

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

Isabela Oliveira dos Anjos Alvim

Adaptation to the Andean and Amazonian environments and
the medical relevance of genetic diversity in Native South
Americans

Belo Horizonte

2021

Isabela Oliveira dos Anjos Alvim

Adaptation to the Andean and Amazonian environments and
the medical relevance of genetic diversity in Native South
Americans

Versão Final

Tese apresentada ao Departamento de
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: Giordano Bruno
Soares-Souza

Belo Horizonte

2021

043

Alvim, Isabela Oliveira dos Anjos.

Adaptation to the Andean and Amazonian environments and the medical relevance of genetic diversity in Native South Americans [manuscrito] / Isabela Oliveira dos Anjos Alvim. - 2021.

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

Orientador: Eduardo Martin Tarazona Santos. Co-orientador: Giordano Bruno Soares-Souza.

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

1. Genética Populacional. 2. Seleção natural. 3. América do sul - nativos. I. Santos, Eduardo Martin Tarazona. II. Soares-Souza, Giordano Bruno. 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	141/2021 entrada
Isabela Oliveira dos Anjos Alvim	2º/2016 CPF: 090.119.726-27

Às treze horas do dia **24 de fevereiro de 2021**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: "**Adaptation to the Andean and Amazonian environments and the medical relevance of genetic diversity in Native South Americans**", 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
Giordano Bruno Soares Souza	Bio Bureau Biotecnologia	065.989.496-37	Aprovada
Francisco Pereira Lobo	UFMG	012.273.736-94	Aprovada
Sibelle Vilaça	Universidade de Ferrara	063.790.716-79	Aprovada
Sandro L Bonatto	UFRGS	41358619034	Aprovada
Jorge Macedo Rocha	Universidade de Porto	P667084	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, 24 de fevereiro de 2021.

Eduardo Martin Tarazona Santos - Orientador (UFMG)

Giordano Bruno Soares Souza - Coorientador (Bio Bureau Biotecnologia)

Francisco Pereira Lobo (UFMG)

Sibelle Vilaça (Universidade de Ferrara)

Sandro L Bonatto (UFRGS)

Jorge Macedo Rocha (Universidade de Porto)

Assinatura dos membros da banca examinadora:



Documento assinado eletronicamente por **Francisco Pereira Lobo, Professor do Magistério Superior**, em 24/02/2021, às 16:54, 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 **Sandro Luis Bonatto, Usuário Externo**, em 24/02/2021, às 16:54, 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 **Giordano Bruno Soares Souza, Usuário Externo**, em 24/02/2021, às 16:55, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Sibelle Torres Vilaca, Usuário Externo**, em 24/02/2021, às 16:55, 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 Martín Tarazona Santos, Professor do Magistério Superior**, em 24/02/2021, às 16:55, 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 **Jorge Macedo Rocha, Usuário Externo**, em 24/02/2021, às 16:57, 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 **0581650** e o código CRC **2A22CA2D**.

Agradecimentos

Agradeço primeiramente ao meu orientador Eduardo Tarazona por todo o conhecimento que me passou durante esse ano, pela paciência e confiança, e pela grande oportunidade de trabalhar em projetos com a equipe excelente que está sob sua coordenação.

A todos da equipe do LDGH que me receberam de braços abertos, estavam dispostos a ajudar a cada momento e me mostraram a eficiência e produtividade de um grupo interdisciplinar bem estruturado. Em especial agradeço a Victor Borda, Marla Mendes, Carolina Silva, Thiago Peixoto e Giordano Soares, com quem trabalhei a maior parte desses 4 anos e foram essenciais para o desenvolvimento deste trabalho, e à Camila Zolini, que faz um trabalho excelente mantendo o LDGH nos eixos.

À minha mãe Adriana, por me dar toda a força que precisei para chegar até aqui, por formar a mulher que eu sou hoje e por não duvidar que eu conseguiria por um minuto sequer. Além disso, sempre com muita compreensão e disposta a fazer o que pudesse para ajudar, ela esteve ao meu lado em muitos momentos de estresse.

Ao meu pai Fausto que me inspirou na busca por conhecimento durante toda a vida e que deu toda ajuda que pedi durante a escrita desta tese.

Às minhas irmãs, tanto às de sangue quanto às de coração, mulheres incríveis e inspiradoras que me permitem ver o mundo de diferentes ângulos, me fazem manter a mente aberta, e me dão todo o apoio que eu poderia querer, na alegria e na tristeza.

À minha querida tia Maria, que acompanhou passo a passo e contribuiu para esse processo mais do que eu imaginava.

Ao meu noivo Igor, com quem passei os últimos 10 meses de pandemia em casa e que, com muito amor e dando todo o suporte possível, ficou ao meu lado dos piores aos melhores momentos desse processo. Não posso imaginar o que seria passar por isso sem ele e mal posso esperar pelos próximos capítulos da nossa vida juntos.

Obrigada a todos que estiveram presentes no meu caminho e me ajudaram a chegar até aqui.

Summary

Agradecimientos	2
Summary	3
List of Figures	4
List of Tables	7
List of Attachments	8
List of abbreviations	9
Resumo	10
Abstract	11
Introduction	12
Chapter 1 - Review: The history behind the mosaic of the Americas	16
Introduction	16
Methodology	17
Published article: The history behind the mosaic of the Americas	20
Chapter 2 - Natural Selection and Genetic variability in Native South Americans	26
Introduction	26
Published article: The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture	28
Complementary Discussion	95
Natural Selection Analyses	95
Genetic Differentiation Analyses	96
Perspectives	99
Genetic differentiation in Native South Americans	99
Functional characterization of natural selection candidate variants	100
Chapter 3 - Collaborations in other projects on American populations	102
Introduction	102
Artigo 1	103
Artigo 2	106
Artigo 3	108
Final remarks	110
References	113
Attachment 1.	123
Attachment 2.	124

List of Figures

Chapter 1

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

Chapter 2

Figure 1. 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 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 S6. ADMIXTURE analysis for 90 worldwide populations including 71 Native American populations, 18 Asian, 2 Oceanian, Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 230K Dataset).

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

Figure S8. fineSTRUCTURE clustering analysis for the Dataset 1.9M dataset.

Figure S9. fineSTRUCTURE clustering analysis for the Dataset 500K dataset.

Figure S10. fineSTRUCTURE clustering analysis for the Dataset 230K dataset.

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

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

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

Figure S14. Quantile-quantile plot comparing Z-scores from D-statistics relating Western (Northern Coast and Arid Andes) and Eastern Andean slope (Amazon Yunga and Amazon Yunga) populations to those expected under a normal distribution (green diagonal) for the Dataset 500K.

Figure S15. Quantile-quantile plot comparing Z-scores from D-statistics relating Western (Northern Coast and Arid Andes) and Eastern Andean slope populations (Amazon Yunga, Lower Amazon and other eastern groups) to those expected under a normal (green diagonal) distribution for the Dataset 230K.

Figure S16. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes.

Figure S17. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes.

Figure S18. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes.

Figure S19. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes.

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 dynamic 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 overlap in Andean populations.

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

Figure S28. Linkage Disequilibrium between rs3775587 and the SNPs found to be under selection in the gene HAND2-AS1.

Figure S29. UCSC Genome Browser view of HAND2-AS1 locus with the SNPs located in regions found to be under selection.

Figure S30. UCSC Genome Browser view of PTPRC locus with the SNPs located in regions found to be under selection.

Figure 4. Manhattan plot showing F_{ST} values between Andean and Amazon populations. The red line delimits the 0,01% highest values ($F_{ST}>0.318$). Only undifferentiated SNPs ($F_{SC}<0.15$) are shown.

Chapter 3

Figure 1. 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.

Figure 2. Fst values distribution of Native Americans vs East Asian populations for 71 SNPs of TMPRSS2 gene.

List of Tables

Chapter 1

Table 1. List of filters and terms used in the search performed with Pubmed and number of results obtained.

Table 2. List of papers analyzed to produce the section “Inferences about natural selection” of the review “The history behind the mosaic of the Americas”, published in the journal Current Opinion in Genetics & Development.

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.

Table 1. Frequency of SNPs in LD in gene ABCB1.

Table 2. PharmGKB annotation level 2A for SNP rs1045642 in gene ABCB1.

List of Attachments

Attachment 1. UCSC Genome Browser view of the locus in chr 17 with highly differentiated SNPs rs16951028 and rs5010295 in gene CA10.

Attachment 2. Linkage disequilibrium (r^2) between SNPs with high F_{ST} values associated with drug metabolism (>0.27) in the pharmacogene gene ABCB1 in (A) Andean and (B) Amazon populations.

List of abbreviations

- aDNA** - Ancient Deoxyribonucleic acid
- ALL** - Acute Lymphoblastic Leukemia
- CEU** - Northern Europeans from Utah
- DNA** - Deoxyribonucleic acid
- EAS** - East Asian
- eQTLs** - Expression quantitative trait loci
- EUR** - European
- GTE_x** - Genotype-Tissue Expression Portal
- GWAS** - Genome Wide Association Study
- KYA** - Kilo-Years Ago
- LD** - Linkage Disequilibrium
- LDGH** - Laboratory of Human Genetic Diversity
- MASSA** - Multi-Agent System for SNP Annotations
- mtDNA** - Mitochondrial Deoxyribonucleic acid
- nPBS** - Normalized Population Branch Statistic
- PharmGKB (PGKB)** - The Pharmacogenetics Knowledge Base
- RNAi** - RNA interference
- SARS** - Severe Acute Respiratory Syndrome
- SIFT** - Sorting Intolerant From Tolerant
- SNP** - Single Nucleotide Polymorphism
- xpEHH** - Cross Population Extended Haplotype Homozygosity
- WAFR** - Western Africa

Resumo

Após a chegada dos primeiros povos na América do Sul, vários grupos culturais foram formados levando ao surgimento de sociedades complexas em diferentes áreas ecológicas. Mais tarde, a chegada dos europeus levou a um processo de miscigenação que contribuiu para a estrutura genômica de várias populações atuais. A inclusão da variabilidade genética dessas populações nos estudos genéticos contribuiu para a melhor compreensão da história humana, bem como para resolver questões biomédicas, como suscetibilidade a doenças, resposta a tratamentos médicos e elucidação de fenótipos complexos. No entanto, as populações sul-americanas permanecem sub-representadas. Esta tese apresenta trabalhos que contribuem para mudar esse cenário. O primeiro capítulo é uma revisão da história da humanidade no continente, contada pela genética de populações atuais e antigas, desde a chegada dos primeiros seres humanos até o processo de miscigenação. No segundo capítulo, apresento o principal projeto que desenvolvi durante meu doutorado, focado na adaptação aos ambientes dos Andes e da Amazônia e na relevância médica da diversidade genética dessas populações. Ao aplicar testes de varredura genômica para seleção natural em um grande conjunto de dados de populações nativas do Peru (177 indivíduos genotipados para 2,5 M SNPs), identificamos: nos Andes, os genes HAND2-AS1 (relacionados à função cardiovascular) e DUOX2 (relacionada à função tireoidiana e imunidade inata); na Amazônia, o gene que codifica a proteína CD45, essencial para o reconhecimento de antígenos pelos linfócitos T/B na interação vírus-hospedeiro. Através da análise de diferenciação genética (Estatísticas F) entre populações nesses dois ambientes, identificamos diferenças acentuadas na frequência de variantes de relevância biomédica relatadas no GWAS Catalog (TMPRSS6) e PharmGKB (ABCG2). No terceiro capítulo, apresento três artigos nos quais participei analisando populações americanas nativas e miscigenadas. Os dois primeiros exploram o mosaico dos genomas das populações americanas miscigenadas para fazer inferências sobre a história da diáspora africana no continente americano, e para detectar variantes associadas ao IMC realizando um *admixture mapping* em coortes populacionais brasileiras. Por fim, apresento um artigo desenvolvido no contexto da atual pandemia de COVID-19 que avalia a diversidade genética de genes relacionados à SARS em nativos da América do Sul.

Abstract

After the arrival of the first human beings in South America, several cultural groups were formed leading to the emergence of complex societies in different ecological areas. Later, the arrival of Europeans led to a process of admixture that contributed to the genomic structure of several current populations. The inclusion of the genetic variability of these populations in genetic studies contributes to the better understanding of human history, as well as to solve biomedical issues such as susceptibility to diseases, response to medical treatments and the elucidation of complex traits. However, South American populations remain underrepresented. This thesis presents works that contribute to changing this scenario. The first chapter is a review on 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. In the second chapter I present the main project I developed during my PhD focused on the adaptation to the environments of the Andes and the Amazon, and the medical relevance of the genetic diversity of these populations. By applying genome-wide scans for natural selection on a large dataset of native populations from Peru (177 individuals genotyped for 2.5 M SNPs), we find: in the Andes, the genes *HAND2-AS1* (related to cardiovascular function) and *DUOX2* (related to thyroid function and innate immunity); in the Amazon, the gene that encodes the CD45 protein, essential for the recognition of antigens by T/B lymphocytes in the virus-host interaction. Through the analysis of genetic differentiation (F-Statistics) between populations in these two environments, we identified sharp differences in the frequency of variants of biomedical relevance reported in the GWAS Catalog (*TMPRSS6*) and PharmGKB (*ABCG2*). In the third chapter, I present three articles in which I participated analysing Native and Admixed American populations. The first two explore the mosaic of the genomes of admixed American populations to make inferences on the history of the African diaspora to the American continent, and to detect variants associated with BMI through an admixture mapping performed on Brazilian population cohorts. Finally, I present an article developed in the context of the current pandemic of COVID-19 that evaluates the genetic diversity of genes related to SARS in Native South Americans.

Introduction

Along the journey of the human species across the continents colonizing the most diverse environments, several evolutionary events have shaped the genetic structure of populations, leaving footprints on the human genome. In the first chapter of this thesis I present a review on the history of the settlement of the Americas told by the genomic footprints that have been detected in the last years with a range of different methods applied to current populations and ancient DNA. This review, entitled “The history behind the mosaic of the Americas”¹, was published in the journal *Current Opinion in Genetics & Development* and provides an evolutionary contextualization of the population dynamic and the events of genetic differentiation and adaptation of the South American populations that will be treated in the second chapter.

In the second chapter, I present the article “The genetic structure and adaptation of Andean highlanders and Amazonians is influenced by the interplay between geography and culture”², in which our group analyzes the genome of Native South American populations to elucidate the genetic relationships between them, the history of gene flow, and the process of adaptation to the Andean and Amazonian environments. The settlement of South America represents the occupancy of different ecological environments and these populations are exposed to specific selective pressures of climates, diets and pathogen diversity³. With increasing available genetic information and the development of new methods, such as genome wide scans for natural selection, several signs of recent selection have been elucidated in different world populations⁴. Such signs regard human genes related to metabolism, physical traits and disease resistance, for example: skin pigmentation⁵, lactose intolerance⁶, high altitude adaptations⁷ and resistance to malaria⁸. However, studies with native populations in South America are still scarce, whether in analyses of natural selection or in genetic studies in general^{4,9}. In the work presented here, I address this issue by carrying out tests of natural selection, and by identifying variants of biomedical relevance that present discrepant frequencies among the populations that inhabit the Andean and Amazonian environments.

The Andean populations addressed in this project live in the arid south region of the Andes, where altitudes are higher and the environment is characterized by the cold, dryness, high UV radiation, and lower oxygen levels (hypoxia). These factors make the Andes an

extreme environment, ideal for the study of human adaptation. In fact, the Andean populations have been the focus of many anthropological and physiological studies in South America ¹⁰. The pressure exerted by hypoxia does not leave much room for cultural adaptations, requiring biological changes ¹¹, and several hematological and respiratory differences between Andeans and lowlanders have already been identified in these groups. For example: lower values of arterial-alveolar O₂, greater total lung volumes, higher levels of hemoglobin, lower leg blood flow during exercise, and greater pelvic blood flow during pregnancy ¹⁰. Some of these adaptations are developmental, such as lowlanders who migrated to high-altitude environments as children and have forced vital capacity (FEV) similar to that of high-altitude natives, which is greater than that of those who migrated as adults ¹². Others result from acclimatization, as occurs when hemoglobin levels rise in lowlanders who have spent time in high altitude environments ¹³. However, some of these adaptations can not be acquired, and require a genetic explanation ¹⁰.

