Parkinson's disease in Latinos. *Ann. Neurol.* doi: 10.1002/ana.26153
- Lohmueller, K. E., Bustamante, C. D., and Clark, A. G. (2011). Detecting directional selection in the presence of recent admixture in African-Americans. *Genetics* 187, 823–835. doi: 10.1534/genetics.110.122739
- Luizon, M. R. (2007). *Dinâmica da Mistura Étnica em Comunidades Remanescentes Quilombo Brasileiras. Tese de Doutorado, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto*. Available online at: <https://www.teses.usp.br/teses/disponiveis/17/17135/tde-31052011-092816/pt-br.php> Mantei, N., Villa, M., Enzler, T., Wacker, H., Boll, W., James, P., et al. (1988). Complete primary structure of human and rabbit lactase-phlorizin hydrolase: implications for biosynthesis, membrane anchoring and evolution of the enzyme. *EMBO J.* 7, 2705–2713. doi: 10.1002/j.1460-2075.1988.tb03124.x
- Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93, 278–288. doi: 10.1016/j.ajhg.2013.06.020
- Mendes, M., Alvim, I., Borda, V., and Tarazona-Santos, E. (2020). The history behind the mosaic of the Americas. *Curr. Opin. Genet. Dev.* 62, 72–77. doi: 10.1016/j.gde.2020.06.007
- Mendoza Torres, E., Varela Prieto, L. L., Villarreal Camacho, J. L., and Villanueva Torregroza, D. A. (2012). Diagnosis of adult-type hypolactasia/lactase persistence: genotyping of single nucleotide polymorphism (SNP C/T-13910) is not consistent with breath test in Colombian Caribbean population. *Arq. Gastroenterol.* 49, 5–8. doi: 10.1590/s0004-28032012000100002
- Montalva, N., Adhikari, K., Liebert, A., Mendoza-Revilla, J., Flores, S. V., Mace, R., et al. (2019). Adaptation to milking agropastoralism in Chilean goat herders and nutritional benefit of lactase persistence. *Ann. Hum. Genet.* 83, 11–22. doi: 10.1111/ahg.12277
- Morales-Ruán Mdel, C., Shamah-Levy, T., Mundo-Rosas, V., Cuevas-Nasu, L., Romero-Martínez, M., Villalpando, S., et al. (2013). [Food assistance programs in Mexico, coverage and targeting]. *Salud Publica Mex.* 55, S199–S205.
- Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J. L., Byrnes, J. K., Gignoux, C. R., et al. (2013). Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* 9:e1003925.
- Moura, C. (2001). *Os Quilombos na Dinâmica Social do Brasil*. Maceio, AL: EDUFAL.






- Muehlhoff, E., Bennett, A., and McMahon, D. (2013). *Milk and Dairy Products in Human Nutrition*. Rome: FAO.
- Nakatsuka, N., Lazaridis, I., Barbieri, C., Skoglund, P., Rohland, N., Mallick, S., et al. (2020). A paleogenomic reconstruction of the deep population history of the andes. *Cell* 181, 1131–1145.e21.
- Ojeda-Granados, C., Panduro, A., Rebello Pinho, J. R., Ramos-Lopez, O., Gleyzer, K., Malta, F. M., et al. (2016). Association of lactase persistence genotypes with high intake of dairy saturated fat and high prevalence of lactase non-persistence among the mexican population. *J. Nutrigenet. Nutrigenomics* 9, 83–94. doi: 10.1159/000446241
- Ongaro, L., Scliar, M. O., Flores, R., Raveane, A., Marnetto, D., Sarno, S., et al. (2019). The genomic impact of european colonization of the Americas. *Curr. Biol.* 29, 3974–3986.e4.
- Paz-Y-Miño, C., Burgos, G., López-Cortés, A., Herrera, C., Gaviria, A., Tejera, E., et al. (2016). A study of the molecular variants associated with lactase persistence in different Ecuadorian ethnic groups. *Am. J. Hum. Biol.* 28, 774–781. doi: 10.1002/ajhb.22865
- Pena, S. D. J., Di Pietro, G., Fuchshuber-Moraes, M., Genro, J. P., Hutz, M. H., Kehdy, F. S. G., et al. (2011). The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS One* 6:e17063. doi: 10.1371/journal.pone.0017063
- Pena, S. D. J., Santos, F. R., and Tarazona-Santos, E. (2020). Genetic admixture in Brazil. *Am. J. Med. Genet. C Semin. Med. Genet.* 184, 928–938. doi: 10.1002/ajmg.c.31853
- Porrata-Maury, C. (2004). “Guías alimentarias para la población cubana mayor de dos años de edad.” in *Educación Alimentaria y Nutricional e Higiene de los Alimentos. Manual de Capacitación*, ed. Ministerio de Salud pública (Havana: Ministerio de Salud Pública).
- Posth, C., Nakatsuka, N., Lazaridis, I., Skoglund, P., Mallick, S., Lamnidis, T. C., et al. (2018). Reconstructing the deep population history of central and South America. *Cell* 175, 1185–1197.e22.
- Primo, A. T. (1992). El ganado bovino ibérico en las Américas: 500 años después. *Arch. Zootec.* 41:13.
- Raggio, A. N. A. Z., Bley, R. B., and Trauczynski, S. C. (eds) (2018). *Abordagem Histórica Sobre a População Negra no Estado do Paraná*. Paraná: SEJU.
- Ranciaro, A., Campbell, M. C., Hirbo, J. B., Ko, W.-Y., Froment, A., Anagnostou, P., et al. (2014). Genetic origins of lactase persistence and the spread of pastoralism in Africa. *Am. J. Hum. Genet.* 94, 496–510. doi: 10.1016/j.ajhg.2014.02.009
- Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., et al. (2012). Reconstructing native American population history. *Nature* 488, 370–374.
- Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., et al. (2014). Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet.* 10:e1004572. doi: 10.1371/journal.pgen.1004572
- Salzano, F. M. (2011). The prehistoric colonization of the Americas: evidence and models. *Evol. Educ. Outreach* 4, 199–204. doi: 10.1007/s12052-011-0330-9
- Salzano, F. M., and Bortolini, M. C. (2005). *The Evolution and Genetics of Latin American Populations*. Cambridge: Cambridge University Press.
- Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.
- Sánchez-Alonso, B. (2019). The age of mass migration in Latin America: the age of mass migration in latin America. *Econ. Hist. Rev.* 72, 3–31. doi: 10.1111/ehr.12787
- Secretaría de Salud de México (2010). *Guía de Alimentos Para la Población Mexicana. {Secretaría de Salud de Mexico}*. Available Online at: <https://www.yumpu.com/es/document/read/63974919/guia-de-alimentos-para-la-poblacion-mexicana>
- Segall, J. J. (1994). Dietary lactose as a possible risk factor for ischaemic heart disease: review of epidemiology. *Int. J. Cardiol.* 46, 197–207. doi: 10.1016/0167-5273(94)90242-9
- Ségurel, L., and Bon, C. (2017). On the evolution of lactase persistence in humans. *Annu. Rev. Genomics Hum. Genet.* 18, 297–319. doi: 10.1146/annurev-genom-091416-035340
- Ségurel, L., Guarino-Vignon, P., Marchi, N., Lafosse, S., Laurent, R., Bon, C., et al. (2020). Why and when was lactase persistence selected for? insights from central Asian herders and ancient DNA. *PLoS Biol.* 18:e3000742. doi: 10.1371/journal.pbio.3000742
- Smith, G. D., Lawlor, D. A., Timpson, N. J., Baban, J., Kiessling, M., Day, I. N. M., et al. (2009). Lactase persistence-related genetic variant: population substructure and health outcomes. *Eur. J. Hum. Genet.* 17, 357–367. doi: 10.1038/ejhg.2008.156
- Soares-Souza, G., Borda, V., Kehdy, F., and Tarazona-Santos, E. (2018). Admixture, genetics and complex diseases in Latin Americans and US Hispanics. *Curr. Genet. Med. Rep.* 6, 208–223. doi: 10.1007/s40142-018-0151-z
- Souza, I. R., and Culp, L. (1992). Valongo, an isolated brazilian black community. i. structure of the population. *Rev. Bras. Genét.* 15, 439–447.
- Souza, I. R., and Culp, L. (2005). Valongo, genetic studies on an isolated Afro-Brazilian community. *Genet. Mol. Biol.* 28, 402–406. doi: 10.1590/s1415-47572005000300012
- Stifel, D., and Alderman, H. (2006). The “glass of milk” subsidy program and malnutrition in peru. *World Bank Econ. Rev.* 20, 421–448. doi: 10.1093/wber/lhl002
- Szpiech, Z. A., and Hernandez, R. D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* 31, 2824–2827. doi: 10.1093/molbev/msu211
- Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature* 590, 290–299.
- Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39, 31–40. doi: 10.1038/ng1946
- Uauy, R., Albalá, C., and Kain, J. (2001). Obesity trends in Latin America: transiting from under- to overweight. *J. Nutr.* 131, 893S–899S.
- Valencia, L., Randazzo, A., Engfeldt, P., Olsson, L. A., Chávez, A., Buckland, R. J., et al. (2017). Identification of novel genetic variants in the mutational hotspot region 14 kb upstream of the LCT gene in a Mexican population. *Scand. J. Clin. Lab. Invest.* 77, 311–314. doi: 10.1080/00365513.2017.1318445
- Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72. doi: 10.1371/journal.pbio.0040072
- Wang, Y., Harvey, C., Rousset, M., and Swallow, D. M. (1994). Expression of human intestinal mRNA transcripts during development: analysis by a semiquantitative RNA polymerase chain reaction method. *Pediatr. Res.* 36, 514–521. doi: 10.1203/00006450-199410000-00018
- Wehling, A., de Macedo Wehling, M. J. C., and da Silva, J. L. W. (1994). *Formação do Brasil Colonial*. Rio de Janeiro: Nova Fronteira.
- Wheeler, J. C. (1995). Evolution and present situation of the South American camelidae. *Biol. J. Linn. Soc. Lond.* 54, 271–295. doi: 10.1016/0024-4066(95)90021-7
- Wiley, A. S. (2015). *Cultures of Milk: the Biology and Meaning of Dairy Products in the United States and India*. Cambridge: Harvard University Press.
- Zhong, Y., Priebe, M. G., Vonk, R. J., Huang, C.-Y., Antoine, J.-M., and He, T. (2004). The role of colonic microbiota in lactose intolerance. *Dig. Dis. Sci.* 49, 78–83. doi: 10.1023/b:ddas.0000011606.96795.40

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Guimarães Alves, Sukow, Adelman Cipolla, Mendes, Leal, Petzl-Erler, Lehtonen Rodrigues Souza, Rainha de Souza, Sanchez, Santolalla, Loesch, Dean, Machado, Moon, Kaplan, North, Weiss, Barreto, Lima-Costa, Guio, Cáceres, Padilla, Tarazona-Santos, Mata, Dieguez, Raggio, Lescano, Tumas, Borges, Ferraz, Rieder, Schumacher-Schuh, Santos-Lobato, Chana-Cuevas, Fernandez, Arboleda, Arboleda, Arboleda-Bustos, O’Connor, Beltrame and Borda. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A large Canadian cohort provides insights into the genetic architecture of human hair colour

Frida Lona-Durazo ¹✉, Marla Mendes^{1,2}, Rohit Thakur^{3,4}, Karen Funderburk³, Tongwu Zhang ^{3,4}, Michael A. Kovacs³, Jiyeon Choi ³, Kevin M. Brown ³ & Esteban J. Parra ¹✉

Hair colour is a polygenic phenotype that results from differences in the amount and ratio of melanins located in the hair bulb. Genome-wide association studies (GWAS) have identified many loci involved in the pigmentation pathway affecting hair colour. However, most of the associated loci overlap non-protein coding regions and many of the molecular mechanisms underlying pigmentation variation are still not understood. Here, we conduct GWAS meta-analyses of hair colour in a Canadian cohort of 12,741 individuals of European ancestry. By performing fine-mapping analyses we identify candidate causal variants in pigmentation loci associated with blonde, red and brown hair colour. Additionally, we observe colocalization of several GWAS hits with expression and methylation quantitative trait loci (QTLs) of cultured melanocytes. Finally, transcriptome-wide association studies (TWAS) further nominate the expression of *EDNRB* and *CDK10* as significantly associated with hair colour. Our results provide insights on the mechanisms regulating pigmentation biology in humans.

¹Department of Anthropology, University of Toronto at Mississauga, Mississauga, Ontario, Canada. ²Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG 31270-901, Brazil. ³Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA. ⁴Integrative Tumor Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA.

✉email: frida.lonadurazo@mail.utoronto.ca; esteban.parra@utoronto.ca

Pigmentary traits in humans (hair, eye and skin pigmentation) have a polygenic architecture, in which pleiotropy and epistasis are common phenomena^{1–5}. Contrary to other complex traits, the environment has little or no effect on pigmentation traits, with the exception of facultative skin pigmentation (i.e. tanning ability)^{6,7}. Therefore, the wide range of diversity in pigmentation traits across worldwide populations is mainly due to genetic variation that has been shaped by a range of evolutionary factors (i.e., drift, gene flow and selection)^{8–12}. Elucidating the genetic architecture of pigmentation traits may aid in the general understanding of biological pathways, gene interactions, their regulation and expression. Likewise, it has the potential to further identify molecular mechanisms associated to important diseases, such as skin cancer (e.g., basal cell carcinoma, squamous cell carcinoma and melanoma).

Current evidence indicates that there is a partial overlap in the genetic architecture of hair, eye and skin pigmentation. Some genes (e.g., *OCA2*, *TYR*, *SLC24A5*) have been associated with multiple pigmentation phenotypes, whereas other genes have been associated with only one of these traits (e.g., *MFSD12*—skin pigmentation)¹³. Adding to this complexity, previous studies have identified the presence of allelic heterogeneity, in which different variants within a single gene are associated with pigmentation variation in different populations (e.g., *MFSD12*, *OCA2*)^{5,13,14}, as well as the effect of multiple independent variants on the same locus associated with a diverse range of pigmentation tones within populations (e.g., *HERC2/OCA2*, *MC1R*, *GRM5/TYR*)^{5,15–19}. Furthermore, some genes associated with pigmentation phenotypes (e.g., *ASIP*, *MC1R*, *TYR*, *SLC45A2*, *OCA2*, *IRF4*, *SLC24A4*) are also known to increase the risk of cutaneous melanoma^{20–26}. Thus, the link between pigmentation and cancer risk also highlights the biomedical importance of efforts to characterise the genetic architecture of pigmentation phenotypes²⁷. Hair colour is a quantitative phenotype that results from differences in the amount and ratio of eumelanin and pheomelanin synthesised in melanocytes located in the hair bulb, which then migrate to the hair shaft^{28–31}. Recently, large-scale genome-wide association studies (GWAS) have uncovered numerous loci associated with hair colour in people of European ancestry by using discrete hair colour categories as an approximation^{32,33}. By analyzing the UK Biobank (UKBB) data, both of these studies identified hundreds of regions associated with the phenotype in question across the genome, some of which had not been previously identified, due to the lack of power to detect significant associations (e.g., *TSPAN10*, *FRMD5*).

Furthermore, Morgan and colleagues provided a detailed analysis of the loci associated with red hair colour, including penetrance and interactions among single-nucleotide polymorphisms (SNPs), offering insights on the genetic complexity of this hair colour tone³³. In spite of these advances, for most of the hair pigmentation-associated loci, the causal variants and the molecular mechanisms underlying pigmentation variation remain to be identified³⁴. This is in fact a challenging task, given that most of the top SNPs associated with hair colour are located in non-protein-coding regions of the genome, with no obvious or direct function on the trait, thus hinting to a regulatory function. Major advances in the characterisation of regulatory elements, along with the development of a diverse set of computational tools using GWAS summary statistics, have facilitated the interpretation of GWAS hits of several polygenic phenotypes^{34,35}. For instance, statistical fine-mapping methods have become computationally feasible and make it possible to model multiple causal variants simultaneously³⁶. Additionally, when putative causal variants are identified, they can be further explored by evaluating their effect on the regulatory profile of target genes on relevant

cell types, which can be statistically tested with colocalization or transcriptome-wide association studies (TWAS) approaches^{37,38}. In this study, we conducted a meta-analysis of genome-wide association studies including 12,741 Canadian participants of European-related ancestry from the Canadian Partnership for Tomorrow's Health. We focused our efforts on identifying candidate causal variants and target genes in complex genomic regions known to alter hair colour, by performing a wide range of post-GWAS analyses. Our main outcomes include the identification of multiple candidate causal variants through fine-mapping of significant loci across distinct regions of the genome, including putative causal variants not previously reported for hair colour, and the colocalization of GWAS loci with gene expression and methylation quantitative trait loci (eQTL and meQTL, respectively) using cultured human primary melanocytes. Finally, we conducted transcriptome-wide association studies (TWAS) of hair colour with cultured melanocyte expression data.

Results

Hair colour distribution. 12,996 participants of the Canadian Partnership for Tomorrow's Health (CanPath) project, who were genotyped using different genome-wide genotyping arrays (See Methods for details), self-reported their natural hair colour (before greying) using six possible answers: black ($N = 824$), dark brown ($N = 5,818$), light brown ($N = 4,429$), blonde ($N = 1,410$), red ($N = 306$) hair colour, NA ($N = 30$). After quality control of the genotypes (i.e. exclusion of poor-quality samples and PCA outliers) we kept 12,741 individuals for further analyses.

The distribution of hair colour categories is similar across all provinces sampled (Fig. 1a; Supplementary Data 1), with black and red hair colour being the least frequent categories and brown (light and dark) being the most common one. In addition, on average across all provinces, the proportion of females with black hair colour is lower, whereas other hair colour categories are proportional between sexes (Fig. 1b; Supplementary Data 1), which is similar to what has been previously reported^{33,39,40}.

Genome-wide association studies and meta-analyses. We performed GWAS of hair colour on each of the five genotyping arrays (genotyped and imputed SNPs) using a logistic mixed model on SAIGE (version 0.38)⁴¹. Specifically, we tested the following models: 1) blonde vs. brown (light and dark) + black hair colour; 2) brown (light and dark) vs. black hair colour; and 3) red vs. brown (light and dark) + black hair colour. For the third model (red vs. brown + black hair colour), the number of individuals with red hair colour on two of the genotyping arrays (GSA 24v1 and Omni 2.5) was less than 20, therefore we excluded these two arrays from the analysis (Supplementary Table 1). In our GWAS models, we included sex, age and the first 10 principal components (PCs) as fixed effects. We did not detect residual population substructure, based on Q-Q plots, in which observed p values did not show an early deviation from the expected p values (Supplementary Figs. 1–3), and the inflation factor (λ : highest value = 1.06; mean = 1.008).

We carried out three meta-analyses using the summary statistics (log of odds ratio and standard error) of each of the GWAS on METASOFT v2.0.1⁴². In total, the number of individuals included on each meta-analysis was: blonde vs. brown + black hair colour: $N = 12,398$; brown vs. black hair colour: $N = 10,990$; and red vs. brown + black hair colour: $N = 10,450$. Q-Q plots of the meta-analyses indicated no inflation of p values (Supplementary Fig. 4). Additionally, the LD Score regression intercepts computed on LDSC (version 1.0.1) were 1.001, 0.994 and 0.999 for

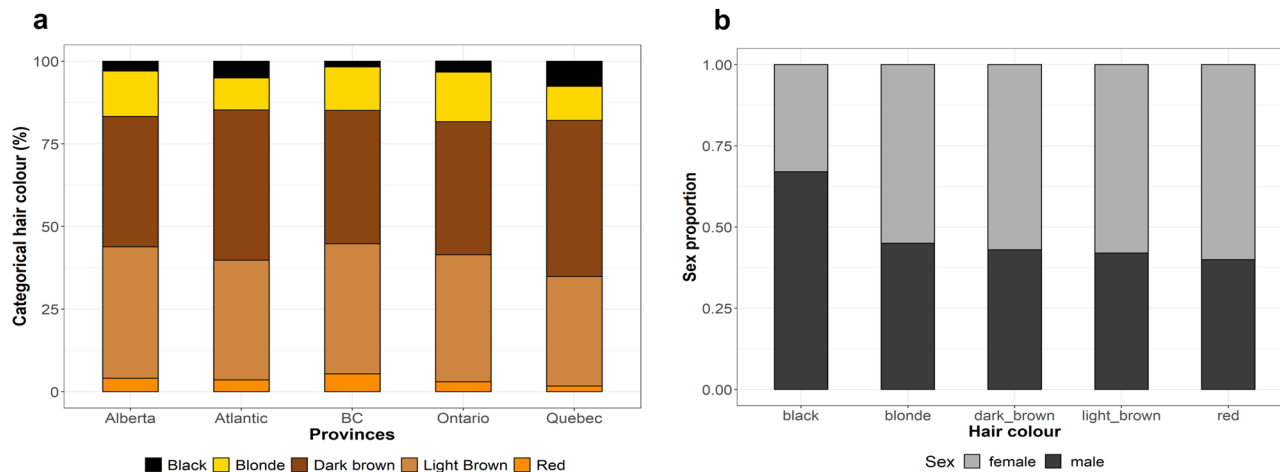


Fig. 1 Distribution of the hair colour categories in the CanPath cohorts. a Percentage of each hair colour category, stratified by province. b Proportion of sexes across the different hair colour categories.

the three models tested, respectively, indicating no residual confounding bias.

Our meta-analyses identified genome-wide significant loci (p value $\leq 1.67e-08$) overlapping or near genes known to affect normal pigmentation variation (Fig. 2; Supplementary Figs. 5 and 6). Supplementary Data 2–4 summarise the suggestive and genome-wide associated SNPs for each meta-analysis. On the blonde hair colour meta-analysis (Fig. 2a; Supplementary Fig. 5; Supplementary Data 2), the most significant locus was *OCA2/HERC2* on chromosome (Chr) 15 (lead SNP: rs12913832; p value = $5.03e-141$; OR = 0.304; 95% CI = 0.389–0.614). Other genome-wide significant loci overlapped *SLC45A2* (Chr 5), *IRF4* (Chr 6), *TPCN2*

(Chr 11), *KITLG* (Chr 12), *SLC24A4* (Chr 14) and *MC1R* (Chr 16). Similarly, on the brown hair colour meta-analysis (Fig. 2b; Supplementary Fig. 5; Supplementary Data 3), the lowest p -value corresponds to a SNP within *HERC2* (rs1129038; p value = $3.36e-52$, OR = 0.411; 95% CI = 0.366–0.461), which is in high LD with rs12913832. Other genome-wide significant regions overlap *SLC45A2* (Chr 5) and *IRF4* (Chr 6).

On the red hair colour meta-analysis (Fig. 2c; Supplementary Fig. 5; Supplementary Data 4), the only genome-wide significant locus was *MC1R* (Chr 16, lead SNP: rs12931267, p value = $2.29e-82$, OR = 0.014; 95% CI = 0.009–0.022), a gene known for its loss-of-function mutations, which switches the production of eumelanin to pheomelanin^{43–45}.

Our linear mixed model meta-analysis, in which hair colour categories ranged from blonde up to black (excluding red hair colour), yielded similar results as the logistic mixed models for blonde hair colour (Supplementary Data 5). Particularly, the regions (*TYR*, *EDNRB*, *BNC2* and *ASIP*) which were not genome-wide significant in the meta-analyses using logistic mixed models, reached genome-wide significance here (Supplementary Fig. 7). In addition, a locus that did not show up in the logistic mixed model meta-analyses (*ARL15*) was genome-wide significant. This gene has only been previously identified in the UKBB hair pigmentation study³³. Given that we linearly regressed ordinal categories, we cannot assume that the differences between categories are equal. For this reason, and considering the similar results obtained with the logistic and linear mixed model approaches, we restricted downstream analyses to the meta-analyses based on the binary logistic mixed models.

Investigating candidate causal variants. We first investigated if the signals for hair colour across significant loci on our CanPath

meta-analyses were being driven by one or more independent SNPs, by conducting approximate conditional and joint analyses of association on GCTA (GCTA-COJO)⁴⁶.