Genome-wide scans for natural selection in Andean populations have identified genes related to the hypoxia-inducible factors pathway (EGLN1, ET-1) ^{14,15}, oxidative stress (FAM213A) ¹⁶, thyroid function (DUOX2) ¹⁷ and cardiovascular function and development (NOS2, VEGFB, TBX5) ^{18,19}. Both genetic and physiological studies show differences in the adaptations found in Andean and Tibetan high altitude populations ¹⁰. Until now, methods focused on monogenic selection have identified only one convergent region between these populations, which is around EPAS1 gene (endothelial protein from the PAS 1 domain - 2p21) that encodes the hypoxia-inducible factor of the 2-alpha protein (HIF-2a) ¹⁴. Using a polygenic approach that aims to detect more subtle signals in multiple correlated genes, Gouy et al., 2017 ²⁰ identified convergent selection in Andean and Tibetan populations in pathways related to vascular process, hypoxia response and blood coagulation.

In contrast to the long history of studies on adaptation in Andean populations, information about natural selection in the Amazon is much more scarce. The low incidence of light, a warm and humid climate, and the high biodiversity, including human pathogens, found in the Amazon, are typical characteristics of the rainforests found around the world ^{3,9}. Amorin et al. 2015 ²¹ explored these environmental similarities by looking for signs of natural selection in rainforests populations in Brazil and Africa and found signals of selection in a gene related to lipid metabolism (SCP2 - sterol transport protein 2) in America and a convergent signal between the continents related to the immune system (CCL28 - CC Motif Chemokine ligand 28). Other studies on adaptation to rainforest mainly targeted African

hunters-gatherers populations and had found genetic regions under selection related to: immune system, reproduction, lipid metabolism, thyroid function, and body growth^{9,22}. The last one is related to the short stature phenotype that is typically associated with populations living in tropical forests and was evidenced by the signs of convergent polygenic selection in Asian and African rainforest populations^{23,24}. Hypotheses about the possible advantages of this phenotype address different aspects of the environment: energy availability (despite the high biodiversity, rainforests are not rich in resources for humans²⁵), improvement in locomotion in dense vegetation, and thermoregulation. The work presented here contributes to fill the gap regarding genomic studies in the Amazon exploring the genetic factors that influenced the human survivor in this environment.

The effect of natural selection in these two different environments has led to the prevalence or exclusion of different alleles in these populations. However, several variants present discrepancies in the frequencies between these groups (showing high values of F_{ST} for those variants) that did not necessarily arise due to the action of natural selection but may have medical relevance, influencing disease susceptibility and response to medical treatments. Population genetics studies are more frequently conducted in populations of European descent, leaving other ethnic groups, such as Native South Americans, underrepresented^{26,27}. This unbalance in representation creates a bias when applying new discoveries for these different populations²⁷. In this context, the identification of variants previously associated with health-related phenotypes that are highly differentiated between these populations, can contribute to improve health policy management in these regions. In addition to identifying genetic regions adaptive to the environments of the Andes and the Amazon, I did a survey of the highly differentiated variants among these populations that are deleterious mutations, or are mentioned in the PharmGKB (<https://www.pharmgkb.org/>)²⁸ and GWAS Catalog databases (<https://www.ebi.ac.uk/gwas/home>)^{28,29}.

In addition to developing the main project on the evolutionary history and adaptation of Native South American populations, during my PhD 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 in South American populations. In chapter 3 I describe my contribution in analyses carried out for 3 articles: “Origins, Admixture Dynamics, and Homogenization of the African Gene Pool in the Americas”³⁰, “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”³¹, and “Human-SARS-CoV-2 interactome and human genetic diversity: TMPRSS2-rs2070788, associated with severe influenza-induced SARS, and its population genetics caveats in Native Americans”³².

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

Introduction

One of the main research topics of the Laboratory of Human Genetic Diversity (LDGH) group is the human evolutionary history of Latin American populations. We aim to increase the knowledge about how the settlement of this continent occurred since the arrival of the first human beings, the dynamic between the complex populations and empires that were raised there, the process of adaptation to new and diverse environments, and the admixture process after the arrival of Europeans. In this context, we were invited by Prof. Sarah Tishkoff from the University of Pennsylvania to write a review of the recent literature on the human evolutionary history in the Americas, that we entitled “The history behind the mosaic of the Americas”¹, published in the journal *Current Opinion in Genetics & Development*. My contribution to this paper was writing the section “Inferences about natural selection” and reading and critically discussing the other sections.

I was in charge of this Section because during my PhD I worked on the analysis of the genetic differentiation and the inference of natural selection in Native South American populations from two very distinct environments: the Andes and the Amazon. This work required a detailed study of the demographic history of these populations and of all aspects of the natural selection process, including the methodologies applied to detect its footprints in the genome, which gave me a good background to collaborate on this review. This article provides an evolutionary contextualization of the events of genetic differentiation and adaptation of the South American population that will be treated in the next chapter. For this reason, although this review was written in the last year of my PhD, I decided to put it in Chapter 1 as an introduction to Chapter 2.

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. Here, we highlight that the most used methods to infer natural selection are based on allelic frequencies and their spectra, or long-range haplotypes/linkage disequilibrium, with few studies using a gene-set approach

to detect more subtle signals from genes in common biological networks. This inspired us to start an exploratory analysis on polygenic selection, conducted by Carolina Silva, a master’s student, and to write a project aiming to develop a method to detect polygenic selection and to apply it to native populations from the Amazon and the Andes. This proposal was submitted by me to the Brazilian National Council for Scientific and Technological Development - CNPq to apply for a Junior Postdoctoral Scholarship (PDJ).

Methodology

The survey for the section “Inferences about natural selection” was carried out through a search in Pubmed for articles published between 2015 and February 2020, using combinations of the following terms: “natural selection”, “human adaptation/evolution”, “North America”, “South America”, “Americas”, “Native American populations” and “admixed populations” (Table 1). After an initial screening, 38 articles (Table 2) related to the theme, including three reviews, were selected to be analyzed in depth. The content of these papers was condensed in a text highlighting the main findings, methodologies, challenges and perspectives in the field.

Table 1. List of filters and terms used in the search performed with Pubmed and number of results obtained.

Filters	Terms	Results	
Year: 2015-2020 Specie: Humans	Natural Selection	Americas	820
		Native American populations	34
		South America	152
		North America	386
		Admixed populations America	15
	Human adaptation evolution	Americas	415
		Native American populations	9
		South America	104
		North America	165
		Admixed populations America	3

*These numbers refer to the results of the respective searches including papers that appear in the other combinations of terms.

Table 2. List of papers analyzed to produce the section “Inferences about natural selection” of the review “The history behind the mosaic of the Americas”, published in the journal *Current Opinion in Genetics & Development*. Reviews are marked with *.

Article	Author	Year
Detection of Convergent Genome-Wide Signals of Adaptation to Tropical Forests in Humans.	Amorin et al. ²¹	2015
Ancestry variation and footprints of natural selection along the genome in Latin American populations.	Deng et al. ³³	2016
Ancient DNA reveals selection acting on genes associated with hypoxia response in pre-Columbian Peruvian Highlanders in the last 8500 years.	Fehren-Schmitz e Georges ³⁴	2016
The role of natural selection in human evolution – insights from Latin America.*	Salzano et al. ³⁵	2016
Genetic signature of natural selection in first Americans.	Amorim et al. ³⁶	2017
Human adaptation to arsenic in Andean populations of the Atacama Desert.	Apata et al. ^{36,37}	2017
Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans.	Crawford et al. ¹⁸	2017
Evidence of Early-Stage Selection on EPAS1 and GPR126 Genes in Andean High Altitude Populations.	Eichstaedt et al. ³⁸	2017
Detecting gene subnetworks under selection in biological pathways.	Gouy et al. ²⁰	2017
Strong Amerindian Mitonuclear Discordance in Puerto Rican Genomes Suggests Amerindian Mitochondrial Benefit.	Massey ³⁹	2017
Measuring high-altitude adaptation.*	Moore ⁴⁰	2017
Genome-Wide Analysis in Brazilians Reveals Highly Differentiated Native American Genome Regions.	Mychaleckyj et al. ⁴¹	2017
Ancestral Variations in the Shape and Size of the Zygoma.	Oettle et al. ⁴²	2017
An assessment of postcranial indices, ratios, and body mass versus eco-geographical variables of prehistoric Jomon, Yayoi agriculturalists, and Kumejima Islanders of Japan.	Seguchi et al. ⁴³	2017
New Insights into the Genetic Basis of Monge’s Disease and Adaptation to High-Altitude.	Stobdan et al. ⁴⁴	2017
Variation in obstetric dimensions of the human bony pelvis in relation to age-at-death and latitude.	Auerbach et al. ⁴⁵	2018
Reconstructing the Deep Population History of Central and South America	Posth et al. ⁴⁶	2018
Genetic ancestry effects on the distribution of toll-like receptors (TLRs) gene polymorphisms in a population of the Atlantic Forest, São Paulo, Brazil.	Guimaraes et al. ⁴⁷	2018
Environmental selection during the last ice age on the mother-to-infant transmission of vitamin D and fatty acids through breast milk.	Hlusco et al. ⁴⁸	2018
Analysis of Type 2 Diabetes and Obesity Genetic Variants in Mexican Pima Indians Marked Allelic Differentiation Among Amerindians at	Hsueh et al. ⁴⁹	2018

HLA.		
Selection scan reveals three new loci related to high altitude adaptation in Native Andeans.	Jacovas et al. ¹⁷	2018
Susceptibility to Plasmodium vivax malaria associated with DARC (Duffy antigen) polymorphisms is influenced by the time of exposure to malaria.	Kano et al. ⁵⁰	2018
Natural Selection Has Differentiated the Progesterone Receptor among Human Populations.	Li et al. ⁵¹	2018
FADS1 and the Timing of Human Adaptation to Agriculture.	Mathieson and Mathieson ⁵²	2018
Genetic ancestry, admixture and health determinants in Latin America.	Norris et al. ⁵³	2018
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.	Oliveira et al. ⁵⁴	2018
Detecting Polygenic Adaptation in Admixture Graphs.	Racimo et al. ⁵⁵	2018
polymorphisms of ADME-related genes and their implications for drug safety and efficacy in Amazonian Amerindians.	Rodrigues et al. ⁵⁶	2018
A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia.	Adhikari et al. ⁵⁷	2019
Population history and gene divergence in Native Mexicans inferred from 76 human exomes.	Avila-Arcos et al. ⁵⁸	2019
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.	Calonga-Solis et al. ⁵⁹	2019
The Genetic Architecture of Chronic Mountain Sickness in Peru.	Grazal et al. ⁶⁰	2019
Evolution of Hominin Polyunsaturated Fatty Acid Metabolism: From Africa to the New World.	Harris et al. ⁶¹	2019
Human Genetic Adaptation to High Altitude: Evidence from the Andes.*	Julian and Moore ¹⁰	2019
Comparing signals of natural selection between three Indigenous North American populations.	Reynolds et al. ⁶²	2019
Complex nature of Hominin Dispersals: ecogeographical and climatic evidence for pre-contact craniofacial Variation.	Ross and Ubelaker ⁶³	2019
Adaptation to Extreme Environments in an Admixed Human Population from the Atacama Desert.	Vicuna et al. ⁶⁴	2019
Investigating mitonuclear interactions in human admixed populations.	Zaide and Makova ^{64,65}	2019



ELSEVIER



The history behind the mosaic of the Americas

Marla 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

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

² Laboratório 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.

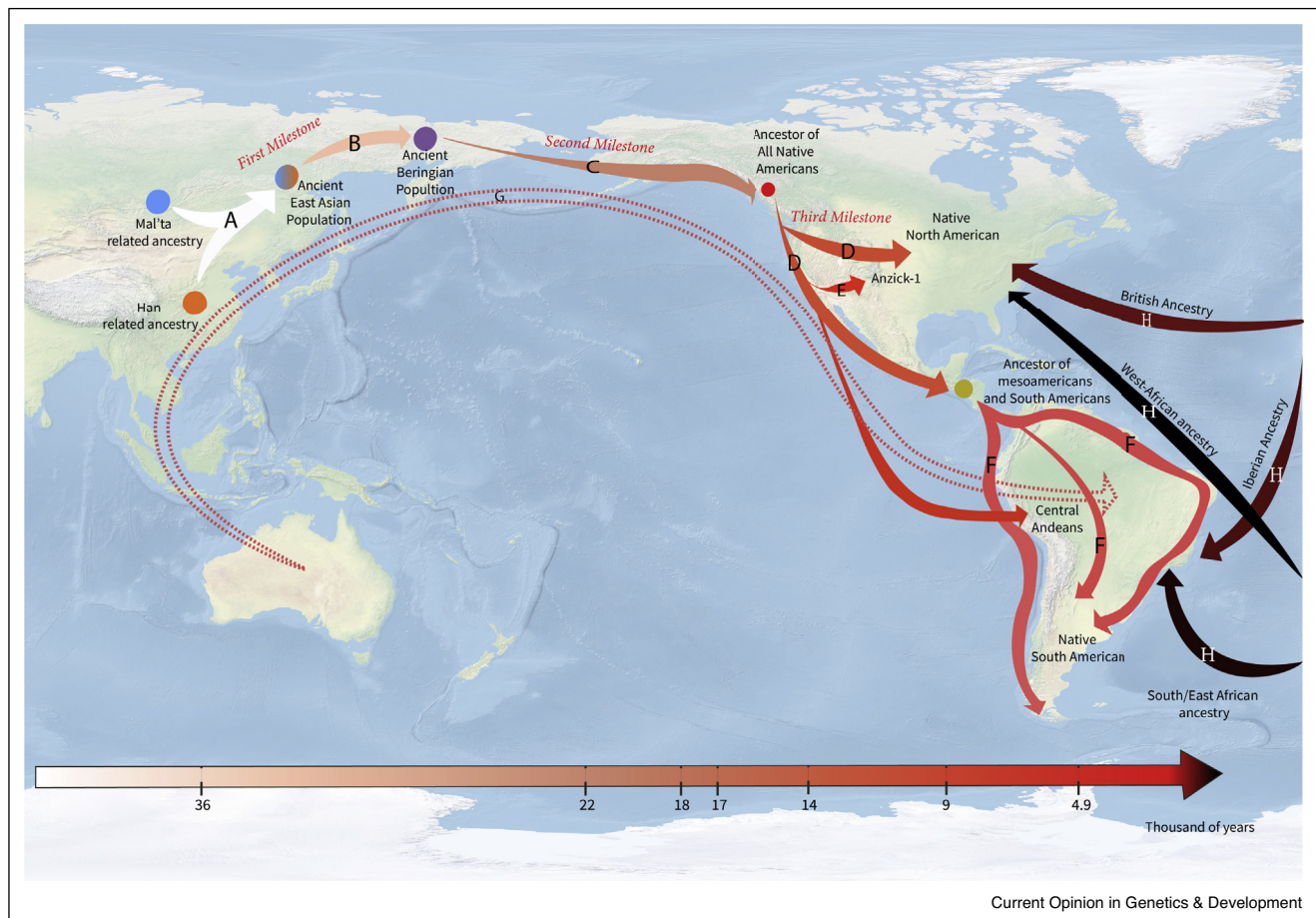
Western South America, being one of the worldwide cradles of civilization [28], has been a focus of population genetics studies [24,29[•],30,31,32^{••}]. The well known Inca Empire was the *tip of the iceberg* of a process that lasted for thousands of years. Notably, Western South America hosts a rich 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 remanent 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 exists. 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, Hanaísa 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 Mitaes 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. González-José 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 Alaskan genome 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, Korneliussen 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, Schifels 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, Mörseburg 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, Martínón-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, Nägele K, 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.* [50,19], 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, Cáceres 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. Gneccchi-Ruscione GA, Sarno S, De Fanti S, Gianvincenzo L, Giuliani C, Boattini A, Bortolini E, Di Corcia T, Sanchez Mellado C, Dávila Francia TJ *et al.*: **Dissecting the pre-Columbian genomic 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 À, Guerra Amorim CE, Krieger JE, Mill JG, Salzano FM, Bortolini MC, Pereira A 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. Chacón-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuña-Alonzo V, Barquera R, Quinto-Sánchez M, Gómez-Valdés J, Everardo Martínez P, Villamil-Ramírez 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-Padrón 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, Gnechi-Ruscione GA, Alarcón-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, Fernández-López JC, Zakharia F, Sikora M, Contreras AV, Acuña-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, Mariño-Ramírez 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, Simões 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.
 49. Calonga-Solís V, Malheiros D, Beltrame MH, De Brito Vargas L, Dourado RM, Issler HC, Wasseem 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. Ávila-Arcos MC, McManus KF, Sandoval K, Rodríguez-Rodríguez JE, Villa-Islas V, Martín AR, Luisi P, Peñaloza-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>.
 51. Reynolds AW, Mata-Míguez J, Miró-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. Vicuña 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, Hünemeier T, Tarazona-Santos E *et al.*: **Reconstructed lost Native American populations from Eastern Brazil are shaped by differential Jê/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, Hünemeier 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.