For blonde hair colour, there was one SNP selected for each of the seven genome-wide significant loci (Supplementary Table 2), some of which are known functional SNPs, such as the rs12913832 (Chr 15) and rs12203592 (Chr 6) variants located within enhancers. Similarly, the brown hair colour conditional analysis highlighted one SNP on each of the genome-wide significant loci (Supplementary Table 2). Notably, the SNP selected on chromosome 15 at the *HERC2* locus is rs1129038, which is a SNP in high LD with rs12913832.

We identified five independent SNPs for red hair colour on chromosome 16, overlapping or near the *MC1R* region. A few of these markers show evidence of heterogeneity (Supplementary Table 2), but have a similar effect size and p -value on the fixed and random-effects models. This result is in agreement with previous studies, as it is well known that there are multiple loss-of-function mutations on this region affecting hair colour variation^{33,44,47,48}. Finally, we performed the approximate conditional analyses a second time using a different reference LD matrix from a subset of our CanPath samples (See Methods for details). Notably, there were no overall differences in the results obtained with either matrices for blonde or brown hair colour, in respect to the independent SNPs per locus and overall summary statistics (Supplementary Data 6). In the case of red hair colour, there were a total of seven SNPs independently associated in the locus on chromosome 16.

We then carried out Bayesian fine-mapping of these loci to identify candidate causal SNPs using FINEMAP v1.4⁴⁹. Compared to GCTA's stepwise conditioning approach, which depends on arbitrary p -value thresholds, Bayesian fine-mapping quantifies the probability of causality by jointly modelling simultaneous effects of multiple SNPs, and considers genotype probabilities for calculating LD³⁶. Each credible set contains a minimum set of candidate causal SNPs with a probability of at least 95%, and we kept the candidate causal variants from each credible set that had considerable evidence of causality (i.e. SNPs with $\log_{10}BF \geq 2$), and included their respective annotations with SNP Nexus^{50,51} (See Methods for details). Based on the combined evidence of fine-mapping and annotation, we defined the candidate causal variants with strong evidence of causality as the most likely candidate causal variants. We have summarised the results for each hair colour model (Supplementary Table 3; Supplementary Data 7–9).

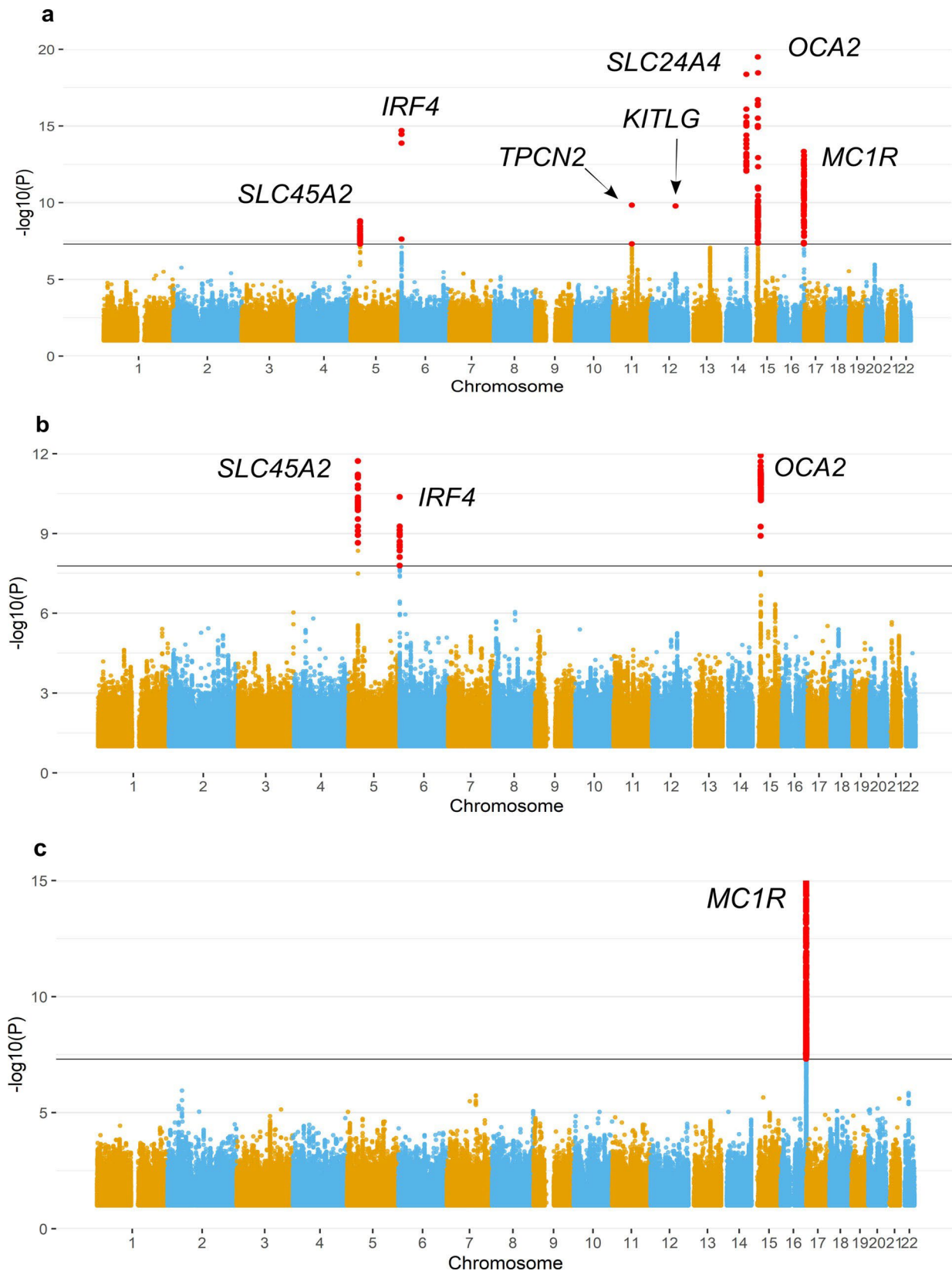


Fig. 2 Manhattan plots of hair colour meta-analyses based on logistic mixed models. a Blonde vs. brown + black hair colour ($n = 12,398$ individuals). b Brown vs. black hair colour ($n = 10,990$ individuals). c Red vs. brown + black hair colour ($n = 10,450$ individuals). The continuous black line denotes the genome-wide significant threshold ($p = 1.66e-8$). Markers in red are genome-wide significant. The Y-axis has been limited to truncate strong signals of association. The full figure is available as Supplementary Fig. 8.

Bayesian fine-mapping analyses highlighted known functional pigmentation SNPs, in line with the results obtained with GCTA-COJO, such as rs12203592 within *IRF4* (Chr 6), which is the most likely candidate causal SNP in the locus, with high posterior probability (identified in the blonde and brown hair colour analyses) (Supplementary Fig. 8; Supplementary Table 3). On the *SLC45A2* locus (Chr 5), the known missense SNP rs16891982 is present within the 95% credible set, although the SNP with the highest posterior probability is rs35391. On the *SLC24A4* locus (Chr 14), the marker highlighted by GCTA-COJO (rs12896471) is among the most likely candidate causal SNPs. This marker is in perfect LD ($r^2 = 1$) with a SNP previously associated with blonde hair colour (rs12896399)⁵², and both appear in the same 95% credible set.

Different from GCTA-COJO, on the *TPCN2* region (Chr 9), we identified the missense SNP rs3829241 as one of the most likely candidates of causality in the locus (Supplementary Fig. 9). Additionally, the SNP rs72932523, highlighted by GCTA-COJO, also has considerable evidence of causality in an independent credible set (Supplementary Fig. 9). This marker is in high LD ($r^2 = 0.83$) with the missense SNP rs72928978. Another two known missense variants (rs3750965 and rs35264875) within *TPCN2* appear in the same 95% credible set as rs3829241, but with a $\log_{10}\text{BF} < 2$, suggesting little evidence of causality.

On the *HERC2/OCA2* region (Chr 15), the only SNP with considerable evidence of causality for blonde hair colour, based on its posterior probability and annotation, is the known regulatory SNP rs12913832 (PIP = 0.978). In contrast, there are two causal signals identified for brown hair colour highlighted by FINEMAP (Supplementary Fig. 10). Similar to the blonde hair colour analysis, one of the credible sets include rs12913832 (PIP = 0.577). The second credible set includes candidate causal SNPs within *OCA2* with low PIP (< 0.5), in which one of them has $\log_{10}\text{BF} > 2$ (rs7168800) (Supplementary Table 3).

Finally, the *MC1R* region (Chr 16) was associated with both blonde and red hair colour. FINEMAP highlighted three causal signals in the locus for blonde hair colour and ten causal signals for red hair colour. The credible sets include known missense SNPs within or near *MC1R* (rs1805005, rs1805007, rs1805008, rs1805009), in which rs1805005 had the highest PIP and $\log_{10}\text{BF}$ on the blonde hair colour analysis (PIP = 0.984; $\log_{10}\text{BF} = 4.970$) (Supplementary Fig. 11); this same SNP does not appear in the red hair colour credible sets. Similarly, the missense SNP rs1805009 is not present in the blonde hair colour credible sets. Additionally, most candidate causal SNPs in the credible sets for red hair colour had a PIP > 0.9 and $\log_{10}\text{BF}$ well above 2 (Supplementary Data 9). When comparing the *MC1R* locus results obtained with GCTA with these results, we noticed an overlap of the candidate causal SNPs, specifically of the missense SNPs, whereas the synonymous SNPs highlighted by each method differ.

Exploring gene expression and methylation using cultured melanocyte data. We conducted colocalization analyses of the GWAS meta-analyses genome-wide signals using hypercoloc⁵³ with gene expression and methylation *cis*-QTLs (eQTLs, meQTLs, respectively) to explore the putative regulatory role of the SNPs identified in our hair colour GWAS and identify candidate genes (Table 1—See Methods for details). We observed colocalization of GWAS signals with eQTLs of *SLC45A2* and *OCA2* (marked by SNPs rs35391 and rs12913832, respectively). Additionally, we also observed colocalization of a GWAS signal with an eQTL of *SLC24A4*, with a regional probability = 1, but there was no candidate SNP selected, suggesting limited evidence of a single colocalized SNP between the traits. In addition, we

identified five GWAS loci colocalized with meQTLs, associated with methylation of CpGs near *MC1R*, *OCA2*, *IRF4*, *SLC45A2* and *SLC24A4*. Notably, we found GWAS colocalization with both eQTL and meQTL in three of these loci (*SLC24A4*, *SLC45A2* and *OCA2* regions). In the case of *SLC24A4*, the SNP rs8022442 is a meQTL for CpG probes, cg11086312 and cg10004481. On the *OCA2* locus, rs12913832 is a meQTL for CpG probes located upstream and downstream of the SNP within *HERC2* (cg05271345, cg25622125 and cg27374167). Furthermore, the SNP rs35391 is a meQTL for the CpG probes in the first exon of *SLC45A2* region (cg14189614 and cg04302388).

We did not find colocalization of QTLs on or near the gene *MC1R* for red hair colour. Given the current evidence, this is likely explained by the fact that known loss-of-function polymorphisms within *MC1R* lead to red hair colour, therefore, they have a direct functional role on the translated protein. However, there is a possibility that we might be missing eQTLs beyond the 500 kb tested region. Nonetheless, we did identify colocalizing meQTLs (rs258322) for blonde hair colour, associated with the CpG methylation near *MC1R*: on or near *CDK10*, *GAS8* and *DPEP1* (cg05714116, cg06907930 and cg00996377),

which may point at an independent regulatory region associated with blonde hair colour, apart from the known missense SNP within *MC1R* (rs1805005)³⁹. Additionally, we observed a colocalized signal between GWAS and meQTL near *IRF4* for both blonde and brown hair colour (cg23785612). Finally, neither of these loci (i.e., *MC1R*, *IRF4*) harbour corresponding eQTLs.