Chapter 2 - Natural Selection and Genetic variability in Native South Americans

Introduction

The main project developed during my PhD aimed to analyze the genetic diversity of native populations in South America with two focuses: (i) the identification of 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 (ii) the analysis of the differentiation between the populations of these regions, with the aim of identifying variants that are highly divergent and that may be of biomedical interest, according to notes from public databases. These analyses are reported in the article “The genetic structure and adaptation of Andean highlanders and Amazonians is influenced by the interplay between geography and culture”², of which I share the first authorship. In this article, our group analyzes the evolutionary history of Native South American populations told by the footprints left in the genome of current populations through the processes of migration, gene flow, population isolation, genetic drift and natural selection. My contribution was reading and critically discussing all the sections, and performing the analyses listed below and described in detail in the Supplementary Material of the article (Section 5).

I developed scripts in Perl and R to prepare input files, run the analyses, filter and plot the results of the following tests:

- ▶ Population Branch Statistics (PBS)^{7,18}: Genome-wide scan for natural selection based on allele frequencies. This analysis was carried out in conjunction with the PhD student Marla Mendes. The parameters needed to run the test were decided together, the scripts to run the PBS and p-value values were made by Marla and those to filter and plot the results by me. The annotation and analysis of the results was performed by me.
- ▶ Cross-population Extended Haplotype Homozygosity (xpEHH)⁶⁶: genome-wide scan for natural selection based on haplotypes.
- ▶ F-statistics⁶⁷: Calculation of the F_{ST} to identify variants with highly differentiated frequencies between the study populations.













The following analyzes were performed by me using public softwares developed for population genetics and genomics analysis:

- ▶ Linkage disequilibrium (LD) analysis with Haploview ⁶⁸.
- ▶ Annotation of the natural selection and genetic differentiation results in public databases with the software MASSA ⁶⁹.
- ▶ Identification of regulatory elements: we used the UCSC genome browser ⁷⁰ to check for the presence of active regulatory elements around the variants with natural selection signals. This analysis was carried out in conjunction with Dr. Marcelo Luizon, from the Pharmacogenomics Lab of the Department of Genetics, Ecology and Evolution at UFMG.

Supplementary tables, as well as the paper and supplementary material in PDF format can be accessed at:

<https://drive.google.com/drive/folders/1Ew2qi6FFIKz3WmU1akdgUvkVtgsrFsgY?usp=sharing>

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

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

^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; ^cLaboratório de Biotecnologia 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 Estudios Avanzados Transdisciplinarios, 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-AS1* (heart and neural crest derivatives expressed 2 antisense RNA1, a noncoding gene related to cardiovascular function) and *rs269868-C/Ser1067* in *DUOX2* (dual oxidase 2, related to thyroid function and innate immunity) genes and, in the Amazon, the gene encoding for the CD45 protein, essential for antigen recognition by T and B lymphocytes in viral–host interaction.

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

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.

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

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Published under the [PNAS license](#).

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>.

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

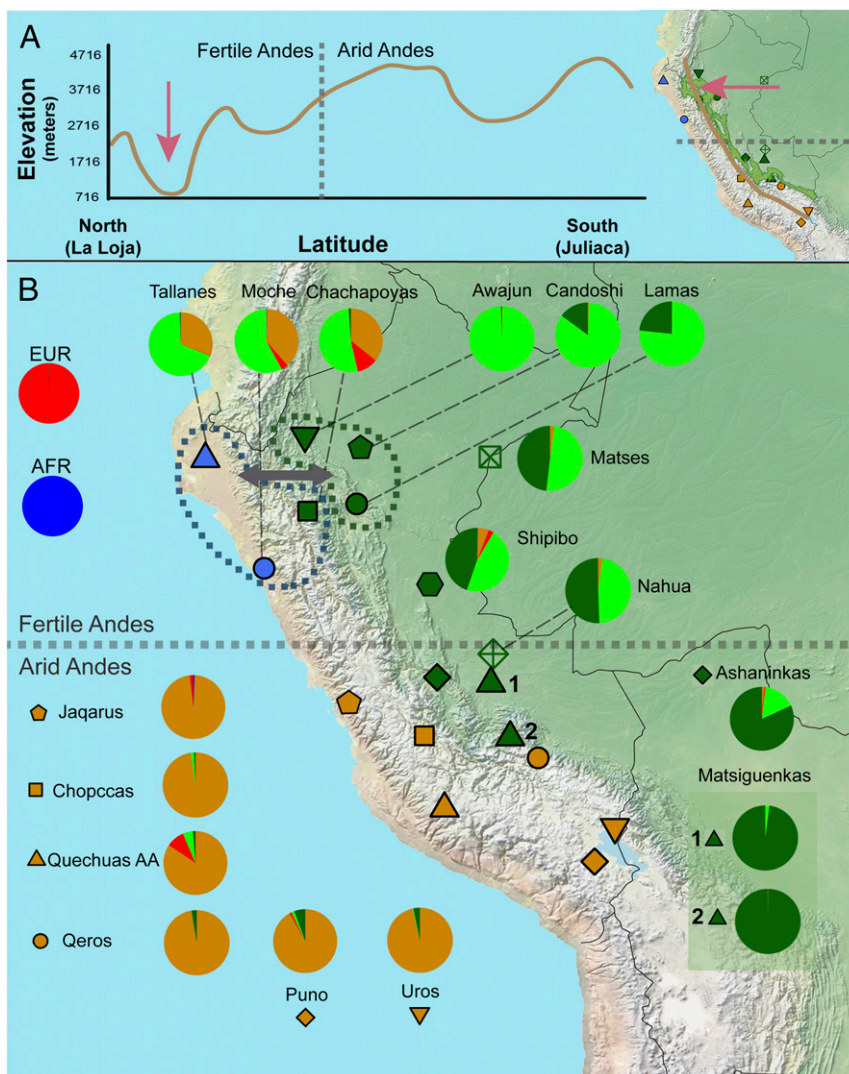


Fig. 1. Genetic and geographic landscape for western South American natives. (A) Elevation vs. latitude plot from a cross-section line in the Andean region from La Loja (Ecuador) to Juliaca (Peru). A vertical line indicates the division between Fertile and Arid Andes (16). The map shows the path used 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, $|Z \text{ score}| > 4$, *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. 14). 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/Globetrotter 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

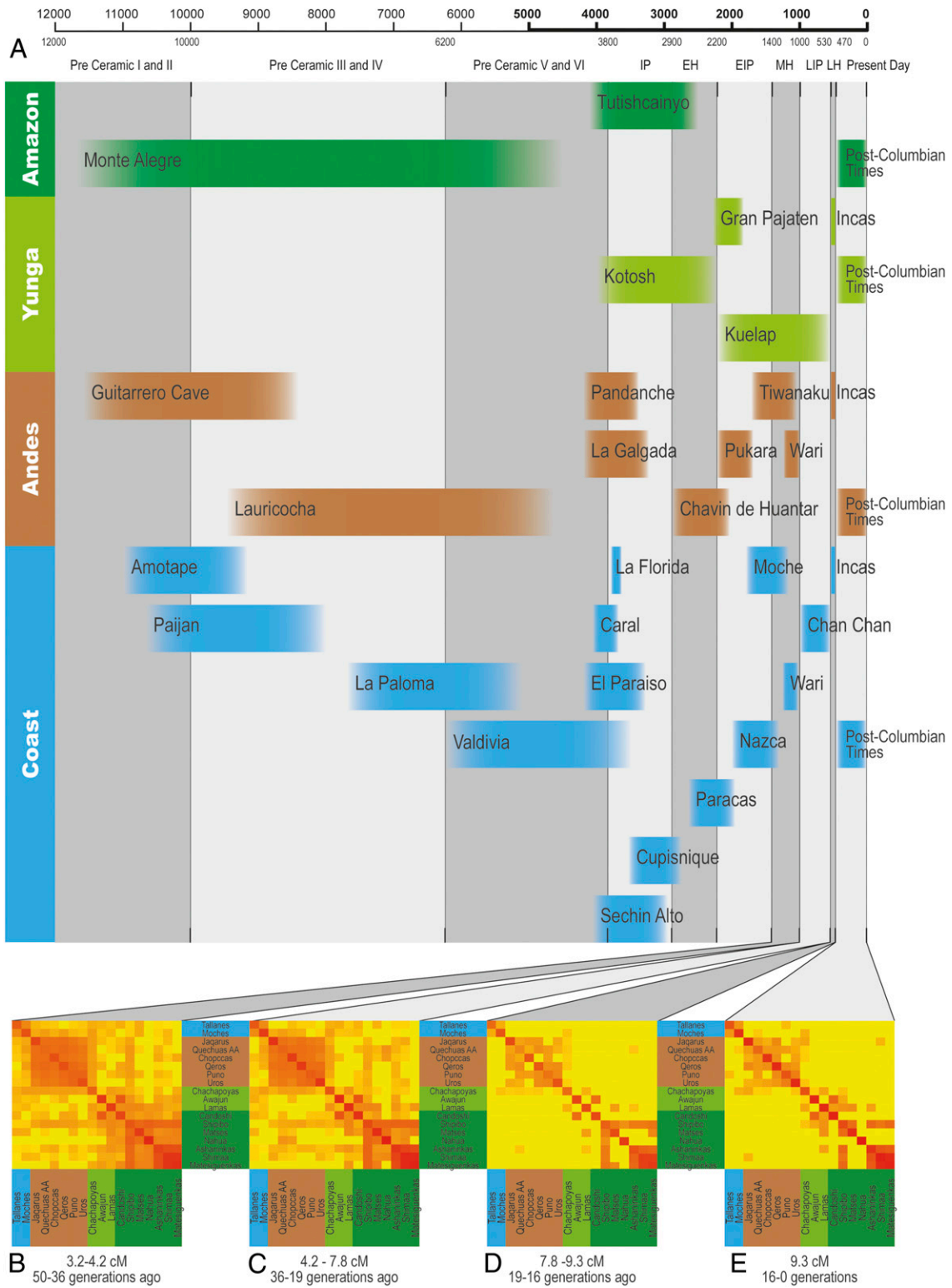


Fig. 2. Changes in IBD sharing over time between the Pacific coast, central Andes, Amazon Yunga, and Amazon and its relationship with the archaeological chronology of the Andes. (A) Key historical events (cultures and archaeological sites) of Peruvian history in four Peruvian longitudinal regions, coast, 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 (Chopccas, 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 *DUOX42*. 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

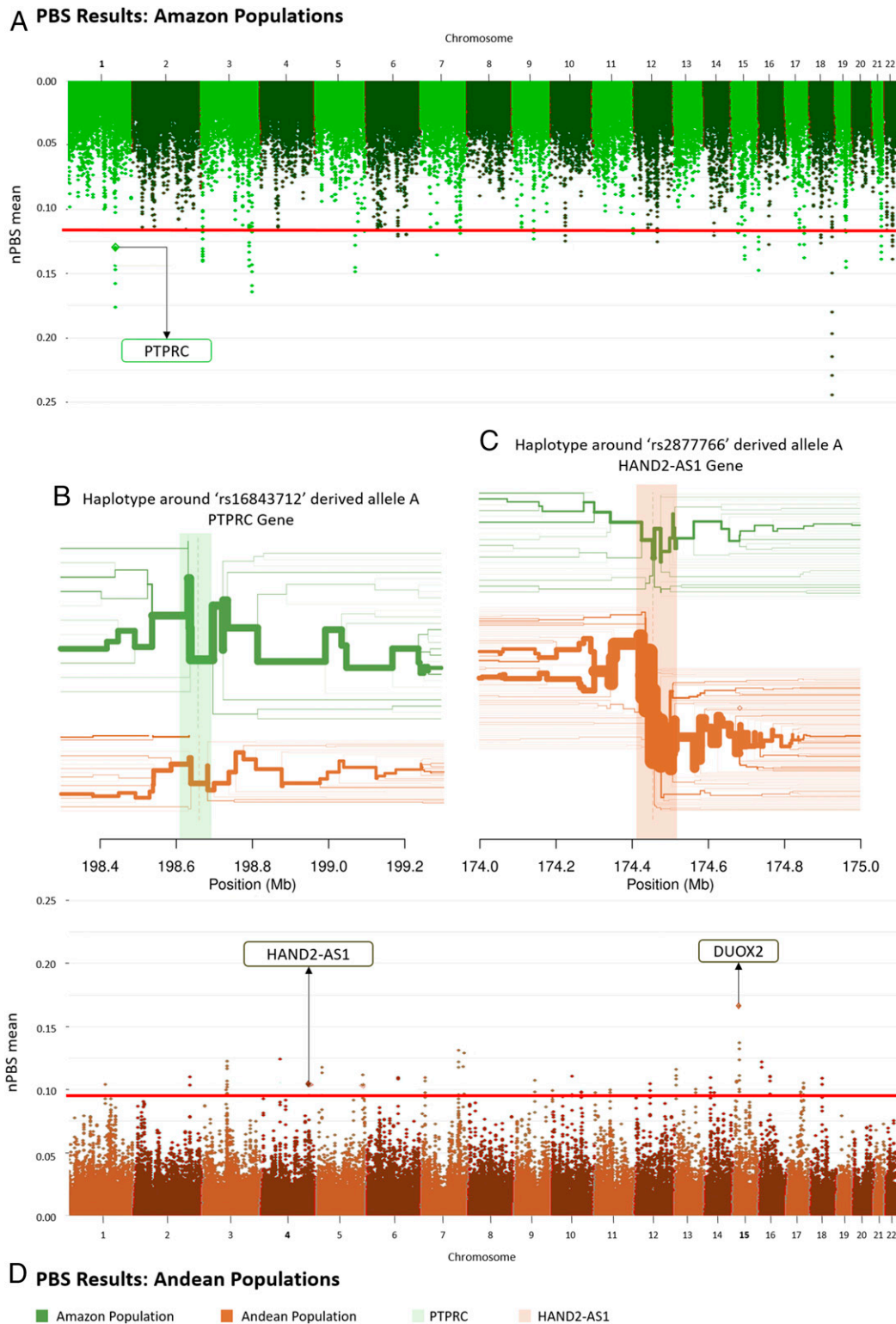


Fig. 3. Sandwich-like representation of natural selection signatures in the arid Andes and the Amazon tropical forest. Manhattan plots (the bread of the sandwich) correspond to the PBSn estimated from the sliding window in Amazon (A) and Andean (D) populations. The horizontal red line shows the 99.95th 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-ASI* (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 Hellenthal, Mark Lipson, Marcia Beltrame, Fabrício Santos, Claudio Struchiner, Ricardo Santos, Luis Guillermo Lumbreras, Sandra Romero-Hidalgo, Víctor Acuña-Alonzo, Miguel Ortega, and Juliana Lacerda, for discussions or technical assistance; Harrison Montejo, Silvia Capristano, Juana Choque, and Marco Galarza from Laboratorio de 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 (PRINT). V.B. was a CAPES/Programa de Estudantes-Convênio de Pós-Graduação (PEC-PG) fellow (grant number 88882.195664/2018-01). P.E.R. was funded by the Fondo Nacional de Desarrollo Científico, Tecnológico y de Innovación Tecnológica (Fondecyt - Perú) (grant number 34-2019, “Proyecto de Mejoramiento y Ampliación de los Servicios del Sistema Nacional de Ciencia, Tecnología e Innovación Tecnológica”). Datasets were