Lastly, we conducted transcriptome-wide association studies (TWAS) using the GWAS summary statistics, and we imputed the expression profile based on the CanPath cohort LD, from the expression weights calculated from the cultured melanocytes RNA-seq data⁵⁴. Similar to the colocalization results, the decreased expression of *OCA2* and *SLC24A4* was significantly associated with blonde hair colour (Table 2; Supplementary Fig. 12). In contrast, the increased expression of *EDNRB* was associated with blonde hair colour, which was not identified through colocalization analyses. In this case, the direction of effect in both the eQTL and GWAS is negative, which is the opposite of what was expected, given that the protein encoded by *EDNRB* is involved in melanocyte development and it induces melanocyte proliferation⁵⁵. This discrepancy could be due to different direction of effects in skin and hair melanocytes, similar to the inverse effect of *IRF4* effect on hair and skin pigmentation⁵⁶. However, further investigation of the *EDNRB* expression patterns in the hair bulb is needed to provide a clear explanation.

The decreased expression of the gene *RIN3* near *SLC24A4* was also significantly associated with blonde hair colour. After performing conditional TWAS analysis to check if these two signals were independent from each other, we found that the gene *RIN3* signal is not independent from that of the nearby gene *SLC24A4* (Supplementary Fig. 13). Lastly, the decreased expression of *CDK10* was significantly associated with red hair colour, which contrasts our colocalization results, in which we did not identify eQTLs for red hair colour. The gene *CDK10* is near *MC1R*, but to the extent of our knowledge, this gene is not implicated in pigmentation. However, it is important to note that TWAS results do not imply gene causality, and further evidence and replication of this signal is needed to have a conclusive result.

Biological pathway gene set enrichment analysis. We performed a gene set analysis to identify relevant biological pathways, using the results of our candidate causal SNP analysis using FINEMAP (Supplementary Table 4; Fig. 3; Supplementary Data 10), for all hair colour loci jointly. Focusing on the gene ontology (GO)

Table 1 Colocalization results of expression and methylation *cis*-QTLs from cultured melanocytes (eQTL and meQTL, respectively) with GWAS SNPs on each hair colour category.

Chromosome	Candidate SNP	Posterior probability	Regional probability	Posterior explained by SNP	Gene/methylation annotation	QTL
Blonde Hair Colour						
5	rs35391	0.83	0.85	0.43	<i>SLC45A2</i>	eQTL
5	rs35391	0.98	0.99	0.54	<i>SLC45A2</i> OpenSea	meQTL
5	rs35391	0.96	0.97	0.55	OpenSea	meQTL
6	NA	NA	0.95	NA	OpenSea	meQTL
14	NA	NA	1.00	NA	<i>SLC24A4</i>	eQTL
14	rs8022442	0.99	1.00	1.00	OpenSea	meQTL
14	rs8022442	0.95	0.95	1.00	<i>SLC24A4</i> (Body) OpenSea	meQTL
15	rs12913832	1.00	1.00	1.00	<i>OCA2</i>	eQTL
15	rs12913832	0.99	0.99	0.94	AC090696.2	eQTL
15	rs12913832	1.00	1.00	0.99	<i>HERC2</i> (Body) OpenSea	meQTL
15	rs12913832	0.98	0.99	0.98	<i>HERC2</i> (Body) S_Shelf	meQTL
15	rs12913832	0.98	0.98	0.98	<i>HERC2</i> (Body) S_Shore	meQTL
16	rs258322	1.00	1.00	1.00	<i>CDK10</i> (TSS1500) N_Shore	meQTL
16	rs258322	1.00	1.00	1.00	LOC100130015 (Body); <i>GAS8</i> (3' UTR) OpenSea	meQTL
16	rs258322	0.94	0.94	1.00	<i>DPEP1</i> (5' UTR) OpenSea	meQTL
Brown Hair Colour						
5	rs35391	0.83	0.85	0.57	<i>SLC45A2</i>	eQTL
5	rs35391	0.98	0.99	0.65	<i>SLC45A2</i> (1st Exon) OpenSea	meQTL
5	rs35391	0.97	0.97	0.66	Open Sea	meQTL
6	rs7773997	0.94	1.00	0.94	Open Sea	meQTL
15	rs12913832	1.00	1.00	0.98	<i>OCA2</i>	eQTL
15	rs1129038	0.99	0.99	0.54	AC090696.2	eQTL
15	rs12913832	1.00	1.00	0.87	<i>HERC2</i> (Body) OpenSea	meQTL
15	rs12913832	0.98	0.98	0.73	<i>HERC2</i> (Body) S_Shelf	meQTL
15	rs12913832	0.97	0.97	0.74	<i>HERC2</i> (Body) S_Shore	meQTL

We show colocalized SNPs with a posterior probability of ≥ 0.8 . We tested all the significant eQTL genes or meQTL probes within ± 250 kb regions flanking the GWAS lead SNPs. The Gene/Methylation Annotation indicates the location of CpG probes with respect to the nearest gene, as well as relative to CpG island. NA = limited evidence of a single SNP driving the colocalization.

Table 2 Genome-wide significant genes in the TWAS of the hair colour categories: blonde, brown and red.

Gene	Chr	GWAS best SNP	GWAS Z-score	eQTL best SNP	eQTL Z-score	# of SNPs	# weighted SNPs	TWAS Z-score	TWAS p-value
Blonde Hair Colour									
<i>EDNRB</i>	13	rs7330412	-5.4	rs7330412	-3.89	1204	1	5.404	6.52E-08
<i>RIN3*</i>	14	rs12896399	11.51	rs12893289	-6.14	986	4	-11.3069	1.21E-29
<i>SLC24A4*</i>	14	rs12896399	11.51	rs61977801	-6.68	933	11	-11.1083	1.14E-28
<i>OCA2</i>	15	rs12913832	-25.28	rs12913832	6.24	489	1	-25.282	5.04E-141
Brown Hair Colour									
<i>OCA2</i>	15	rs1129038	-15.2	rs12913832	6.24	489	1	-15.203	3.38E-52
Red Hair Colour									
<i>CDK10</i>	16	rs1805007	18.67	rs11538871	-5.69	987	5	-7.23615	4.62E-13

Genome-wide significant threshold: p-value $\leq 4.17 \times 10^{-6}$. GWAS/eQTL best SNP is the most significant SNP on each analysis. Chr chromosome. *Signals are not independent from each other, as evidenced by conditional TWAS.

processes, most gene sets correspond to pigmentation processes (e.g., developmental pigmentation, melanin biosynthetic process, melanocyte differentiation) or are a parental process of a pigmentation-related process (e.g., secondary metabolite biosynthetic process, phenol-containing compound biosynthetic process).

Notably, the DNA repair process includes genes surrounding the *MC1R* gene (i.e., *FANCA*, *SPIRE2*), as well as *HERC2*. It is well known that *MC1R* signalling reduces UV-induced DNA damage by mediating a cascade of reactions after the activation of cAMP-dependent Protein Kinase A (PKA)⁵⁷. Likewise, loss-of-function mutations on *MC1R* diminish the UV-induced DNA damage repair⁵⁷. *HERC2* is known to be involved in DNA repair induced by UV³⁸, but the expression of this gene likely does not

play a role in hair colour variation, as the associated SNPs are within introns, serving as enhancers to regulate the expression of *OCA2*.

Pigmentation traits, such as red/blonde hair, fair skin and response to sun exposure, are risk factors for skin cancer (melanoma and nonmelanoma)^{27,59}. We assessed if our genome-wide significant signals associated with hair colour have also been significantly associated with sun tanning (i.e., response to UV radiation exposure), melanoma, basal cell carcinoma (BCC) and squamous cell carcinoma (SCC), based on the databases available in the NHGRI-EBI GWAS Catalog⁶⁰. We tested a total of 327 SNPs across all phenotypes (melanoma: 153, BCC: 103, SCC: 29, and sun tanning: 42 SNPs), of which 21 unique SNPs overlap our GWAS hits (melanoma: 13, BCC: 5, SCC: 9, and sun tanning: 5 SNPs).

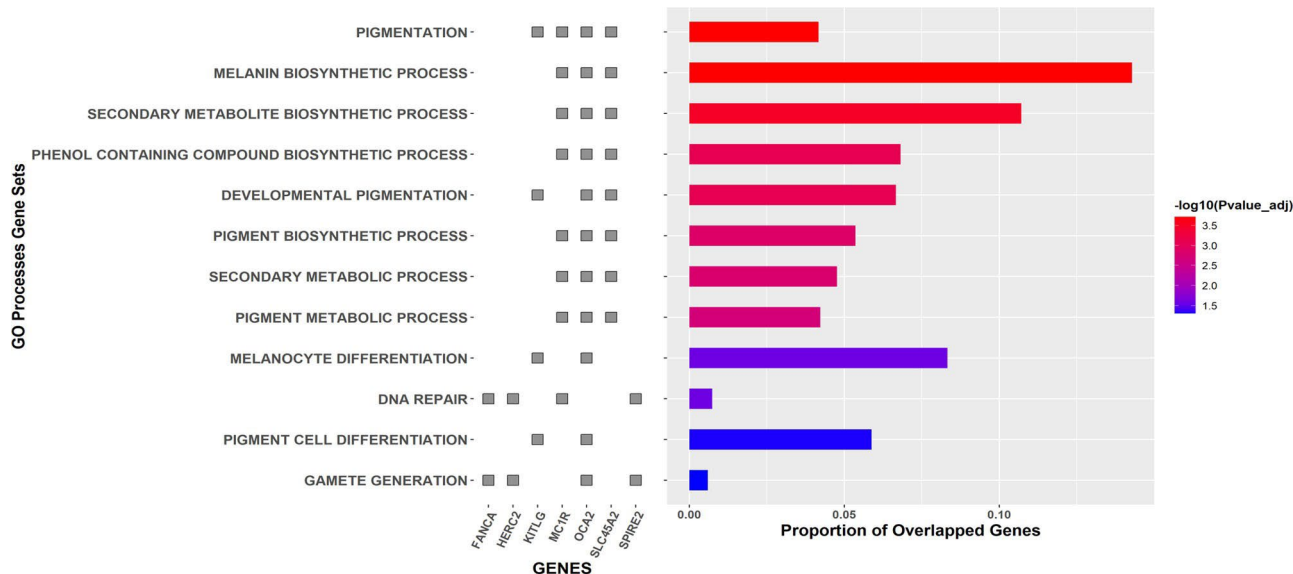


Fig. 3 Enrichment pathway analysis of genome-wide significant hair colour signals. Analysis performed with FUMA using Gene Ontology processes for all genes significantly associated with hair colour, as indicated by FINEMAP results ($n = 13$ genes).

These SNPs are distributed across key pigmentation loci that overlap genes, such as *SLC45A2*, *IRF4*, *SLC24A4*, *OCA2*, *HERC2* and *MC1R* (Supplementary Table 5). However, we note that an overlap of significant signals is not indicative of a shared causal signal between traits, therefore the biological relevance of these loci on skin cancer should be functionally investigated.

Discussion

In this study, we conducted meta-analyses of genome-wide association studies of hair colour on a Canadian cohort of European ancestry. Furthermore, we performed several post-GWAS analyses to investigate in more detail the genetic architecture of this trait, by focusing on the genome-wide associated loci. By applying statistical fine-mapping methods, we identified pigmentation loci harbouring more than one causal signal with high confidence, similar to what has been previously observed^{5,19,61}. Additionally, we have incorporated colocalization analyses of expression and methylation QTLs, as well as TWAS, both of which provide insights on the regulatory mechanisms of hair colour variation and pigmentation biology.

We note that hair colour is a naturally quantitative trait, which shows substantial variation as a result of differences in the amount of melanin in the hair, as well as differences in the ratio between the two types of melanin, eumelanin and pheomelanin. In an ideal scenario, quantification of the phenotype on a large sample would yield more accurate association results given the nature of the trait. However, recent studies from large-scale biobanks have shown a drastic increase in power to detect significant associations when the sample sizes increase, even if a self-reported, qualitative approximation of hair colour is used^{5,32,33}. To investigate the genetic architecture of hair colour categories, we relied on our binary logistic mixed model analyses, in which we identified several known genes involved in pigmentation variation, such as *SLC45A2*, *KITLG*, *SLC24A4*, *OCA2*, *MC1R*, among others. However, by using a linear GWAS model based on four eumelanin ordinal categories (i.e. blonde, light brown, dark brown, black), we replicated the signal of the gene *ARL15* that was previously identified in a recent UKBB study³³ (Supplementary Data 5).