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. Gnechi-Ruscione *et al.*, Dissecting the pre-Columbian genomic ancestry of Native Americans along the Andes-Amazonia divide. *Mol. Biol. Evol.* **36**, 1254–1269 (2019).
16. L. G. Lumbreras, *Los orígenes de la civilización en el Perú*; (Instituto Andino de Estudios Arqueológico-Sociales, 2015).
17. A. Roosevelt, “The maritime, highland, forest dynamic and the origins of complex culture” in *The Cambridge History of the Native Peoples of the Americas*, F. Salomon, S. B. Schwartz, Eds. (Cambridge University Press, 1999), pp. 264–349.
18. C. Barbieri *et al.*, The current genomic landscape of western South America: Andes, Amazonia and Pacific coast. *Mol. Biol. Evol.* **36**, 2698–2713 (2019).
19. H. Silverman, W. Isbell, Eds., *The Handbook of South American Archaeology* (Springer, New York, 2008).
20. J. Haas, S. Pozorski, T. Pozorski, *The Origins and Development of the Andean State* (Cambridge University Press, 1987).
21. E. P. Lanning, *Peru before the Incas* (Prentice-Hall, 1967).
22. W. H. Isbell, “Wari and Tiwanaku: International identities in the central Andean Middle Horizon” in *The Handbook of South American Archaeology*, H. Silverman, W. H. Isbell, Eds. (Springer, New York, 2008), pp. 731–759.
23. G. Valverde *et al.*, Ancient DNA analysis suggests negligible impact of the Wari Empire expansion in Peru’s central coast during the Middle Horizon. *PLoS One* **11**, e0155508 (2016).
24. E. Tarazona-Santos, M. Lavine, S. Pastor, G. Fiori, D. Pettener, Hematological and pulmonary responses to high altitude in Quechuas: A multivariate approach. *Am. J. Phys. Anthropol.* **111**, 165–176 (2000).
25. L. G. Moore, Measuring high-altitude adaptation. *J. Appl. Physiol.* (1985) **123**, 1371–1385 (2017).
26. C. E. G. Amorim, J. T. Daub, F. M. Salzano, M. Foll, L. Excoffier, Detection of convergent genome-wide signals of adaptation to tropical forests in humans. *PLoS One* **10**, e0121557 (2015).
27. G. R. Abecasis *et al.*; 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
28. D. Reich *et al.*, Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
29. M. Raghavan *et al.*, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
30. S. Mallick *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
31. W. C. S. Magalhães *et al.*; Brazilian EPIGEN Consortium, EPIGEN-Brazil initiative resources: A Latin American imputation panel and the scientific workflow. *Genome Res.* **28**, 1090–1095 (2018).
32. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
33. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
34. G. Hellenthal *et al.*, A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
35. N. Patterson *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
36. P. F. Palamara, T. Lencz, A. Darvasi, I. Pe’er, Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.* **91**, 809–822 (2012).
37. N. D. Cook, “Migration in colonial Peru: An overview” in *Migration in Colonial Spanish America*, D. J. Robinson, Ed. (Cambridge University Press, 1990), pp. 41–61.
38. N. Sanchez-Albornoz, *The Population of Latin America: A History* (University of California Press, Berkeley, 1974).
39. J. Lindo *et al.*, The genetic prehistory of the Andean highlands 7000 years BP 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. Corvilain, J. E. Dumont, F. Miot, Roles of DUOX-mediated hydrogen peroxide in metabolism, host defense, and signaling. *Antioxid. Redox Signal.* **20**, 2776–2793 (2014).
52. Y. Maruo *et al.*, Natural course of congenital hypothyroidism by dual oxidase 2 mutations from the neonatal period through puberty. *Eur. J. Endocrinol.* **174**, 453–463 (2016).
53. T. Ueyama *et al.*, The extracellular A-loop of dual oxidases affects the specificity of reactive oxygen species release. *J. Biol. Chem.* **290**, 6495–6506 (2015).
54. E. A. Pretzell *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).

Supplementary Information for

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

Victor Borda^{1,2,3,19}, Isabela Alvim^{1,19}, Marla Mendes^{1,19}, Carolina Silva-Carvalho^{1,19}, Giordano B Soares-Souza¹, Thiago P Leal¹, Vinicius Furlan¹, Marilia O Scliar^{1,4}, Roxana Zamudio¹, Camila Zolini^{1,5,6}, Gilderlanio S Araujo⁷, Marcelo R Luizon¹, Carlos Padilla³, Omar Cáceres^{3,8}, Kelly Levano³, César Sánchez³, Omar Trujillo⁹, Pedro O. Flores-Villanueva³, Michael Dean¹⁰, Silvia Fuselli¹¹, Moara Machado^{1,10}, Pedro E. Romero¹², Francesca Tassi¹¹, Meredith Yeager¹⁰, Timothy D O'Connor^{13,14,15}, Robert H Gilman^{12,16}, Eduardo Tarazona-Santos^{1,17,20}, Heinner Guio^{3,8,18,20}

Eduardo Tarazona-Santos, Heinner Guio.

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

This PDF file includes:

- Supplementary text
- Figures S1 to S30
- Legends for Datasets S1 to S10
- SI References

Other supplementary materials for this manuscript include the following:

- Datasets S1 to S10

Supplementary Information Text

Section 1: Sampling, quality control and Datasets

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

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

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

1.2. DNA sampling: For INS data, after collecting the 10ml blood sample from participants (we only collect blood) we proceed to extract the DNA using commercial kits (Qiagen, USA). For this procedure, we travel to the community with our supplies and equipment to perform the DNA extraction, when we obtain the DNA purified we aliquoted, send to our Laboratory (Laboratorio de Biotecnología y Biología Molecular) in Lima, Perú to be stored in -80 °C until we continued with the next project phase genotyping. Specifically, for LDGH data, we extracted genomic DNA from blood samples using the Gentra Puregene blood kit (Qiagen, USA) in the LDGH.

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

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

Family structure affects the analysis of population structure as a familiar cluster can be confounded with a discrete population (8). To overcome this issue, we estimated the kinship coefficients (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:

- a) **Natives 1.9M Dataset** (1,927,769 SNPs/673 individuals): Dataset with maximum number of genotyped SNPs. This dataset includes just Peruvian Native individuals from INS and LDGH and 107 Iberian (IBS), 108 Yoruba (YRI) and 100 East Asian individuals (CDX) from 1000 Genomes Project (Dataset S1).
- b) **Natives 500K Dataset** (567,718 SNPs/849 individuals): This dataset includes individuals from **Natives 1.9M Dataset**, Native American, Siberian, South Asian (Onge) and Oceanian (Bougainville and Papuan), individuals from the Simons Project (13), Raghavan *et al.*, 2015 and 79 individuals from Guatemala of Michael Dean NCI lab (Dataset S2) genotyped for 600K

SNPs. The Guatemalan sample includes individuals from the Kaqchikel native population and non-native individuals with more than 99% of Native American ancestry.

- c) **Natives 230K Dataset** (235,352 SNPs/1,286 individuals): Dataset with maximum number of individuals. This data includes individuals from Natives 1.9M dataset and all Native Americans from the unmasked data of Reich *et al.* (2012) (~300K SNPs), which includes HGDP individuals (Dataset S3).

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

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

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

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

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

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

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

Section 2: Genetic relationships in Western South America

To address our scientific questions:

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

and

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

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

2.1. Methods

2.1.1. Population Structure using genotype based methods

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

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

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

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

Natives 1.9M dataset

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

For the PCA (Fig. S3), we excluded Africans (YRI) due to its high level of differentiation that masks the relationships in Native Americans. The first principal component (PC1, Variance explained=2.36%) showed an axis of differentiation between the European and Native American groups. We observed that some Andean, Moches, and Chachapoyas individuals have some degree of European ancestry. The PC2 (Variance explained=1.2%) separated Western (Andean and SPPC

populations) and Eastern (Amazon) South American natives. Chachapoyas showed affinity with SPPC populations. Jivaroan populations (Awajun and Candoshi), were intermediate in the axis Western-Eastern. Furthermore, the PC2 showed a cline for the genetic diversity of the Amazon populations, from North (Matses) to South (Matsigenkas). As in ADMIXTURE, both Matsigenkas groups, Shimaa (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 a specific ancestry cluster (darkgreen) which is not shared with other populations of the same linguistic group (Ashaninkas). Panoan populations (Shipibo, Matses and Nahua) showed a predominant ancestry associated with the Ashaninkas population. Jivaroan groups showed a specific ancestry which was predominant in the Awajun population.

For the Principal Component Analysis (Fig. S5), the PC1 separated Native Americans from Europeans. Some Chachapoyas, Quechuas_AA, 1 Moche, 1 Mixtec and 1 Shipibo individuals showed affinity to Europeans due to admixture. The PC2 separated Amazon from non-Amazon populations; Jivaroan and Tupian individuals were observed as intermediate between these groups. The PC3 separates a group that includes Andean and 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.

Natives 230K dataset

The following description was focused on clusters related to South American natives. The lower CV error was obtained for 18 ancestral clusters (K = 18, Fig. S6). The ADMIXTURE plots (Fig. S6) from K=3 to K=5 inferred continental ancestry clusters. The ADMIXTURE plot K=5, five ancestry clusters related to each continental region were identified: Africa, Europe, Asia, Oceania (light purple, pops 90-91) and America. ADMIXTURE K=6 showed a cluster (dark green) associated with Arawakan groups (Ashaninkas and Matsigenkas). ADMIXTURE K=8 identified a cluster (light pink) for Costa Rican natives (Bribri, Cabecar, Chorotega, Guaymi, Huetar, Maleku, Teribe). ADMIXTURE K=9 inferred an ancestral cluster (black) related to Mesoamericans, predominantly in Pima. ADMIXTURE K=10 showed an ancestry cluster (light green) related to the Awajun population which represented most of non Andean natives. ADMIXTURE K=12 identified an ancestry cluster (gray) associated with Tupian populations (Surui and Karitiana) also observed in Mesoamericans and in non Andean natives. Furthermore, a brown cluster is predominant in the Andean populations. ADMIXTURE K=13 showed a light blue cluster associated with Pima natives. ADMIXTURE K=15 showed an ancestry

cluster, darkgreen and forest green, for each Peruvian Arawakan individuals (Ashaninkas and Matsigenkas). ADMIXTURE K=17 detected a cluster (light blue) associated with the SPPC population, and which is also present in Chachapoyas and Lamas. The ADMIXTURE K=18 with lowest cross validation showed differentiation in Asian natives. ADMIXTURE K=21 showed an ancestry cluster related to Lamas sample.

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

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

2.1.2. Population Structure using haplotype based methods

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

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

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

Since our main question is about the history of Native Americans, we excluded individuals with >5% of Non-Native ancestry in the Natives 230K dataset except for Chachapoyas individuals. We did this because Chachapoyas population has almost all individuals with more than 5% of Non-Native ancestry, so we maintained all individuals (see ADMIXTURE results). We removed slightly admixed individuals just in the 230K dataset because being the most numerous dataset, the exclusion of 202 individuals did not represent a considerable loss. We determined which individuals had to be removed

based on ADMIXTURE results (K=3). For Natives 1.9M and Natives 500K datasets we maintained the complete dataset.

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

Natives 1.9M dataset

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

Natives 500K dataset

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

Natives 230K dataset

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

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

2.1.3. Ancestry profiles inferred by GLOBETROTTER and SOURCEFIND

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

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

Natives 1.9M dataset

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

Natives 500K Dataset

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

Natives 230K Dataset

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

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

- Maya (n=29)
- Kaqchikel (n=6)
- Mixe (n=17)
- Zapotec1 (n=6)
- Zapotec2 (n=21)
- Mixtec (n=5)
- Karitiana (n=8)
- Surui (n=10)
- Guahibo (n=6)
- Palikur (n=3)
- Piapoco (n=6)
- Ticuna (n=4)
- Embera (n=4)
- Kogi (n=3)
- Waunana(n=3)
- Wayuu (n=2)
- Toba (n=4)
- East_Amazon_Brazil [Arara (n=1) - Parakana (n=1)]
- Inga (n=2)
- 1 Chaco cluster [Guarani (n=3) - Chane (n=2)]
- Wichi (n=4)
- Maleku (n=2)
- Cabecar (n=16)
- Guaymi (n=5)
- Teribe (n=3)
- Chipewyan (n=3)
- East Greenland (n=3)
- IBS cluster (n=107)
- CDX cluster (n=93)

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

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

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

2.2. Conclusions

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

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

3.1. Introduction

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

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

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

3.2. Methods

3.2.1. *D* statistics

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

$$\text{Outgroup } (P_3 \quad (P_1, \quad P_2))$$

This treeness test considers, as a null hypothesis, that P_1 , P_2 , P_3 and an outgroup have a tree relationship. In this null hypothesis, P_1 and P_2 diverged earlier from the ancestor of P_3 , with no gene flow between P_3 and P_1 or P_2 after the divergence. The alternative hypothesis is that P_3 was involved in gene flow with P_1 or P_2 after the divergence. This analysis is restricted to biallelic sites and considered an allele "A" as the ancestral allele of the outgroup (O) and an alternative allele "B" in P_3 . Considering the order O- P_3 - P_1 - P_2 , the *D* test is focused on the ABBA or ABAB pattern. The first pattern (ABBA) corresponds to the total number of sites with the alternative allele (B) shared only by P_1 and P_3 . The second configuration (ABAB) is the total number of sites with the alternative allele (B) shared only by P_2 and P_3 . The *D* statistic is calculated by the relationship of: the difference of ABBA-ABAB counts (numerator) and the total count ABBA+ABAB (denominator). The numerator of this relationship indicates the signal of the statistic which is interpreted as the direction of gene flow. If the *D* statistic is not significantly different from zero, we accept the null hypothesis of treeness. If the result differs significantly from zero, we reject the null hypothesis and consider the possibility of gene flow between P_3 with P_1 or P_2 . For *D* statistics estimation we used ADMIXTOOLS (16). The results are

interpreted as follows: negative values of D are interpreted as gene flow between P_1 and P_3 and positive values indicate a gene flow between P_2 and P_3 . A D value is considered statistically significant if the absolute value of the relationship between the D value and its standard deviation is equal or above 3 ($|Z\text{-score}| \geq 3$). We applied D statistics to the masked Datasets 500K and 230K. Considering the huge number of combinations for D inferences and even obtaining highly significant values, we construct Q-Q plots to analyze the Z-score distribution, which is expected to be approximately normal under the hypothesis that our results could be expected by chance (H_0) (33).

Cultural interactions and gene flow across the Andes

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

- (Outgroup, Eastern (Western₂, Western₁))
- (Outgroup, Western (Eastern₁, Eastern₂))

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

Natives_500K Dataset (Masked)

Configuration (**Outgroup, Western (Eastern₂, Eastern₁)**):

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

Configuration (**Outgroup, Eastern (Western₂, Western₁)**):

When we explored the possible signal of gene flow from Eastern (Amazon yunga and Lower Amazon) to Western, we found results similar to the first configuration, a significant deviation from the diagonal involving SPPC, with Chachapoyas, Lamas and Jivaroan populations (Fig. S14B).

Natives_230K Dataset (Masked)

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

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

3.2.2. Admixture graphs

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

We test two contrasting hypotheses of gene flow. **The first one** ($W \rightarrow E$), Amazon groups (Eastern) receives a contribution from the Coast (Western) and **the second one** ($E \rightarrow W$), Coast (Western) receives a contribution from Amazon (Eastern). As a result, qpGraph offers a log-likelihood of the fit, Z-score values for the f -statistics and the parameters involved in admixture events (contributions) if

these are tested. We accept an hypothetical tree if the absolute value of the highest f -statistic is less than 3 ($|Z\text{-score}| < 3$). Moreover, if we obtain trees with similar Z -score values, we select the tree with fewer zeroed branches and lower log-likelihood.

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

African:	YRI
Siberian:	MA1
South Asian:	Onge
East Asian:	CDX
North America:	Clovis
Mesoamerican:	Mixe
Amazon:	Ashaninkas and Matses
Andes:	Uros

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

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

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

3.3. Conclusion

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

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

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

4.1. Methods

4.1.1 Identical-by-descent segment analysis

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

$$E \approx 3/2L$$

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 present day) |

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

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

$$abIBD = \frac{\sum_{ij} L_{ij}}{N_{pairs}}$$

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

N_{pairs} : $N_a * N_b$ (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 homogenous population. For this reason we performed the analysis just for the arid Andean group. To avoid the underestimate of effective population size, we restricted the analysis to segments larger than 4cM, as suggested by the authors for array data (41). We inferred this parameter (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.

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):

$$PBS_n = PBS_1 / (1 + PBS_1 + PBS_2 + PBS_3)$$

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 PBS_n.

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 Csi2 (58) to simulate the genetic data under a neutral model as described:

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

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

```
gene_conversion_relative_rate 0.0000000045
```

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

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

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

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

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

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

```
#coast
```

```

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

pop_event exp_change_size "Andean second expansion" 2 4 9 8064 2500
pop_event bottleneck "Andean bottleneck due to European conquest" 2 29 0.067
pop_event bottleneck "Coast bottleneck due to European conquest" 4 29 0.067
pop_event bottleneck "Amazon bottleneck" 1 479 0.067
pop_event exp_change_size "Andean expansion" 2 30 450 7000 2426
pop_event exp_change_size "Coast expansion" 4 4 9 6975 1500
pop_event split "andean and amazon split" 2 1 516
pop_event split "andean and coast split" 2 4 508
pop_event bottleneck "native bottleneck" 2 959 0.067
pop_event split "asian and native split" 3 2 960
pop_event bottleneck "asian bottleneck" 3 1998 0.067

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

```

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

5.2.2. xpEHH: cross population extended haplotype homozygosity

Positive selection events increase the frequency of a genetic variant and, consequently, the frequency of the variants around it (59). This process occurs faster than the haplotypes are broken down by recombination, leading to the emergence of an unusual high frequency long-range haplotype. Considering this, we decided to perform an extended haplotype homozygosity (EHH) test to select the most likely candidates for natural selection.

Sabeti *et al.* (60, 61) developed methods to detect natural selection signatures calculating the EHH, defined as the probability of finding homozygosity of all SNPs around the haplotype of interest choosing two random chromosomes containing this haplotype in a population:

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

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

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

We considered as positive signals for natural selection the SNPs representing the 99.5 percentile of the xpEHH results of an Andean vs Amazon comparison (xpEHH > 2.97 for the Andes and xpEHH <

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

To assess whether the results obtained from xpEHH analysis between Andean and Amazon populations are corroborated when comparing each group with an outgroup, we performed xpEHH analysis between each group and East Asian populations from 1000 genomes (xpEHH_{ANDvsEAS} and xpEHH_{AMZvsEAS}). The results for Andean populations show a high score for gene HAND2-AS1 (higher xpEHH_{ANDvsEAS}=1,59, p-value=0.001), but not for gene RARS (higher xpEHH_{ANDvsEAS}=0,79 p-value=0,105). For Amazon populations, the highest signal was from an intergenic region near the gene PTPRC (rs1326288 higher xpEHH_{AMZvsEAS}=-1,23 p-value=0.026)(Tables S3, S4).

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

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

5.2.3 Linkage Disequilibrium Patterns in natural selection signals

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

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

5.2.4 Identification of regulatory elements located around *HAND2-AS1* locus

GeneHancer track available in the UCSC Genome Browser (90 80) was used to identify active regulatory elements (enhancers and promoters) that may target *HAND2-AS1* (Fig. S29). GeneHancer database was created by integrating >1 million regulatory elements from seven genome-wide databases: ENCODE project Z-Lab Enhancer-like regions (version v3); Ensembl regulatory build (version 92); functional annotation of the mammalian genome (FANTOM5) atlas of active enhancers; VISTA Enhancer Browser; dbSUPER super-enhancers; Eukaryotic Promoter Database (EPDnew) promoters; and UCNEbase ultra-conserved noncoding elements. Genes were linked to enhancers by GeneHancer using five methods: eQTLs from GTEx (v6p); Capture Hi-C promoter-enhancer long range interactions; FANTOM5 co-expression of enhancers in the form of noncoding enhancer RNA;

transcription factor co-expression; and gene target distance. For this analysis a “double elite” dataset was considered, which is composed of regulatory elements derived from more than one database (elite enhancers) that are associated to genes from more than one method (elite association).

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

5.2.5 F-statistics

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

5.3. Conclusions

- We have confirmed a natural selection signal from a gene previously reported in Andean populations, *DUOX2* (76) ($PBSn=0.22$ p-value=0.002, $xpEHH=-2.647$ p-value=0.991).
- We identified Natural selection signals Andeans in genes related to (Tab. S4):
 - High altitude adaptation: *SULT1A1*: $PBSn=0.167$ p-value=0.007; *RARS*: $PBSn=0.15$ p-value=0.010, $xpEHH=2.980$ p-value=0.0025 (82, 83),
 - Heart development: *HAND2-AS1*: $PBSn=0.21$ p-value=0.003, $xpEHH=4.481$ p-value $<2e-5$ (84),
 - Immune response: *UBQLN4*: $PBSn=0.17$ p-value=0.007, $xpEHH=-0.217$ p-value=0.607; *SSR2*: $PBSn=0.17$ p-value=0.007, $xpEHH=-0.215$ p-value=0.606; *DUOX2*: $PBSn=0.22$ p-value=0.002, $xpEHH=-2.647$ p-value=0.991 (85–87).
- We identified Natural selection signals in Amazon populations related to (Tab. S5):
 - Immune response: *PTPRC*: $PBSn=0.265$ p-value=0.004, $xpEHH=-4.222$ p-value=0.0003 (88),
 - Food intake regulation: *MCHR1*: $PBSn=0.26$ p-value=0.004 (89),
 - Lipid transport: *ABCA9*: $PBSn=0.21$ p-value=0.008, $xpEHH=-1.570$ p-value=0.060, *ABCA6*: $PBSn=0.19$ p-value=0.011, $xpEHH=-1.362$ p-value=0.084 (90, 91).

Supplementary Figures

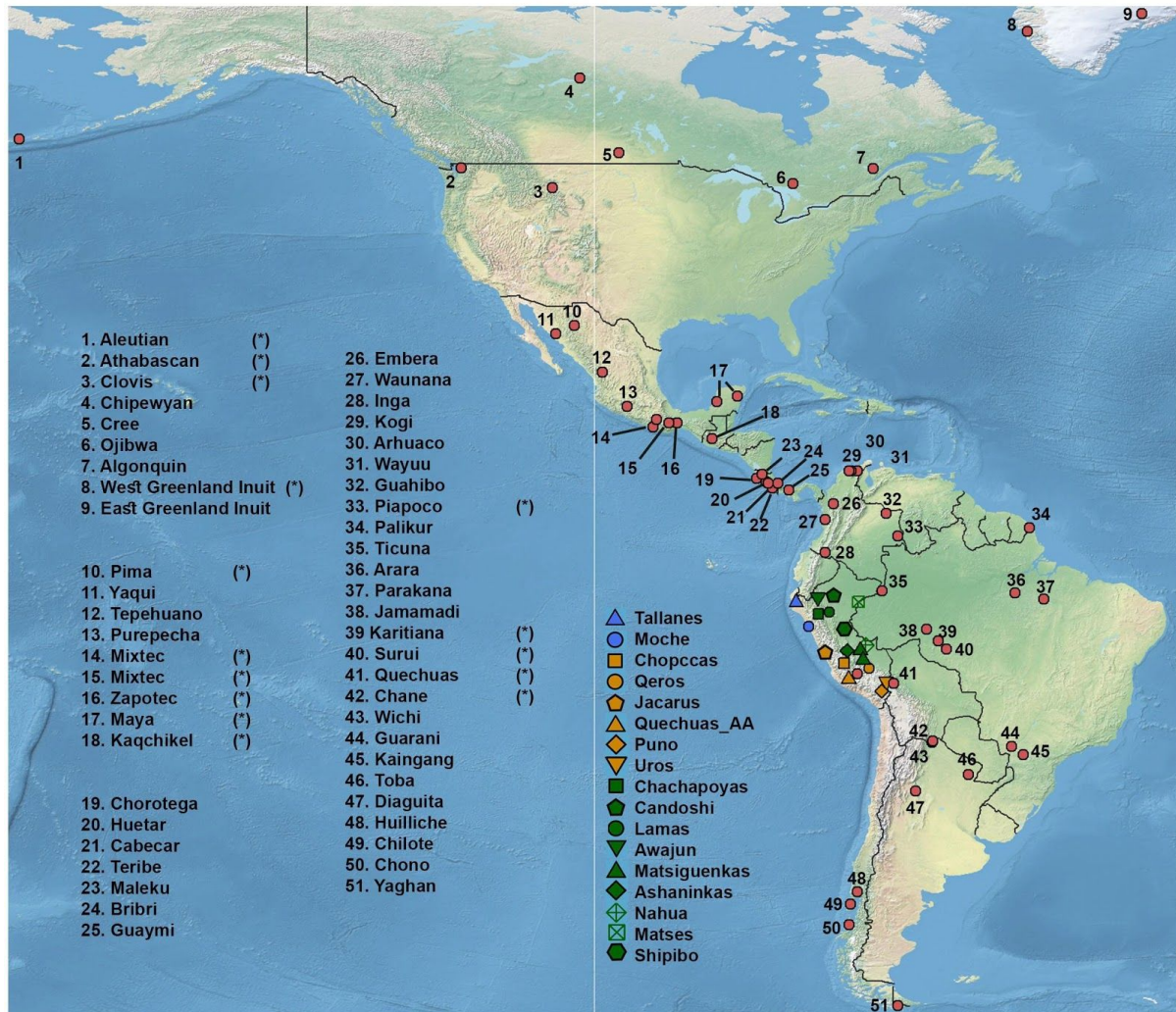


Figure S1. Geographical distribution for the 18 Peruvian Native populations sampled, plus the 65 sampled Native American populations and public data sets (Mallick *et al.* 2016, Raghavan *et al.* 2015, Reich *et al.* 2012). All samples except Clovis and Athabascan were included in a data set of ~ 230,000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

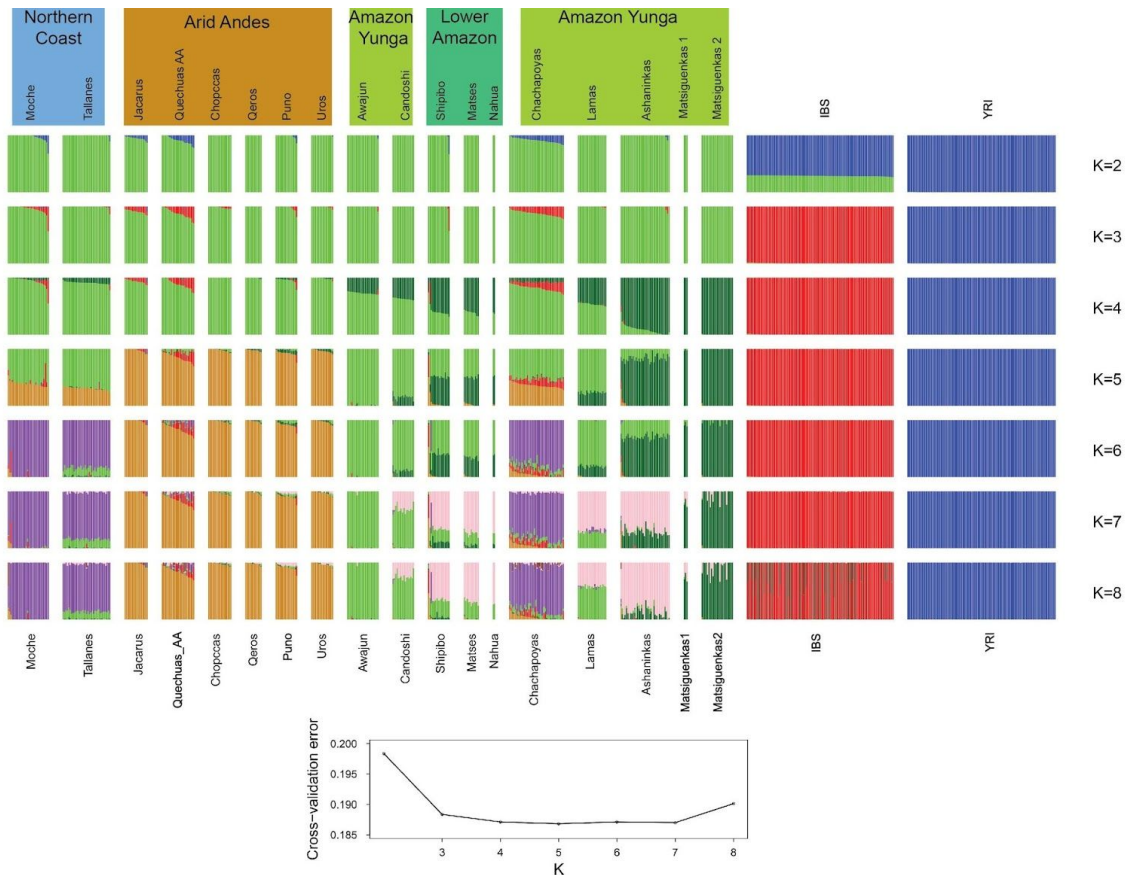


Figure S2. ADMIXTURE analysis for 18 Native American populations, as well as Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 1.9M Dataset). Figure shows results for 2 to 8 ancestral clusters (K) and a plot (Bottom) with the ADMIXTURE cross-validation errors as a function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

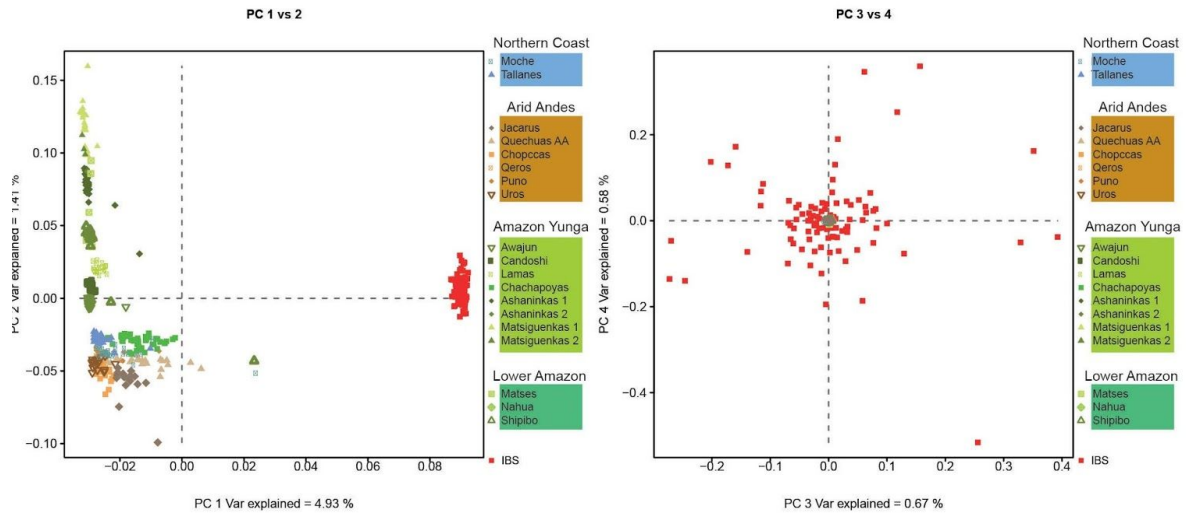


Figure S3. Principal Component Analysis for 18 Native American Peruvian populations and Iberian individuals (IBS) from 1000 Genomes Project (Natives 1.9M Dataset). Shades of blue are related to Coast populations. Orange-brown colors are related to Andean populations and green colors are related to Amazon.