We identified through fine-mapping analysis known pigmentation variants, which are likely candidates of causality, such as

the missense SNP rs16891982 on *SLC45A2*, rs12821256 near *KITLG*, rs12203592 on *IRF4* and rs12913832 on *HERC2*, the latter three located within known enhancers affecting the transcription of *IRF4*, *KITLG*, and *OCA2*, respectively^{56,62–65}.

Our fine-mapping analyses highlighted several regions that harbour more than one causal signal. On the *TPCN2* locus, we identified two likely causal signals that include or are in high LD with missense SNPs: rs3829241 and rs72932523 (in LD with rs72928978). It is known that *TPCN2* encodes a cation transporter channel in melanosomes that regulates the pH of melanosomes to downregulate melanogenesis^{66,67}. There are three nonsynonymous SNPs (rs3829241, rs35264875, and rs3750965) associated with blonde hair colour reported in the GWAS literature^{33,52}. Two of these nonsynonymous SNPs (rs3829241 and rs35264875) have been experimentally shown to modify the pH of the melanosome, an important factor for tyrosinase's activity⁶⁸.

Additionally, it is possible that the missense SNP rs72928978, which has been recently associated with eye colour⁶⁹, may also have a functional effect on the TPC2 protein, as predicted by SIFT and PolyPhen (deleterious and possibly damaging, respectively). In addition, the A-allele of rs72928978 shows high population differentiation: it is common only in European and European admixed populations of the UKGP (mean MAF in EUR = 0.14)

and it is absent from other continental populations of the UKGP (Supplementary Fig. 14). Therefore, future experimental analyses on the missense SNP rs72928978 may provide more details regarding its functional effect on melanin synthesis.

The *HERC2* rs12913832 SNP is one of the most important determinants of eye colour variation in human populations, and more particularly blue vs. non-blue eye colour. This SNP is located within an enhancer on an intron of *HERC2*, which regulates the expression of the downstream gene *OCA2*⁶². We report here that rs12913832 is both an expression and methylation QTL in melanocytes. However, we cannot be certain of a correlation between the meQTL target CpG and *OCA2* expression, given the current evidence.

We were particularly interested in studying in more detail the *OCA2/HERC2* region due to previous evidence pointing to several independent candidate variants in this region^{5,18,19}. In addition to the signal described above for rs12913832, our fine-mapping

results indicate that there is evidence of another candidate causal signal associated with brown hair colour on the *OCA2* gene. The second credible set comprises SNPs within *OCA2*, including the intronic SNPs rs72714118 and rs72714121 (PIP = 0.12 and 0.10, respectively), both of which overlap histone marks identified in foreskin melanocytes (H3K27ac, H2K4me1) associated with enhancer signals⁷⁰. This result contrasts to observations by Adhikari *et al.*, in which they detect a secondary variant associated with eye and skin pigmentation on the *HERC2/OCA2* locus, but not with hair colour⁵.

The red hair colour phenotype is mainly a consequence of missense polymorphisms on the *MC1R* locus, some of which have high penetrance and are termed “R” variants (rs1805007, rs1805008, rs1805009, rs1805006) and others that have low penetrance and are termed “r” variants (rs1805005, rs2228479, rs885479). Compared to “R” variants, which impair the function of the protein and lead to the synthesis of pheomelanin in homozygote or compound heterozygote state, “r” variants only reduce the protein’s efficiency, leading to low levels of eumelanin synthesis^{17,71}. We have identified here candidate causal R and r variants associated with red hair colour (rs1805008, rs1805009), but also with blonde hair colour (rs1805005, rs1805007, rs1805008). The SNP rs1805005 is one of the most likely candidate causal SNPs, although there is a high probability of additional causal signals in the region. In fact, a SNP within *CDK10* (rs258322) is a colocating meQTL for blonde hair colour. ThemeQTL is in moderate LD with a candidate causal SNP with $\log_{10}BF > 2$ (rs75570604; $r^2 = 0.62$).

These results provide insights about possible regulatory variants leading to blonde hair colour. Within the solute carrier family, there are at least three transmembrane proteins involved in ion transport (*SLC24A5*, *SLC45A2* and *SLC24A4*), which are also involved in normal pigmentation variation. Light skin pigmentation in people of European ancestry is driven mainly by two nonsynonymous mutations (rs1426654 and rs16891982) on

SLC24A5 and *SLC45A2*, respectively^{2,14,64,72}. Notably, our colocalization analyses suggest that there may be additional SNPs in the *SLC45A2* region regulating the expression of the gene.

Several reports have identified the association of *SLC24A4* with hair and eye colour in the same population^{2,47,73–75}, highlighting an upstream SNP (~15 kb from the transcription start site) with the largest effect (rs12896399), which is a common variant in all 1KGP continental populations, except in African populations. Based on the Genotype-Tissue Expression (GTEx) Project, rs12896399 is significantly associated with the expression of *SLC24A4* in skin tissue (skin not sun-exposed; p value = 2.3e-6). However, to the extent of our knowledge, there is no clear evidence of the molecular process by which this SNP regulates the expression of *SLC24A4*. By conducting colocalization and TWAS analyses, we identified both an eQTL and meQTL in the *SLC24A4* locus, in which the candidate meQTL is in moderate LD with the fine-mapped candidate causal SNP rs12896471 ($r^2 = 0.65$).

Our colocalization results highlighted meQTLs for blonde hair colour, associated with the methylation of CpGs near known pigmentation genes (i.e., *MC1R*, *IRF4*). These loci do not harbour colocating eQTLs, which suggests that other mechanisms may be involved, such as *trans*-QTLs⁷⁶, which were not considered in the current analysis. Alternatively, it is possible that some CpG probes capture the status of poised enhancers (i.e., enhancers in a *latent* state), which may not yet have any influence on gene expression in actively growing melanocytes. This is a possible scenario, given that the melanocytes used in the QTL analyses were from newborns. However, none of the candidate colocating SNPs in these loci (rs258322 and rs7773997) are eQTLs of *MC1R* and *IRF4*, respectively, in adult skin tissue based on the GTEx Project. Experimental histone modification marker assays may

provide support for the alternative hypothesis, as it is known that poised enhancers lose H3K27me3 and acquire acetylation at the same amino acid residue upon activation⁷⁷.

Several of the genome-wide significant SNPs we identified in the CanPath cohort overlap SNPs associated with different types of skin cancer and response to UV radiation, and it is particularly the case for melanoma. Additionally, a few of the overlapped SNPs are also colocating eQTLs and/or meQTLs (e.g., rs4904871), which highlights the importance of investigating the genetic and epigenetic mechanisms involved in the pigmentation pathway, such as hair colour. In fact, a recent study has provided a first glance into the epigenetic mechanisms (i.e., DNA methylation and gene expression) of pigmentation genes mediating skin cancer using whole blood tissue⁷⁸. By applying summary-based Mendelian randomisation and colocalization analyses, they colocating 9 DNA methylation sites (DNAm) with pigmentation traits (skin cancer, hair colour and sun exposure), as well as with the expression of genes.

We followed-up their QTLs (Table 3 of Bonilla *et al.*⁷⁸) on our colocalization results, but none of their SNPs was present in our colocalization results. The differences may lie in the fact that we used cultured melanocytes, which provide a cell-specific expression and methylation profile, best suited for the traits being tested⁷⁶. However, it is relevant to note that the authors also successfully colocating a DNAm site near the gene *CDK10* with blonde hair colour, as well as with several skin cancer traits (i.e., melanoma, basal cell carcinoma), which further reinforces the evidence we reported here, regarding putative regulatory variants in that locus.

Overall, our results indicate that the performance of GCTA-COJO and FINEMAP is largely concordant, although the credible sets and posterior probabilities computed by FINEMAP, combined with appropriate annotations, provide a broader approach to prioritise the most likely causal variants for further functional validation. It is worth noting that our analyses focused on SNPs due to the nature of the data and imputation approach, therefore we may be missing important structural variants that contribute to pigmentation variation, such as small indels. Finally, the lack of sex chromosome data also poses a limitation in the current study. In conclusion, by taking advantage of a relatively large cohort like the CanPath, we conducted GWAS meta-analyses of hair colour in which we identified candidate causal variants and provided insights into the genetic architecture that modulates hair colour variation. Many of these variants also affect other pigmentation traits, such as normal skin pigmentation variation, tanning response, as well as different types of skin cancer. We took advantage of expression and methylation data to characterise nonprotein-coding GWAS hits, and we believe that further experimental assays that include other epigenetic elements will provide further details on the genomic mechanisms regulating pigmentation variation. Our results provide insights on the general mechanisms regulating pigmentation biology in humans.

Methods

Canadian partnership for tomorrow’s health participants. This study was approved by the University of Toronto Ethics Committee (Human Research Protocol # 36429) and data access was granted by the Canadian Partnership for Tomorrow’s Health (Application number DAO-034431). All relevant ethical regulations were followed, and informed consent was obtained from CanPath participants. The samples in this study correspond to a subset of 12,996 individuals from the Canadian Partnership for Tomorrow’s Health (CanPath), which were sampled in different provinces: Alberta ($N = 969$; 7.45%), Atlantic Coast Provinces (i.e., New Brunswick, Newfoundland, Nova Scotia and Prince Edward Island) ($N = 937$; 7.21%), British Columbia ($N = 986$; 7.59%), Ontario ($N = 941$; 7.24%), and Quebec ($N = 9,163$; 70.51%). We selected individuals who self-reported having European-related ancestry and for whom self-reported hair colour was available. Among all participants included here, 53.78% were females and the average age was 53 years old (SE \pm 7.85).

Genotyping of participants and quality control. An overview of the methodological workflow is shown in Supplementary Fig. 15. Individuals who self-reported as having European-related ancestry were genotyped between 2012 and 2018 using five different genotyping array chips: (i) Axiom 2.0 UK Biobank (Affymetrix) ($N = 4,821$), (ii) Global Screening Array (GSA, Illumina) 24v1 ($N = 438$), (iii) 24v2 + MDP ($N = 2,594$), (iv) 24v1 + MDP ($N = 4,617$), and (v) Omni 2.5 (Illumina) ($N = 526$) by the Canadian Partnership for Tomorrow's Health (CanPath) project. The number of single-nucleotide polymorphisms (SNPs) of these chip arrays ranges between 626,377 and 2,349,746 SNPs.

We performed genotype quality control for each array chip separately by first filtering out variants that deviated in minor allele frequency >0.2 from the 1000 Genomes Project Phase 3 European sample (1KGP-EUR), GC/TA variants with minor allele frequency >0.4 in the 1KGP-EUR and flipping alleles according to the 1KGP-EUR, using a Perl script (version 4.2)⁷⁹. Afterwards, we used PLINK (version 1.9)^{80,81} to filter out variants with minor allele frequency $<1\%$, high missing genotyping rate ($-\text{geno } 0.05$), high missing individual rate ($-\text{mind } 0.05$) or variants that significantly deviated from the Hardy-Weinberg Equilibrium (HWE) ($-\text{hwe } 1e-06$). Then, we also identified second-degree relatives ($-\text{genome_PI_HAT } > 0.2$) using a pruned set of variants in linkage disequilibrium (LD) ($-\text{indep-pairwise } 100 \ 10 \ 0.1$), and filtered out, from each pair, the individual with the lowest genotyping rate. We performed a Principal Components Analysis (PCA) of a pruned set of common variants of our study samples projected on the 1KGP European Phase 3 samples on PLINK (version 1.9)^{80,81} (Supplementary Figs. 16 and 17). Finally, we performed a PCA with the full 1KGP Phase 3 samples and removed individual outliers that did not cluster within the European sample of the 1KGP by inspecting the first three principal components (total PCA outliers across genotyping arrays = 81). Amongst the outliers, 63 individuals are from Quebec, 8 from British Columbia, 5 from the Atlantic Provinces, 5 from Alberta and none from Ontario.

The final number of SNPs (n) and individuals (N) remaining after quality control for each chip array were: (i) Axiom 2.0 UK Biobank ($n = 630,508$; $N = 4,745$), (ii) GSA 24v1 ($n = 558,183$; $N = 438$), (iii) GSA 24v2 + MDP ($n = 596,061$; $N = 2,553$), (iv) GSA 24v1 + MDP ($n = 596,061$; $N = 4,480$), and (v) Omni 2.5 ($n = 2,081,743$; $N = 525$). (Supplementary Table 6), yielding a total number of 12,741 individuals from different provinces in Canada.