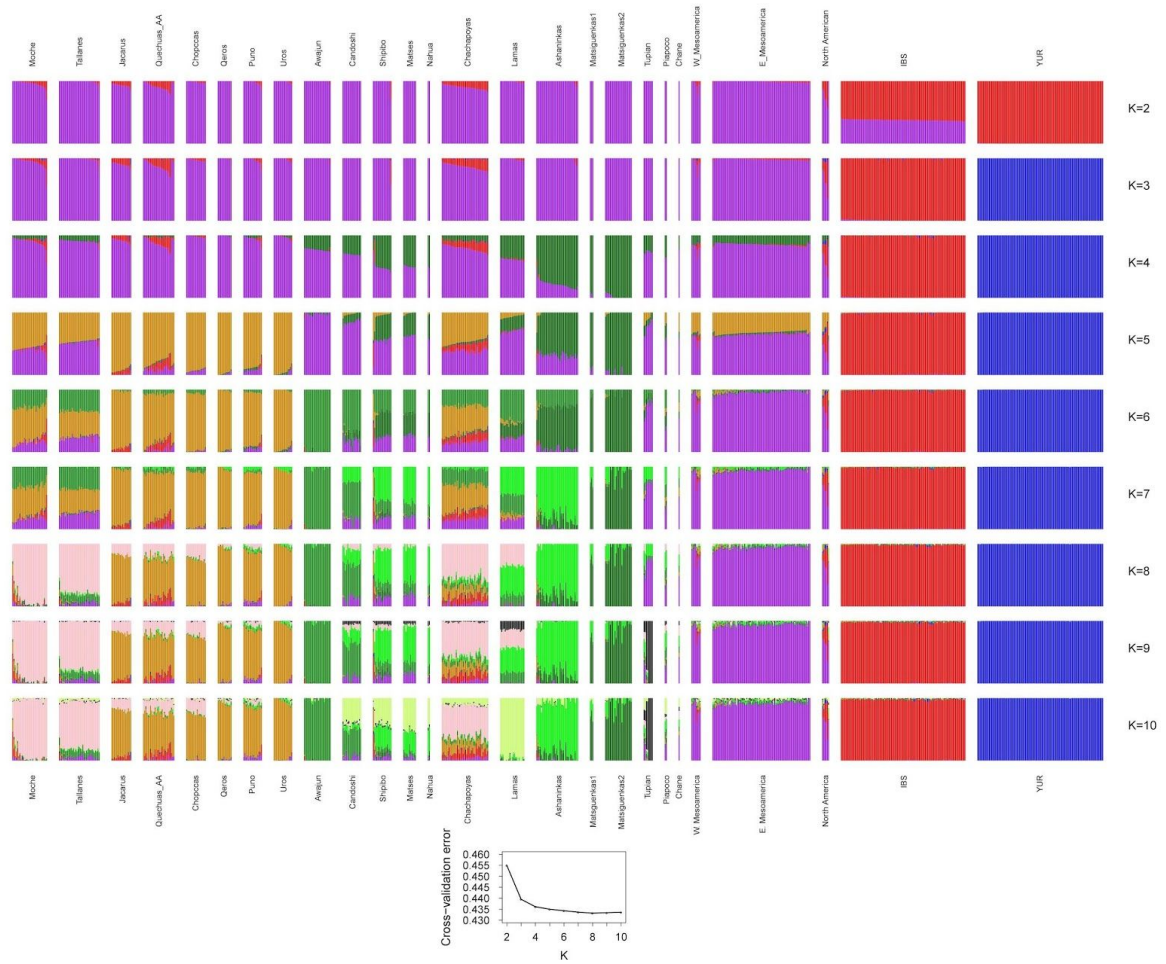


Figure S4. ADMIXTURE analysis for 18 Native American Peruvian populations, Guatemala samples, Native Americans from Raghavan *et al.* 2015 and the Simons Project (Mallick *et al.* 2016) Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 500K Dataset). Figure shows results for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

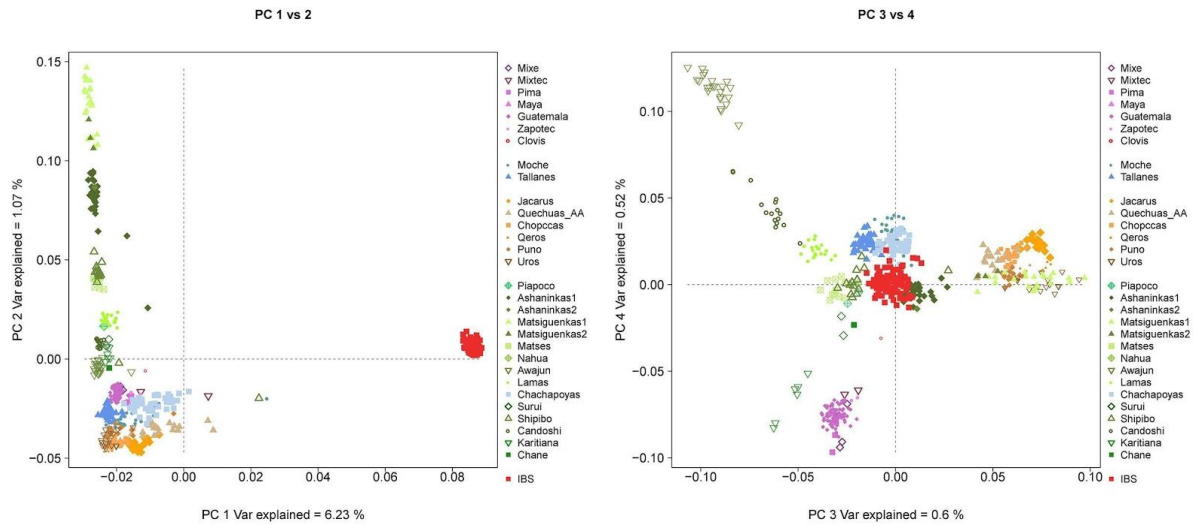


Figure S5. Principal Component Analysis for 18 Native American Peruvian populations, Guatemala samples, Native Americans from Raghavan *et al.* 2015 and the Simons Project (Mallick *et al.* 2016) and Iberian (IBS) populations from 1000 Genomes Project (Natives 500K Dataset). Shades of blue are related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

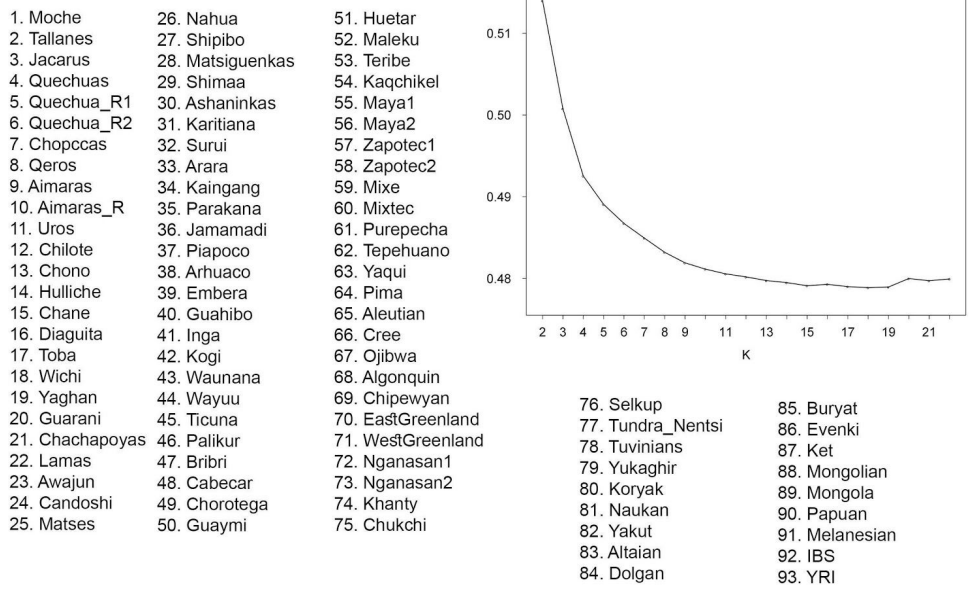
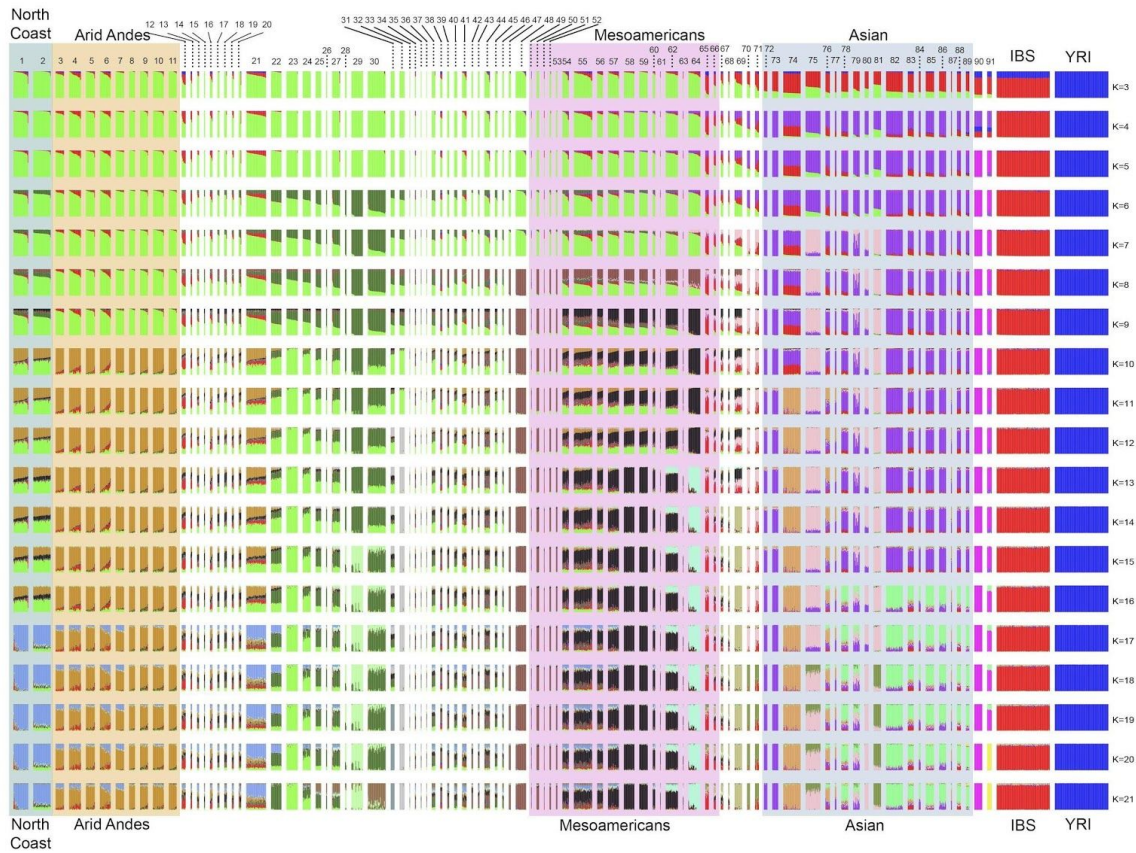


Figure S6. (Top) ADMIXTURE analysis for 90 worldwide populations including 71 Native American populations, 18 Asian, 2 Oceanian, Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 230K Dataset). Figure shows results for 3 to 21 ancestral clusters (K). (Bottom) ADMIXTURE cross-validation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

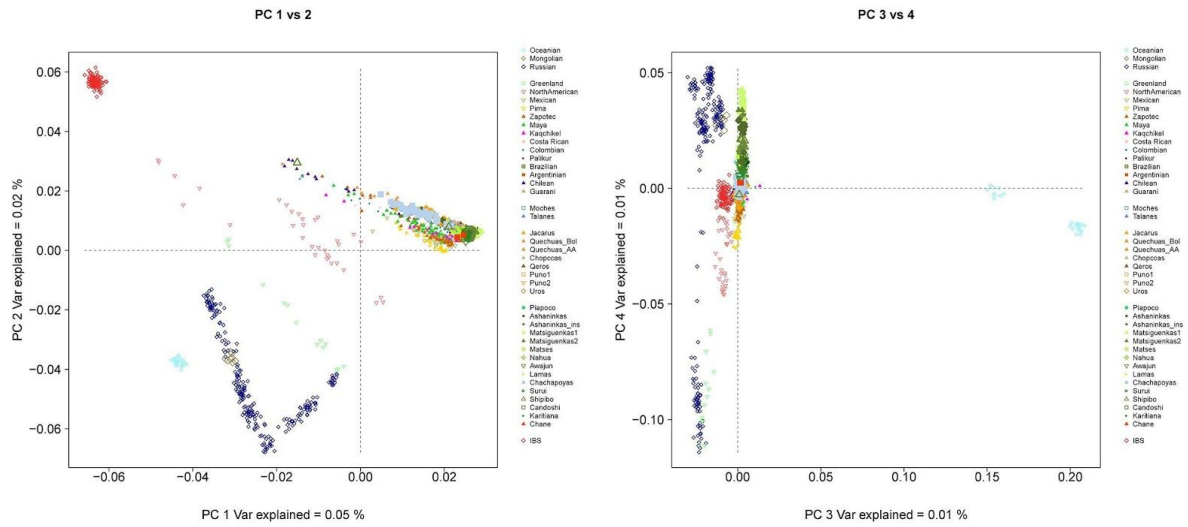


Figure S7. Principal Component Analysis for 89 worldwide populations including 68 Native American populations, 18 Asian populations, 2 Oceanian populations, Iberian (IBS) populations (Native 230K Dataset).