Imputation of Genotypes. Each genotyping array was first phased with EAGLE2 (version 2.0.5)⁸² using the Sanger Imputation Server⁸³. After phasing, samples on each genotyping array were imputed on the Sanger Imputation Server using the positional Burrows-Wheeler transform (PBWT) algorithm⁸⁴ and the Haplotype Reference Consortium (HRC) release 1.1 dataset as reference⁸³. The HRC includes $\sim 64,000$ haplotypes and $\sim 40,000,000$ autosomal SNPs of $\sim 32,000$ individuals predominantly of European ancestry, which makes it ideal for the imputation of our datasets, which are of European-related ancestry. Post-imputation quality control consisted of filtering out variants with INFO score <0.3 . Briefly, the INFO score is a measure of the imputation certainty across samples, in which INFO = 1 indicates complete certainty. We also filtered out variants with MAF <0.01 , missing genotyping rate $>5\%$ or variants that significantly deviated from the HWE. The final number of markers included in the GWAS for each genotyping array were: (i) Axiom 2.0 UK Biobank = 6,880,138, (ii) GSA 24v1 = 6,185,935, (iii) GSA 24v2 + MDP = 6,214,597, (iv) GSA 24v1 + MDP = 6,204,261, and (v) Omni 2.5 = 7,391,256.

Phenotyping. Participants of the CanPath answered a questionnaire that included self-report on natural hair colour (before greying) using the following discrete categories: black, dark brown, light brown, blonde, red hair colour or NA. These categories were then transformed into binary categories using R (version 3.5.1)⁸⁵ to build logistic mixed models to compare the presence (1) or absence (0) of: 1) blonde vs. brown (light and dark) + black hair colour; 2) brown (light and dark) vs. black hair colour; and 3) red vs. brown (light and dark) + black hair colour, similar to the approach used by Morgan and colleagues³³. Supplementary Table 1 shows the number of individuals on each hair colour category by genotyping array, and Supplementary Fig. 15 shows an overview of the methodology. In addition, participants also reported their age and sex.

Genome-wide association studies (GWAS) and meta-analyses. Genome-wide association studies of hair colour were performed for each genotyping array with binary logistic mixed models on SAIGE (version 0.38)⁴¹, using an additive genetic model (i.e. the effect size is a linear function of the number of effect alleles), and considering the genotypes' dosages. Specifically, the three hair colour models used were: 1) blonde vs. brown (light and dark) + black hair colour; 2) brown (light and dark) vs. black hair colour; and 3) red vs. brown (light and dark) + black hair colour. We performed a PCA of a pruned set of genotyped variants for each genotyping array after quality control, keeping only SNPs with MAF >0.05 and excluding regions of high LD, using PLINK (version 1.9)^{80,81}. We included 10 PCs as fixed effects in the logistic mixed models for all genotyping arrays. We also included in the model sex and age as fixed effects. Additionally, we included as random effects a genetic relationship matrix of independent markers to account for subtle structure, computed on PLINK (version 1.9)^{80,81}. We did not perform a

GWAS for the presence/absence of red hair colour in the two small samples (Omni 2.5 and GSA 24v1) due to the low number of cases ($N < 20$). To evaluate the case of residual population substructure, we computed the inflation factor (λ) and visualised the expected vs. observed p values using Q-Q plots on R (version 3.5.1)⁸⁵.

We performed a meta-analysis for each hair colour model (blonde, brown and red hair colour) using the beta coefficient (i.e. $\log(\text{odds ratio})$) and standard error (SE) of each study on the software METASOFT (version 2.0.1)⁴². METASOFT conducts a meta-analysis using a fixed-effects model (FE), which works well when there is no evidence of heterogeneity (i.e., assumes the same effect size across studies), and an optimised random-effects model (RE2), which works well when there is evidence of heterogeneity among studies⁴². Additionally, METASOFT computes two estimates of statistical heterogeneity, Cochran's Q statistic and I^2 statistic⁸⁶. We considered a SNP as heterogeneous across studies at an alpha level of 0.05 and $K-1$ degrees of freedom, where K is the number of studies included in the meta-analysis. Similarly, values of $I^2 > 50\%$ are considered to represent notable heterogeneity⁸⁷.

After conducting the meta-analyses results, we generated Manhattan and Q-Q plots using the qqman⁸⁸ and ggplot2⁸⁹ R packages. We used a genome-wide significant threshold of $1.67e-08$ (i.e., $5e-08/3$) to account for the three models tested. We focused our results on the fixed-effects model, but we also report the RE2 on the summary statistics of the top signals as Supplementary Data 2-4, and compared the statistical significance between both the models when there was evidence of heterogeneity based on Cochran's Q and I^2 statistics. We performed LD Score regression on LDSC⁹⁰ (version 1.0.1) on each of the meta-analyses summary statistics to evaluate possible inflation, using the LD Scores from the European population of the 1000 Genomes Project.

Additionally, we conducted GWAS and meta-analyses of hair colour using a linear mixed model approach, similar to the methods used in previous hair pigmentation studies^{5,32}. We coded the hair colour categories as a quantitative trait, spanning from low to high eumelanin (1 = blonde, 2 = light brown, 3 = dark brown, 4 = black). For this analysis, we excluded red hair colour, given that red hair colour is characterised by its high levels of pheomelanin, and does not fall within the low to high eumelanin spectrum⁹¹. We performed the GWAS on GCTA 1.26.0^{92,93}, including as fixed-effects sex, age and the first 10 PCs and a genetic relationship matrix as random effects. Subsequently, we performed a meta-analysis on METASOFT (version 2.0.1)⁴². We generated Manhattan plots for the meta-analysis results.

Annotation of significant loci. We used the web-based programme SNPnexus^{50,51} to annotate the genome-wide significant signals (p -value $\leq 1.67e-08$) from each meta-analysis. Specifically, gene and variant type annotation were done using the University of California Santa Cruz (UCSC) and Ensembl databases (human genome version hg19); assessment of the predictive effect of nonsynonymous coding variants on protein function was done with SIFT and PolyPhen scores. Both SIFT and PolyPhen output qualitative prediction scores (i.e. probably damaging/deleterious, possibly damaging/deleterious-low confidence, tolerated/benign). Noncoding variation scoring was assessed using CADD score, which is based on ranking a variant relative to all possible substitutions of the human genome. In addition, we explored the effect of significant loci on RNA and protein expression using the GTEx database⁹⁴ and the effect of significant genes using the Protein Atlas⁹⁵.

Approximate conditional analyses of association. In order to identify if the genome-wide significant loci of our original logistic meta-analyses were driven by one or more independent variants, we conducted approximate conditional and joint analyses of association (COJO) with the programme Genome-Wide Complex Trait Analysis (GCTA)⁴⁶. We performed the analysis ($-\text{cojo-slc1}$) using as input the summary statistics of our hair colour meta-analyses (fixed effects, FE) and the weighted average effect allele frequency from all studies. In addition, the programme requires a reference sample for computing LD correlations and, in the case of a meta-analysis, it is suggested to use one of the study's large samples⁴⁶. Therefore, we ran the analysis twice: 1) using as a reference the sample genotyped with the Axiom UKBB array ($N = 4745$), and 2) using as a reference the sample genotyped with the GSA 24v1 + MDP ($N = 4480$), including in both cases only high imputation-score SNPs (INFO > 0.8). We assumed that variants farther than 10 Mb are in complete linkage equilibrium, hence we ran the programme on each chromosome separately to speed up the analyses and used the genome-wide significant p -value threshold of $1.67e-08$.

Statistical fine-mapping of significant loci. We used the programme FINEMAP (version 1.4)⁴⁹ to identify candidate causal variants in the genome-wide associated loci across the genome for each binary phenotype. FINEMAP is based on a Bayesian framework, which uses summary statistics and LD correlations among variants to compute the posterior probabilities of causal variants, with a shotgun stochastic search algorithm⁴⁹. Compared to other methods, FINEMAP allows a maximum of 20 causal variants per locus. To run the programme, we used as input the meta-analyses summary statistics of each binary phenotype, including the weighted average MAF among all studies, and an LD correlation matrix from one of the large samples in our study (Axiom UKBB array, $N = 4,745$). The LD correlation matrix was computed using LDStore (version 2.0), which considers

genotype probabilities⁹⁶. We defined regions for fine-mapping as ± 500 kb regions flanking the lead SNP, based on the genome-wide signals of association from the meta-analyses, and allowing a maximum number of 10 causal signals for each locus (i.e., a maximum of 10 credible sets). A credible set is comprised of SNPs that cumulatively reach a probability of at least 95%. The SNPs within a credible set are referred to as candidate causal variants and each of them has a corresponding posterior inclusion probability (PIP).

We filtered FINEMAP results by removing candidate causal variants with a $\log_{10}\text{BF} < 2$ from each of the 95% credible sets, where a $\log_{10}\text{BF}$ indicates considerable evidence of causality. We annotated the remaining SNPs using SNPnexus⁵¹ to obtain information about the overlapping/nearest genes, overlapping regulatory elements and CADD scores. Annotation of gene expression on ENCODE, Roadmap Epigenomics and Ensembl Regulatory Build was restricted to melanocytes, keratinocytes and fibroblasts, which are the relevant cell types involved in hair pigmentation. Based on the combined evidence of fine-mapping and posterior annotation, we defined the candidate causal variants with strong evidence of causality (based on their $\log_{10}\text{BF}$ and annotation) as the most likely candidate causal variants. We computed LD correlations among the candidate causal SNP(s) on each locus using LDStore (version 2.0) and plotted the Posterior Inclusion Probability (PIP) and $\log_{10}\text{BF}$ results on R (version 3.5.1)⁸⁵ using ggplot2⁸⁹.

Gene expression and methylation using cultured melanocyte data. We conducted colocalization analyses of our GWAS meta-analyses signals using gene expression and methylation *cis*-QTL data from primary cultures of foreskin melanocytes, isolated from the foreskin of 106 newborn males^{54,76}. *Cis*-QTLs were assessed for variants in the ± 1 Mb region of each gene or CpG^{54,76}. Foreskin melanocytes are recurrently the most adequate choice to study regulatory mechanisms involved in hair colour due to the shared pigmentation pathways in skin and hair. We used the programme *hyprocoloc*⁵³ to obtain the posterior probability of a variant being shared between the GWAS and the expression or methylation QTLs. We tested all the significant eQTL genes or meQTL probes within ± 250 kb regions flanking the most significant GWAS SNP on each of the genome-wide regions of association (p -value $\leq 1.67 \times 10^{-8}$) from the logistic meta-analyses summary statistics (11 different loci across the three GWAS models). We used as LD reference the matrix obtained from the CanPath's Axiom UKBB Array (INFO score > 0.3), computed on PLINK (version 1.9; $-r$ square)^{80,81}. We kept and report colocalized regions that reached a posterior probability ≥ 0.8 , indicating high confidence of shared signal.

We performed three transcriptome-wide association studies (TWAS) by imputing the expression profile of the CanPath cohort using GWAS summary statistics and melanocyte RNA-seq expression data⁵⁴. Using the programme FUSION³⁷, we used as LD reference the CanPath's Axiom UKBB genotyping array computed in binary PLINK format (version 1.9; $-make-bed$)^{80,81}. As recommended by FUSION, we used the LDSC *munge_sumstats.py* script to check the GWAS summary statistics⁹⁰. Before running the script, we filtered out SNPs with MAF < 0.01 , SNPs with a genotyping missing rate > 0.01 and SNPs that failed HWE test at significance threshold of 10^{-7} using PLINK (version 1.9; $-maf 0.01$, $-geno 0.01$, $-hwe 10^{-7}$)^{80,81}. We computed expression weights from our melanocyte RNA-seq data one gene at a time. Genes that failed quality control during a heritability check (using minimum heritability p -value of 0.01) were excluded from the further analyses, yielding a total of 3998 genes. We restricted the locus to 500 kb on either side of the gene boundary. We applied a significance cut-off to the final TWAS result of 4.17×10^{-6} (i.e. $0.05/(3998 \text{ genes} \times 3 \text{ models tested})$). Finally, we performed conditional analysis on FUSION (FUSION_post.process.R script) if more than one gene in a locus was significant, to identify if these were independent signals. We also ran a follow-up permutation test with a maximum of 100,000 permutations to assess if the random distribution of QTL effect sizes could yield a significant association by chance.