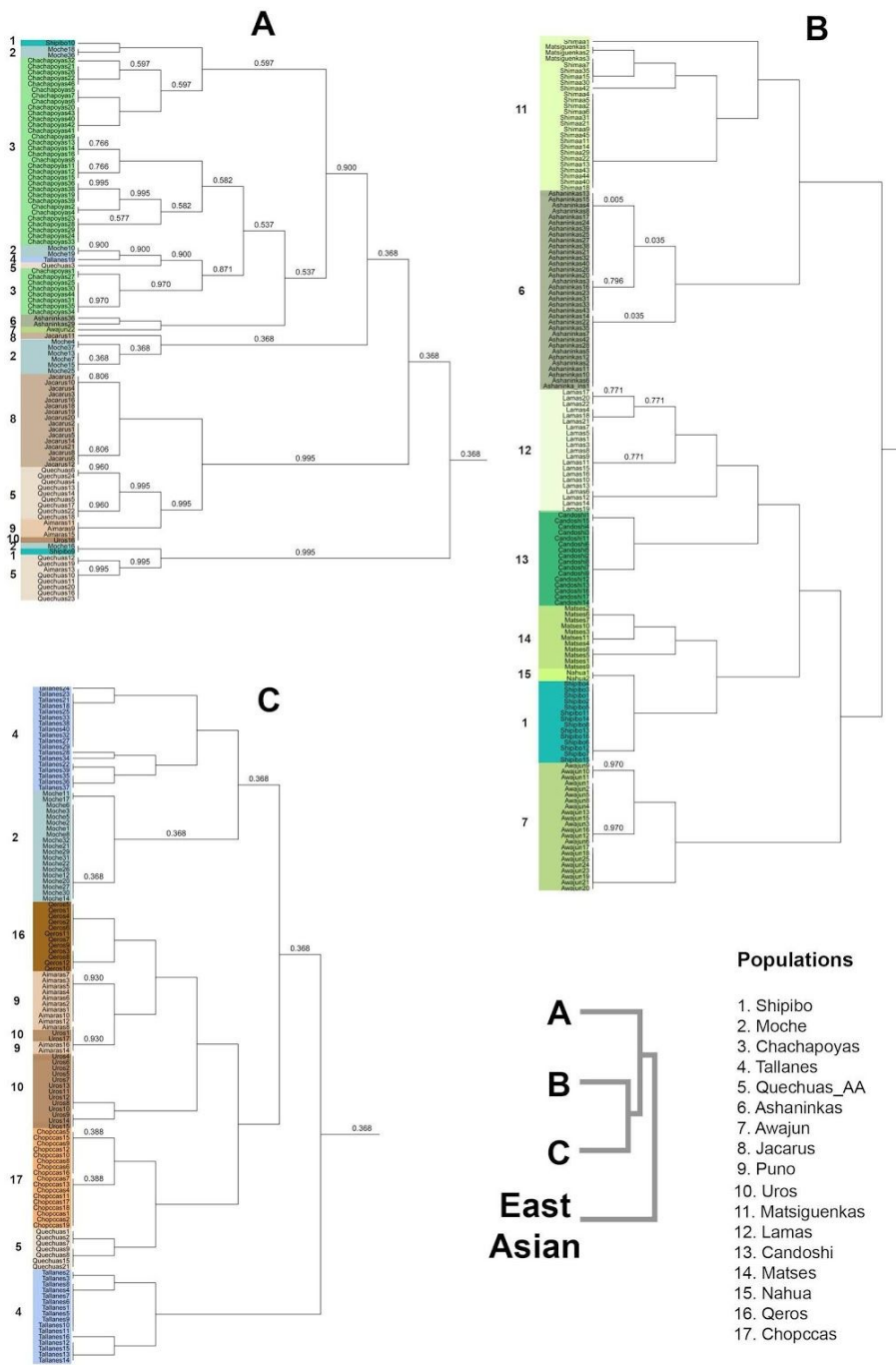


Figure S8. fineSTRUCTURE clustering analysis for the Dataset 1.9M dataset. The tree shows the haplotype sharing between Native Americans and East Asian samples. Figures A, B, and C represent the clusters A, B, and C, respectively, in the tree on the right. East Asian clusters grouped all Asian samples.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

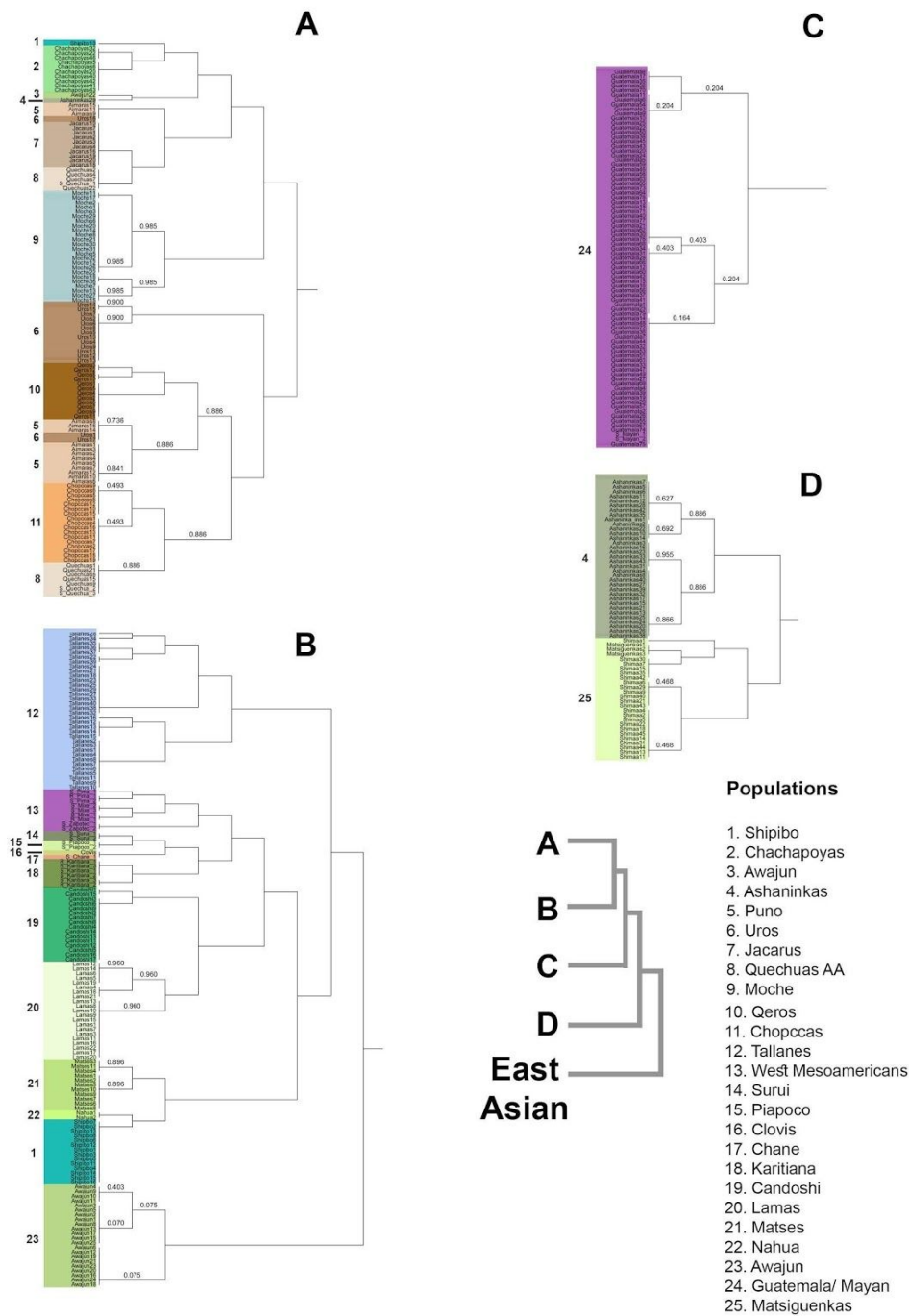


Figure S9. fineSTRUCTURE clustering analysis for the Dataset 500K dataset. The tree shows the haplotype sharing between Native Americans and East Asian samples. Figures A, B, C and D represent the clusters A, B, C and D, respectively, in the tree on the right. East Asian clusters grouped all Asian samples.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

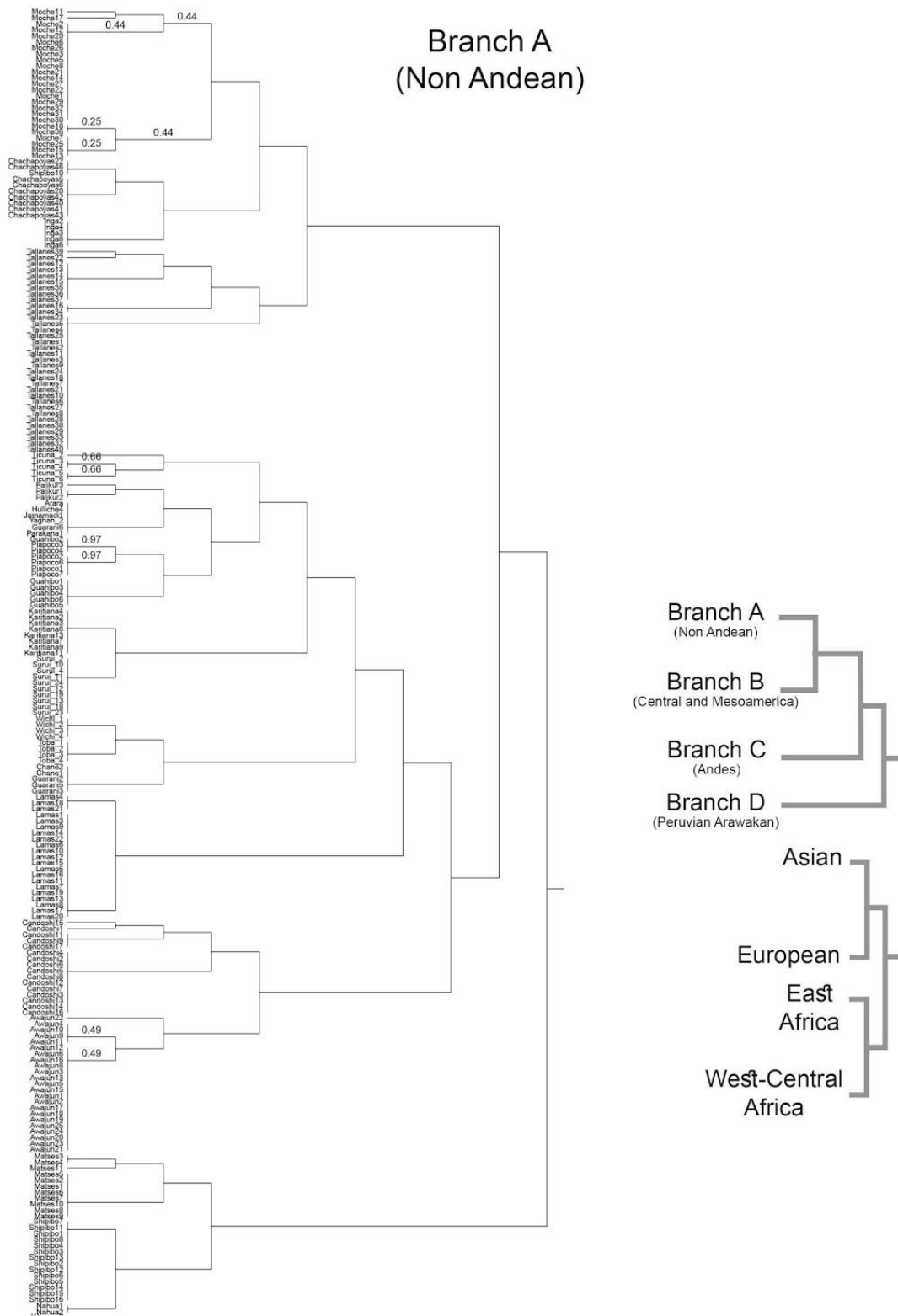


Figure S10. fineSTRUCTURE clustering analysis for the Dataset 230K dataset. A) Branch A of the tree showing the clustering of the Non Andean populations of South America. B) Branches B (Central Americans and Mesoamericas), C (Andean populations) and D (Peruvian Arawakan Ashaninkas and Matsigenkas).

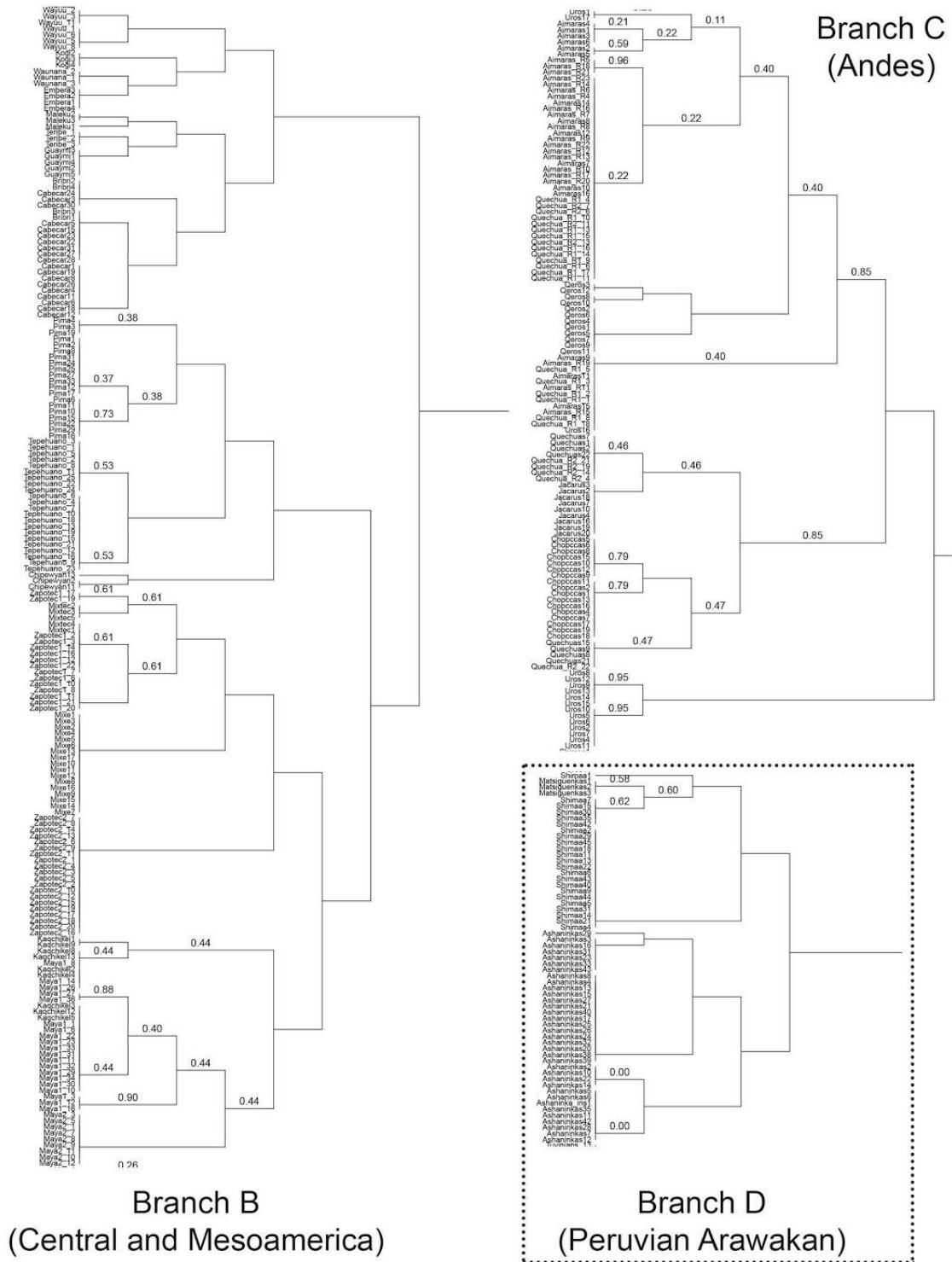


Figure S10 (Continued). fineSTRUCTURE clustering analysis for the Dataset 230K dataset. A) Branch A of the tree showing the clustering of the Non Andean populations of South America. B) Branches B (Central Americans and Mesoamericas), C (Andean populations) and D (Peruvian Arawakan Ashaninkas and Matsigenkas).