Biological pathway gene set enrichment analysis. We performed pathway enrichment analysis using the GENE2FUNC application available on the web-based programme FUMA⁹⁷, to annotate relevant gene sets in a biological context. FUMA tests if genes are overrepresented in any pre-defined Gene Ontology (GO) gene set, and the significance ($p < 0.05$) is adjusted for multiple testing per biological category separately, using the Bonferroni method. We used as input the gene IDs that overlapped candidate causal variants defined by FINEMAP for all models jointly (Supplementary Table 4), all genes as a background set (19,283 protein-coding genes), and a minimum overlap of two genes per gene set.

Statistics and reproducibility. Statistical analyses were performed for the GWAS and downstream analyses as described in the corresponding Methods section, including all parameters used to allow reproducibility.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

We provide the genome-wide ($p \leq 1.67 \times 10^{-8}$) and suggestive ($p \leq 1 \times 10^{-6}$) signals identified in the meta-analyses as a Supplementary Data (Supplementary Data 2–4). Further

information and requests for data published here should be directed to CanPath, which regulates the access to the data and biological materials (<https://canpath.ca/>). Melanocyte genotype data, RNA-seq expression data, and all meQTL association results are deposited in Genotypes and Phenotypes (dbGaP) under accession dbGaP: phs001500.v1.p1. The raw data of Illumina HumanMethylation450 BeadChips from 106 primary human melanocytes have been submitted to the Gene Expression Omnibus (GEO) database with the accession numbers: GSE101771 and GSE166069, respectively.

Received: 15 February 2021; Accepted: 8 October 2021;
Published online: 04 November 2021

References

1. Pospiech, E., Draus-Barini, J., Kupiec, T., Wojas-Pelc, A. & Branicki, W. Gene-gene interactions contribute to eye colour variation in humans. *J. Hum. Genet.* 56, 447–455 (2011).
2. Sulem, P. et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* 39, 1443–1452 (2007).
3. Visconti, A. et al. Genome-wide association study in 176,678 Europeans reveals genetic loci for tanning response to sun exposure. *Nat. Commun.* 9, 1684 (2018).
4. Pospiech, E. et al. The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction. *Forensic Sci. Int. Genet.* 11, 64–72 (2014).
5. Adhikari, K. et al. A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation. *Nat. Commun.* 10, 1–16 (2019).
6. Byard, P. J. Quantitative Genetics of Human Skin Color. *Yearb. Phys. Anthropol.* 24, 123–137 (1981).
7. Quillen, E. E. The Evolution of Tanning Needs Its Day in the Sun. *Hum. Biol.* 87, 352–360 (2017).
8. Sturm, R. A. & Duffy, D. L. Human pigmentation genes under environmental selection. *Genome Biol.* 13, 248 (2012).
9. Hider, J. L. et al. Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC Evol. Biol.* 13, 150 (2013).
10. Jablonski, N. G. & Chaplin, G. The colours of humanity: the evolution of pigmentation in the human lineage. *Philos. Trans. R. Soc. B Biol. Sci.* 372, 20160349 (2017).
11. Quillen, E. E. et al. Shades of complexity: New perspectives on the evolution and genetic architecture of human skin. *Am. J. Phys. Anthropol.* 168, 4–26 (2018).
12. Martin, A. R. et al. An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. *Cell* 171, 1340–1353.e14 (2017).
13. Crawford, N. G. et al. Loci associated with skin pigmentation identified in African populations. *Science*. 8433, eaan8433 (2017).
14. Norton, H. L. et al. Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol. Biol. Evol.* 24, 710–722 (2007).
15. Donnelly, M. P. et al. A global view of the OCA2-HERC2 region and pigmentation. *Hum. Genet.* 131, 683–696 (2012).
16. Flanagan, N. et al. Pleiotropic effects of the melanocortin 1 receptor (MC1R) gene on human pigmentation. *Hum. Mol. Genet.* 9, 2531–2538 (2000).
17. Makova, K. & Norton, H. Worldwide polymorphism at the MC1R locus and normal pigmentation variation in humans. *Peptides* 26, 1901–1908 (2005).
18. Beleza, S. et al. Genetic Architecture of Skin and Eye Color in an African-European Admixed Population. *PLoS Genet.* 9, e1003372 (2013).
19. Lona-Durazo, F. et al. Meta-analysis of GWA studies provides new insights on the genetic architecture of skin pigmentation in recently admixed populations. *BMC Genet.* 20, 1–16 (2019).
20. Amos, C. I. et al. Genome-wide association study identifies novel loci predisposing to cutaneous melanoma. *Hum. Mol. Genet.* 20, 5012–5023 (2011).
21. Law, M. H. et al. Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nat. Genet.* 47, 987–995 (2015).
22. Brown, K. M. et al. Common sequence variants on 20q11.22 confer melanoma susceptibility. *Nat. Genet.* 40, 838–840 (2008).
23. Bishop, D. T. et al. Genome-wide association study identifies three loci associated with melanoma risk. *Nat. Genet.* 41, 920–925 (2009).
24. Barrett, J. H. et al. Genome-wide association study identifies three new melanoma susceptibility loci. *Nat. Genet.* 43, 1108–1114 (2011).
25. Duffy, D. L. et al. Multiple pigmentation gene polymorphisms account for a substantial proportion of risk of cutaneous malignant melanoma. *J. Invest. Dermatol.* 130, 520–528 (2010).
26. Antonopoulou, K. et al. Updated Field Synopsis and Systematic Meta-Analyses of Genetic Association Studies in Cutaneous Melanoma: The MelGene Database. *J. Invest. Dermatol.* 135, 1074–1079 (2015).

27. Scherer, D. & Kumar, R. Genetics of pigmentation in skin cancer — A review. *Mutat. Res.* 705, 141–153 (2010).
28. Matamá, T., Gomes, A. C. & Cavaco-Paulo, A. Hair Coloration by GeneRegulation: Fact or Fiction? *Trends Biotechnol.* 33, 707–711 (2015).
29. Rees, J. L. Genetics of hair and skin color. *Annu. Rev. Genet.* 37, 67–90 (2003).
30. Lin, J. Y. & Fisher, D. E. Melanocyte biology and skin pigmentation. *Nature* 445, 843–850 (2007).
31. Parra, E. J. Human Pigmentation Variation: Evolution, Genetic Basis, and Implications for Public Health. *Yearb. Phys. Anthropol.* 50, 85–105 (2007).
32. Hysi, P. G. et al. Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. *Nat. Genet.* 50, 652–656 (2018).
33. Morgan, M. D. et al. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nat. Commun.* 9, 5271 (2018).
34. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* 102, 717–730 (2018).
35. Cannon, M. E. & Mohlke, K. L. Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci. *Am. J. Hum. Genet.* 103, 637–653 (2018).
36. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* 19, 491–504 (2018).
37. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Publ. Gr.* 48, 245–252 (2016).
38. Hormozdiari, F. et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* 99, 1245–1260 (2016).
39. Mengel-From, J., Wong, T. H., Morling, N., Rees, J. L. & Jackson, I. J. Genetic determinants of hair and eye colours in the Scottish and Danish populations. *BMC Genet.* 10, 88 (2009).
40. Shekar, S. N. et al. Spectrophotometric Methods for Quantifying Pigmentation in Human Hair — Influence of MC1R Genotype and Environment. *Photochem. Photobiol.* 84, 719–726 (2008).
41. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* 50, 1335–1341 (2018).
42. Han, B. & Eskin, E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *Am. J. Hum. Genet.* 88, 586–598 (2011).
43. Valverde, P., Healy, E., Jackson, I., Rees, J. L. & Thody, A. J. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nat. Genet.* 11, 328–330 (1995).
44. Box, N. F., Wyeth, J. R., O’Gorman, L. E., Martin, N. G. & Sturm, R. A. Characterization of melanocyte stimulating hormone receptor variant alleles in twins with red hair. *Hum. Mol. Genet.* 6, 1891–1897 (1997).
45. Rees, J. L. The Melanocortin 1 Receptor (MC1R): More Than Just Red Hair. *Pigment Cell Res.* 13, 135–140 (2000).
46. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Publ. Gr.* 44, 369–375 (2012).
47. Han, J. et al. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet.* 4, e1000074 (2008).
48. Rana, B. K. et al. High polymorphism at the human melanocortin 1 receptor locus. *Genetics* 151, 1547–1557 (1999).
49. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 32, 1493–1501 (2016).
50. Ullah, A. Z. D., Lemoine, N. R. & Chelala, C. SNPexus: A web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic Acids Res.* 40, 65–70 (2012).
51. Ullah, A. Z. D. et al. SNPexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Res.* 46, 109–113 (2018).
52. Sulem, P. et al. Two newly identified genetic determinants of pigmentation in Europeans. *Nat. Genet.* 40, 835–837 (2008).
53. Foley, C. N. et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat. Commun.* 12, 764 (2021).
54. Zhang, T. et al. Cell-type specific eQTL of primary melanocytes facilitates identification of melanoma susceptibility genes. *Genome Res.* 28, 1621–1635 (2018).
55. Takeo, M. et al. EdnrB Governs Regenerative Response of Melanocyte Stem Cells by Crosstalk with Wnt Signaling. *Cell Rep.* 15, 1291–1302 (2016).
56. Praetorius, C. et al. A Polymorphism in IRF4 Affects Human Pigmentation through a Tyrosinase-Dependent MITF / TFAP2A Pathway. *Cell* 155, 1022–1033 (2013).
57. Jarrett, S. G., Horrell, E. M. W., Boulanger, M. C. & Orazio, J. A. D. Defining the Contribution of MC1R Physiological Ligands to ATR Phosphorylation at Ser435, a Predictor of DNA Repair in Melanocytes. *J. Invest. Dermatol.* 135, 3086–3095 (2015).
58. Mohiuddin et al. The role of HERC2 and RNF8 ubiquitin E3 ligases in the promotion of translesion DNA synthesis in the chicken DT40 cell line. *DNA Repair (Amst)*. 40, 67–76 (2016).
59. Gordon, R. Skin Cancer: An Overview of Epidemiology and Risk Factors. *Semin. Oncol. Nurs.* 29, 160–169 (2013).
60. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, 1005–1012 (2019).
61. Landi, M. T. et al. Genome-wide association meta-analyses combining multiple risk phenotypes provide insights into the genetic architecture of cutaneous melanoma susceptibility. *Nat. Genet.* 52, 494–504 (2020).
62. Visser, M., Kayser, M. & Palstra, R. J. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res.* 22, 446–455 (2012).
63. Guenther, C. A., Tasic, B., Luo, L., Bedell, M. A. & Kingsley, D. M. A molecular basis for classic blond hair color in Europeans. *Nat. Genet.* 46, 748–752 (2014).
64. Lamason, R. L. et al. SLC24A5, a putative cation exchanger, affects pigmentation in Zebrafish and humans. *Science (80-)*. 310, 1782–1786 (2005).
65. Graf, J., Hodgson, R. & Van Daal, A. Single Nucleotide Polymorphisms in the MATP Gene Are Associated With Normal Human Pigmentation Variation. *Hum. Mutat.* 28, 278–284 (2005).
66. Ambrosio, A. L., Boyle, J. A., Aradi, A. E., Christian, K. A. & Di, S. M. TPC2 controls pigmentation by regulating melanosome pH and size. *Proc. Natl. Acad. Sci.* 113, 1–6 (2016).
67. Bellono, N. W., Escobar, I. E. & Oancea, E. A melanosomal two-pore sodium channel regulates pigmentation. *Sci. Rep.* 6, 26570 (2016).
68. Chao, Y. et al. TPC2 polymorphisms associated with a hair pigmentation phenotype in humans result in gain of channel function by independent mechanisms. *Proc. Natl. Acad. Sci.* 114, E8595–E8602 (2017).
69. Simcoe, M. et al. Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color. *Sci. Adv.* 7, 1–12 (2021).
70. Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797 (2012).
71. Swope, V. B. & Abdelmalek, Z. A. Significance of the Melanocortin 1 and Endothelin B Receptors in Melanocyte Homeostasis and Prevention of Sun-Induced Genotoxicity. *Front. Genet.* 7, 1–11 (2016).
72. Soejima, M., Tachida, H., Ishida, T., Sano, A. & Koda, Y. Evidence for recent positive selection at the human AIM1 locus in a European population. *Mol. Biol. Evol.* 23, 179–188 (2006).
73. Liu, F. et al. Digital quantification of human eye color highlights genetic association of three new loci. *PLoS Genet.* 6, 34 (2010).
74. Eriksson, N. et al. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* 6, 1–20 (2010).
75. Zhang, M. et al. Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. *Hum. Mol. Genet.* 22, 2948–2959 (2013).
76. Zhang, T. et al. Cell-type-specific meQTLs extend melanoma GWAS annotation beyond eQTLs and inform melanocyte gene-regulatory mechanisms. *Am. J. Hum. Genet.* 108, 1631–1646 (2021).
77. Caglio, G., Triglia, E. T. & Pombo, A. PRC2 Poises Enhancer-Promoter Interactions at Anterior Neuronal. *Genes. Stem Cell* 20, 573–575 (2017).
78. Bonilla, C. et al. Investigating DNA methylation as a potential mediator between pigmentation genes, pigimentary traits and skin cancer. *Pigment Cell Melanoma Res.* 0, 1–13 (2020).
79. Rayner, W. McCarthy Group Tools (2019). Available at: <https://www.well.ox.ac.uk/~wrayner/tools/index.html#Checking>. (Accessed: 1st August 2019).
80. 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).
81. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7 (2015).
82. Loh, P. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* 48, 1443–1450 (2016).
83. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* 48, 1279–1283 (2016).
84. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* 30, 1266–1272 (2014).
85. R Core Team. *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, 2019).
86. Han, B. & Eskin, E. Interpreting Meta-Analyses of Genome-Wide Association Studies. *PLoS Genet.* 8, e1002555 (2012).
87. Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* 21, 1539–1558 (2002).
88. Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and Manhattan plots. *J. Open Source Softw.* 3, 1–2 (2018).
89. Wickham, H. et al. Welcome to the Tidyverse Tidyverse package. *J. Open Source Softw.* 4, 1–6 (2019).

90. Bulik-Sullivan, B. K. et al. LD sJ. Open Source Software regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295 (2015).
91. Ito, S. & Wakamatsu, K. Quantitative Analysis of Eumelanin and Pheomelanin in Humans, Mice, and Other Animals: a Comparative Review. *Pigment Cell Melanoma Res.* 16, 523–531 (2003).
92. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* 46, 100–106 (2014).
93. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82 (2011).
94. Consortium, Gt. Genetic effects on gene expression across human tissues. *Nature* 550, 204–213 (2017).
95. Uhlén, M. et al. Tissue-based map of the human proteome. *Science (80-)*. 347, 1260419 (2015).
96. Benner, C. et al. Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *Am. J. Hum. Genet.* 101, 539–551 (2017).
97. Watanabe, K. & Taskesen, E. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* 8, 1–10 (2017).

Acknowledgements

The data used in this research were made available by CanPath—Canadian Partnership for Tomorrow's Health (formerly CPTP), CARTaGENE, Alberta's Tomorrow Project, Ontario Health Study, BC Generations Project and Atlantic PATH. The authors would like to thank all the participants of the Canadian Partnership for Tomorrow's Health. FLD is supported by the National Council for Science and Technology (CONACYT) in Mexico. MM was supported by a Mitacs Globalink Research Award (FR37903) and by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (88887.474324/2020-00). EJP received funding from the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant). RT, KF, MAK, JC, TZ, and KMB are supported by the Intramural Research Programme of the NIH, National Cancer Institute, Division of Cancer Epidemiology and Genetics; <https://dceg.cancer.gov/>; the content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organisations imply endorsement by the U.S. Government. Computations were performed on the GPC supercomputer at the SciNet HPC Consortium, Canada and at the UTM High-Performance Computing server at Mississauga, ON, Canada. This work also utilised the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>). SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund—Research Excellence; and the University of Toronto.

Author contributions

E.J.P. and F.L.D. designed the study. F.L.D., M.M., and R.T. performed statistical analyses. F.L.D. wrote the draft of the manuscript. F.L.D., E.J.P., K.F., T.Z., M.A.K., J.C., and K.M.B. aided in the interpretation of the results and in the preparation of the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-021-02764-0>.

Correspondence and requests for materials should be addressed to Frida Lona-Durazo or Esteban J. Parra.

Peer review information *Communications Biology* thanks Michael Morgan and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editors: Melanie Bahlo and George Inglis. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

Final remarks

This thesis focused on the integration of population history, natural selection and biomedical advances to increase the genetic knowledge of underrepresented people in genetic studies. In the first chapter, we present a review of the human history on the American continent told by the genetics of current and ancient populations, from the arrival of the first humans to the process of admixture that occurred after the arrival of Europeans (Mendes et al, 2020). We saw that the increase in the availability of genetic data and bioinformatic tools allowed a great advance in studies on the history of the human species. Still, there are many to be discovered and the generation of datasets from populations that are usually neglected is an essential step for it. This effort can be especially helpful to take the advances coming from the inference of the polygenic risk score to the populations that need it most, as we discuss in detail in Chapter 3. In the following chapters, we present articles that exemplify the importance of this diversity by revealing new information on human history, natural selection and health through the analysis of large datasets of different neglected populations.

The use of datasets that take into account the genetic diversity as the Peruvian Genome Project, LDGH data, the Epigen initiative and the West Maharashtra data enable us to increase the knowledge about how different evolutionary events affect the human genome. This is shown with the genetic data from 18 Peruvian populations in chapter 2, where we reveal important aspects of the human history covered before by the gap in the genetic studies of these peoples. Such as the importance of other human empires, as the Wari Empire (600 to 1,000 years before present), for the genetic homogenization between the populations of the arid Andes. We have also shown that Native American populations, which are often treated as a single group in genetic studies, have a structure that implies differences in important pharmacogenetic variants and needs to be considered. We took one more step to fill the need for studies in non-European, not only bringing new insights into the history of those populations and the genetics of complex traits but by (which may be one of our greatest contributions) making publicly available the largest database of non-admixed South American native populations that we have today, which will allow we, and other groups around the world to develop more inclusive studies.

Another example of evolutionary events that shape our genome, and where the data from underrepresented people can fill important gaps in the human adaptation to different

environments, whether in the Americas (Andes, Amazonia) or in South Asia (India), which shows the importance of the development of pipelines to uncover this knowledge. However, as illustrated by the literature review and the lack of information available for some genes reported in chapter 2 and 3, there is also a gap of functional studies to characterize the selection signals found in genome scans, so our results represent just possible candidates of selection signatures and need to be carefully interpreted to avoid misunderstandings.

The identification of loci that evolved under the effect of natural selection is an efficient method of identifying possible functional genomic regions that have played an essential role in survival, and possibly have consequences for human health. However, other evolutionary factors generate genetic diversity that is associated with different outcomes in medical treatments and diseases development. The lack of diversity in genetic studies makes it difficult to understand the real effect of new discoveries, and how to project this knowledge to different populations. We exemplified that in our chapter 4, where we make an effort to take the advances of polygenic risk score inferences to the admitted populations, applying this calculation to Brazilian cohorts from the Epigen Initiative.

Exploring the diversity of human genetics is not only important for neglected populations, but it is also essential for increasing the capacity to make new discoveries. The five articles in attachments are examples of how the genomic mosaic of non-European populations is a rich source of information from different perspectives including medical, history and evolutionary aspects. The work presented here shows in several ways how the effort to generate data from underrepresented populations, brings important contributions to science. These results are a practical example that reinforces the importance of diversity in genetic studies, which in recent years has been declared by population geneticists in several journals, but which is still very low. The vast majority of studies are still focused on populations of European origin. This prevents new discoveries, such as new signatures of natural selection, the differences of populations specific allele frequency in the impact of a disease as the Covid19, and the evolution of a tract as the Lactase persistence.

Works Cited

- Alexander, David H., et al. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” *Genome Research*, vol. 19, no. 9, 2009, pp. 1655–1664., doi:10.1101/gr.094052.109.
- Bustamante, Carlos D., et al. “Genomics for the World.” *Nature*, vol. 475, no. 7355, 2011, pp. 163–165., doi:10.1038/475163a.
- Gurdasani, Deepti, et al. “Genomics of Disease Risk in Globally Diverse Populations.” *Nature Reviews Genetics*, vol. 20, no. 9, 2019, pp. 520–535., doi:10.1038/s41576-019-0144-0.
- Ioannidis, Alexander G., et al. “Native American Gene Flow into Polynesia Predating Easter Island Settlement.” *Nature*, vol. 583, no. 7817, 2020, pp. 572–577., doi:10.1038/s41586-020-2487-2.
- Martin, Alicia R., et al. “Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations.” *The American Journal of Human Genetics*, vol. 100, no. 4, 2017, pp. 635–649., doi:10.1016/j.ajhg.2017.03.004.
- Mendes, Marla, et al. “The History behind the Mosaic of the Americas.” *Current Opinion in Genetics & Development*, vol. 62, 2020, pp. 72–77., doi:10.1016/j.gde.2020.06.007.
- Mendes, Marla, et al. “The History behind the Mosaic of the Americas.” *Current Opinion in Genetics & Development*, vol. 62, 2020, pp. 72–77., doi:10.1016/j.gde.2020.06.007.
- Need, Anna C., and David B. Goldstein. “Next Generation Disparities in Human Genomics: Concerns and Remedies.” *Trends in Genetics*, vol. 25, no. 11, 2009, pp. 489–494., doi:10.1016/j.tig.2009.09.012.
- Palamara, Pier Francesco, et al. “High-Throughput Inference of Pairwise Coalescence Times

- Identifies Signals of Selection and Enriched Disease Heritability.” *Nature Genetics*, vol. 50, no. 9, 2018, pp. 1311–1317., doi:10.1038/s41588-018-0177-x.
- Parikh, Pulkit, et al. “Categorizing Sexism and Misogyny through Neural Approaches.” *ACM Transactions on the Web*, vol. 15, no. 4, 2021, pp. 1–31., doi:10.1145/3457189.
- Popejoy, Alice B., and Stephanie M. Fullerton. “Genomics Is Failing on Diversity.” *Nature*, vol. 538, no. 7624, 2016, pp. 161–164., doi:10.1038/538161a.
- Refoyo-Martínez, Alba, et al. “Identifying Loci under Positive Selection in Complex Population Histories.” 2018, doi:10.1101/453092.
- Roberts, Steven O., and Michael T. Rizzo. “The Psychology of American Racism.” *American Psychologist*, vol. 76, no. 3, 2021, pp. 475–487., doi:10.1037/amp0000642.
- Roullier, C., et al. “From the Cover: Cozzarelli Prize Winner: Historical Collections Reveal Patterns of Diffusion of Sweet Potato in Oceania Obscured by Modern Plant Movements and Recombination.” *Proceedings of the National Academy of Sciences*, vol. 110, no. 6, 2013, pp. 2205–2210., doi:10.1073/pnas.1211049110.
- Scliar, Marilia O., 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.” *International Journal of Obesity*, vol. 45, no. 5, 2021, pp. 1017–1029., doi:10.1038/s41366-021-00761-1.
- Teh, Bin Tean. “The Importance of Including Diverse Populations in Cancer Genomic and Epigenomic Studies.” *Nature Reviews Cancer*, vol. 19, no. 7, 2019, pp. 361–362., doi:10.1038/s41568-019-0158-0.
- Weissbrod, Omer, et al. “Leveraging Fine-Mapping and Non-European Training Data to Improve Cross-Population Polygenic Risk Scores.” 2021, doi:10.1101/2021.01.19.21249483.

Wojcik, Genevieve L., et al. “Genetic Analyses of Diverse Populations Improves Discovery for Complex Traits.” *Nature*, vol. 570, no. 7762, 2019, pp. 514–518.,
doi:10.1038/s41586-019-1310-4.