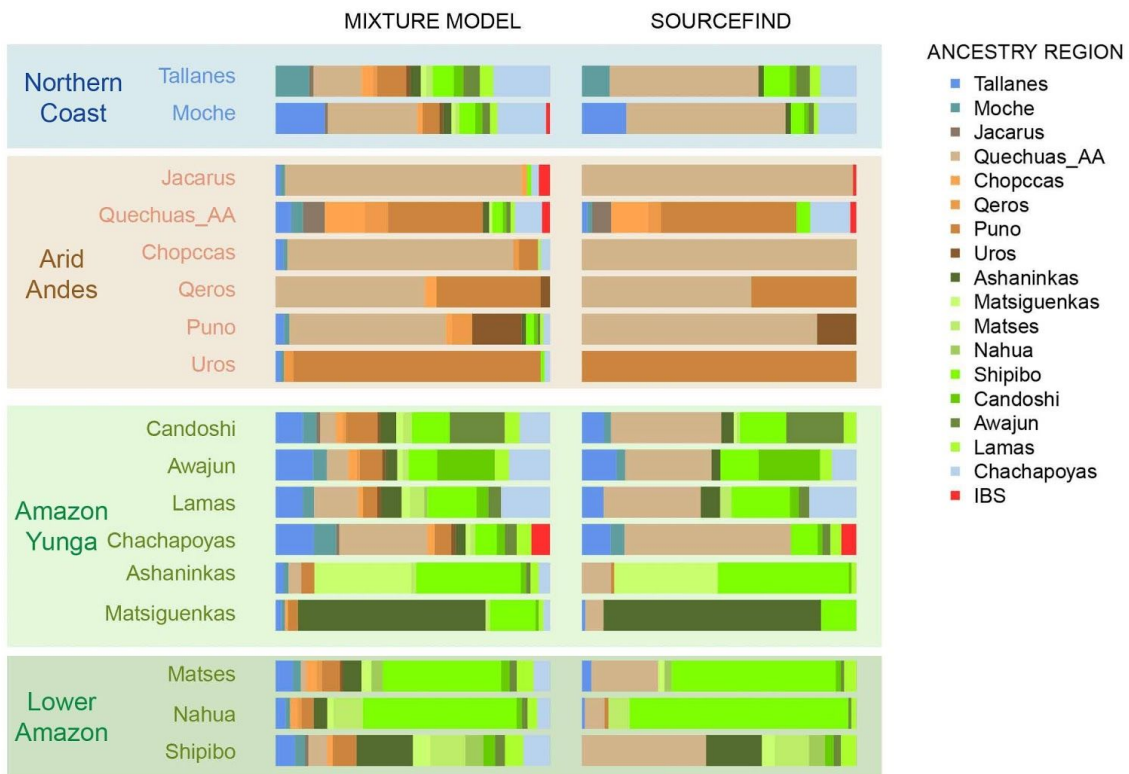


Figure S11. Proportions of haplotype sharing for each target population respect to Native Americans, Europeans and Africans donors populations, for the Dataset Natives 1.9M, inferred by two approaches: A non negative regression (MIXTURE MODEL) and a Bayesian approach (SOURCEFIND). Colored bars indicate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

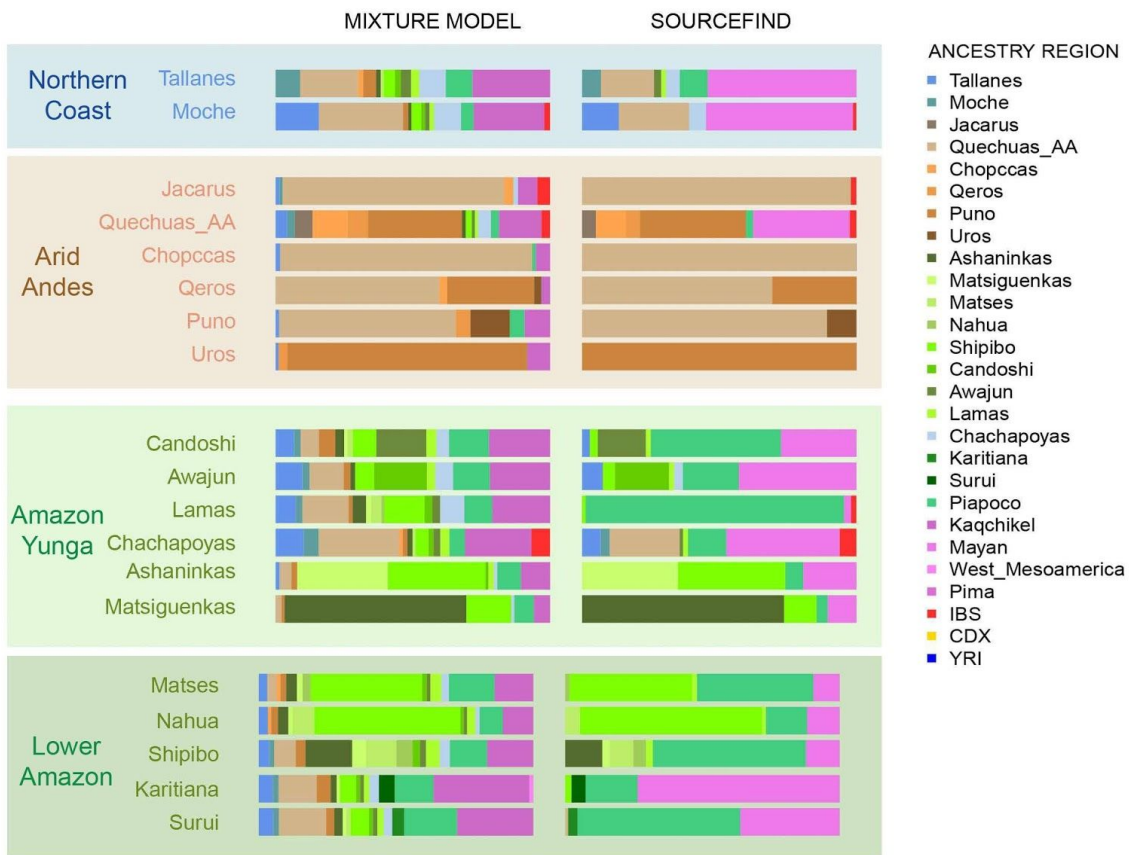


Figure S12. Proportions of haplotype sharing for each target population respect to Native Americans, Europeans and Africans donors populations, for the Dataset Natives 500K, inferred by two approaches: A non negative regression (MIXTURE MODEL) and a Bayesian approach (SOURCEFIND). Colored bars indicate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

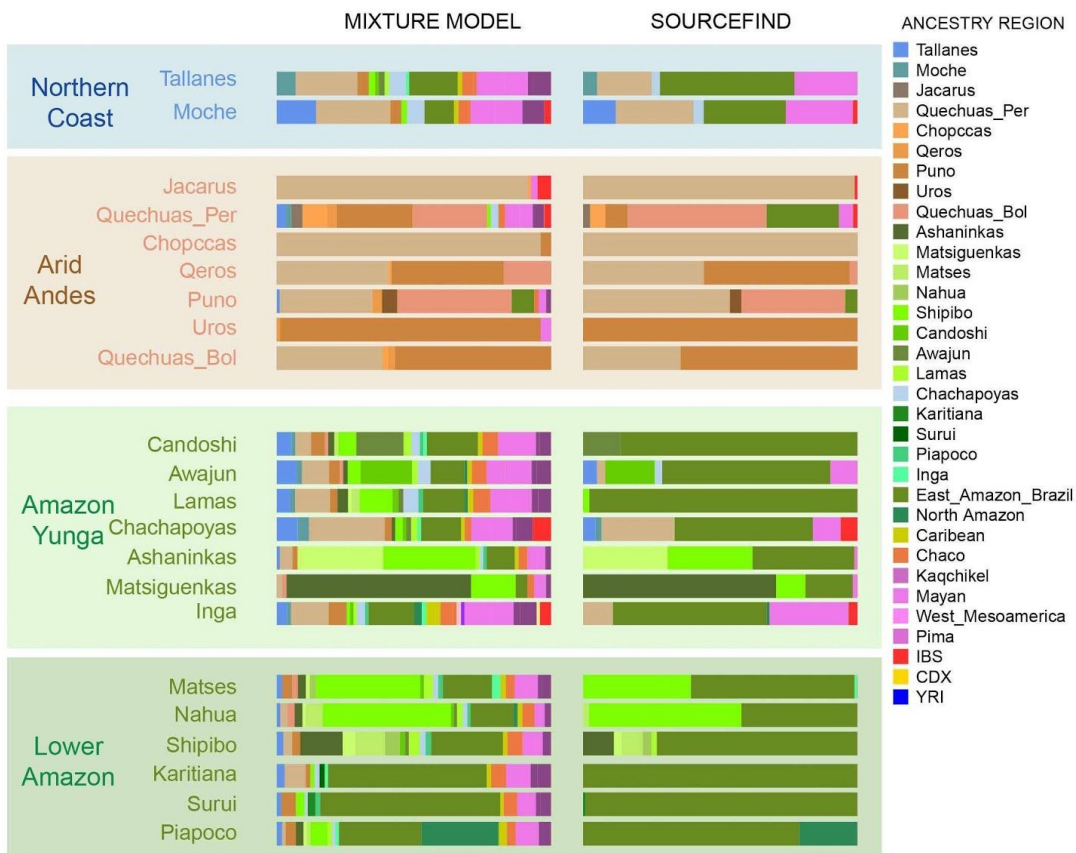


Figure S13. Proportions of haplotype sharing for each target population respect to Native Americans, Europeans and Africans donors populations, for the Dataset Natives 230K, inferred by two approaches: A non negative regression (MIXTURE MODEL) and a Bayesian approach (SOURCEFIND). Colored bars indicate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

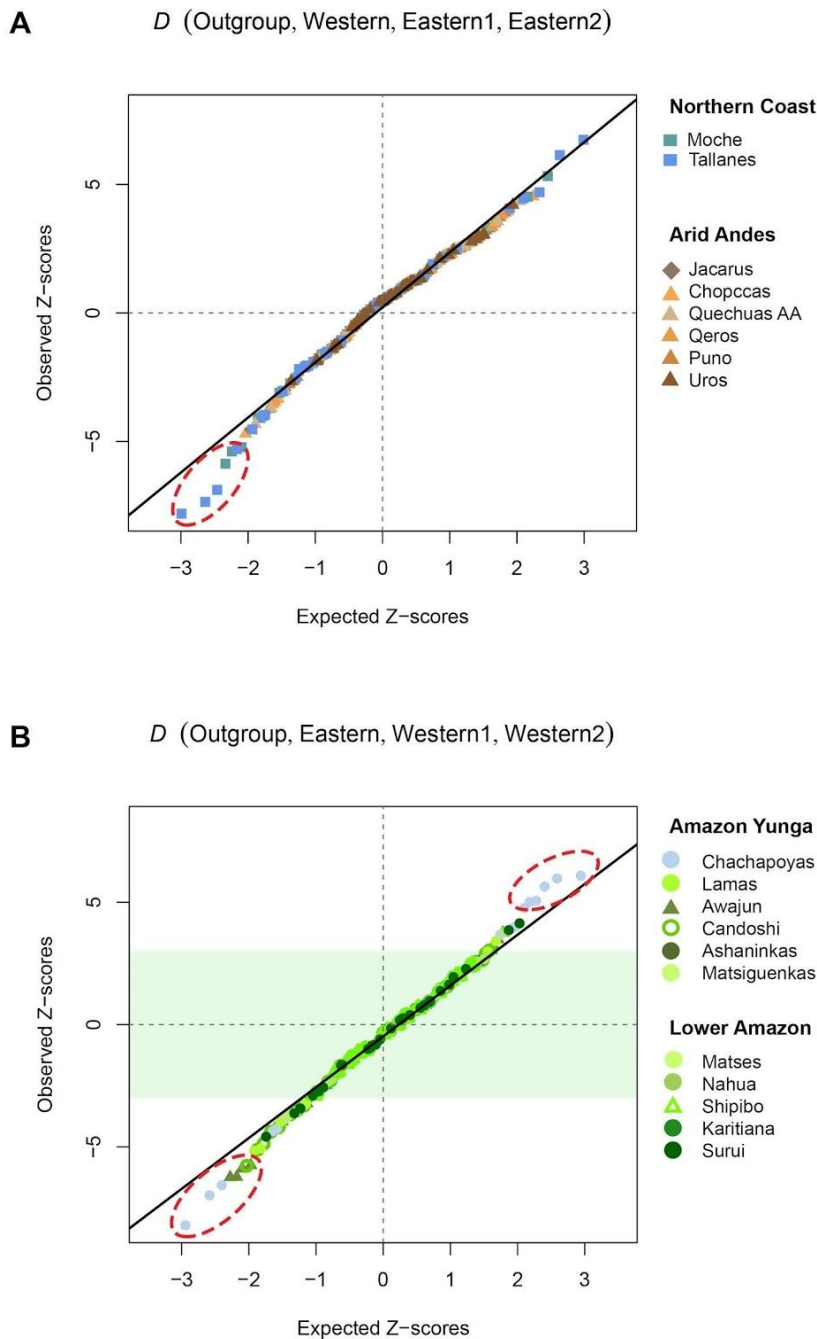


Figure S14. Quantile-quantile plot comparing Z-scores from D -statistics relating Western (Northern Coast and Arid Andes) and Eastern Andean slope (Amazon Yunga and Amazon Yunga) populations to those expected under a normal distribution (green diagonal) for the Dataset 500K. Red dashed circles show the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

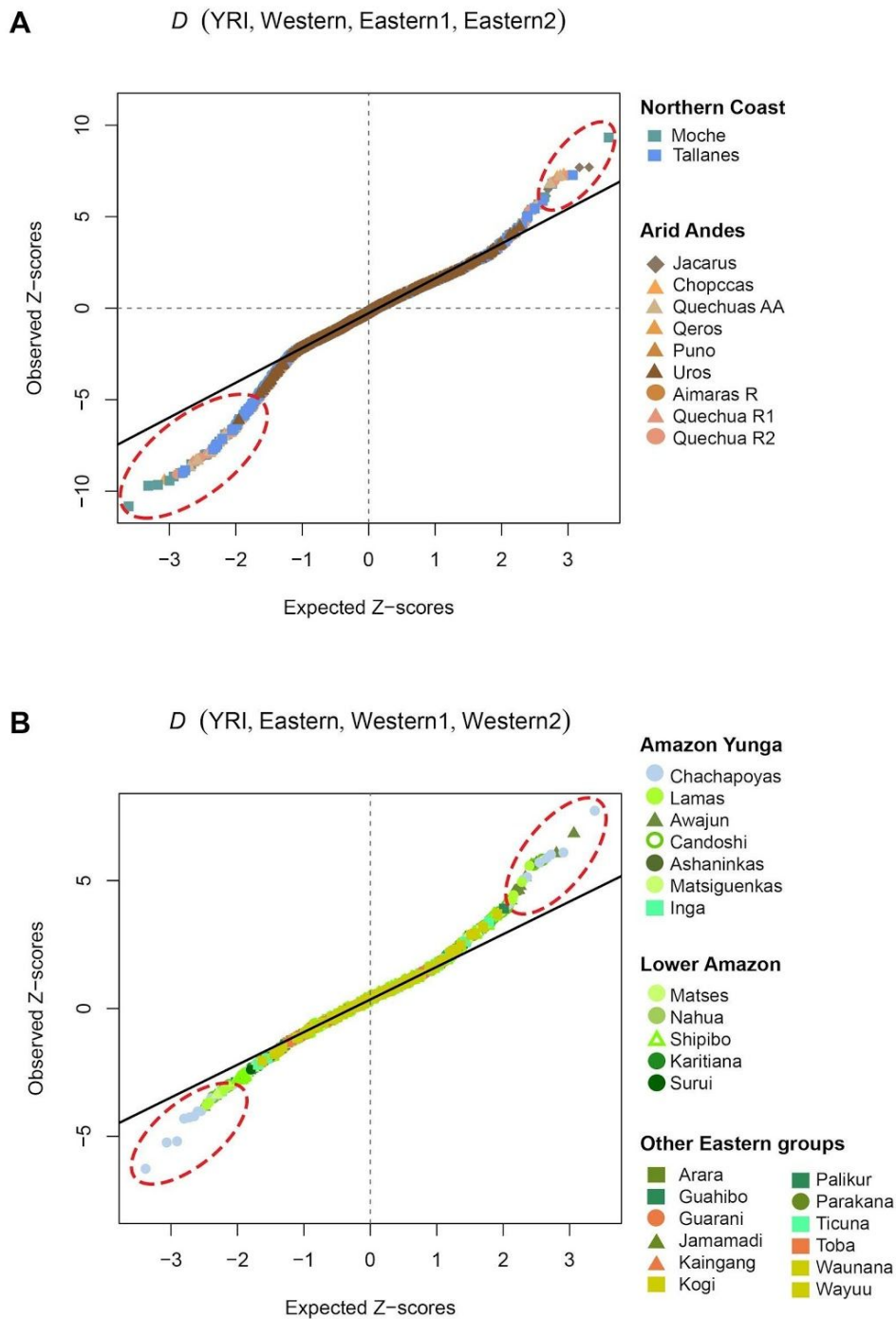


Figure S15. Quantile-quantile plot comparing Z-scores from D -statistics relating Western (Northern Coast and Arid Andes) and Eastern Andean slope populations (Amazon Yunga, Lower Amazon and other eastern groups) to those expected under a normal (green diagonal) distribution for the Dataset 230K. Red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

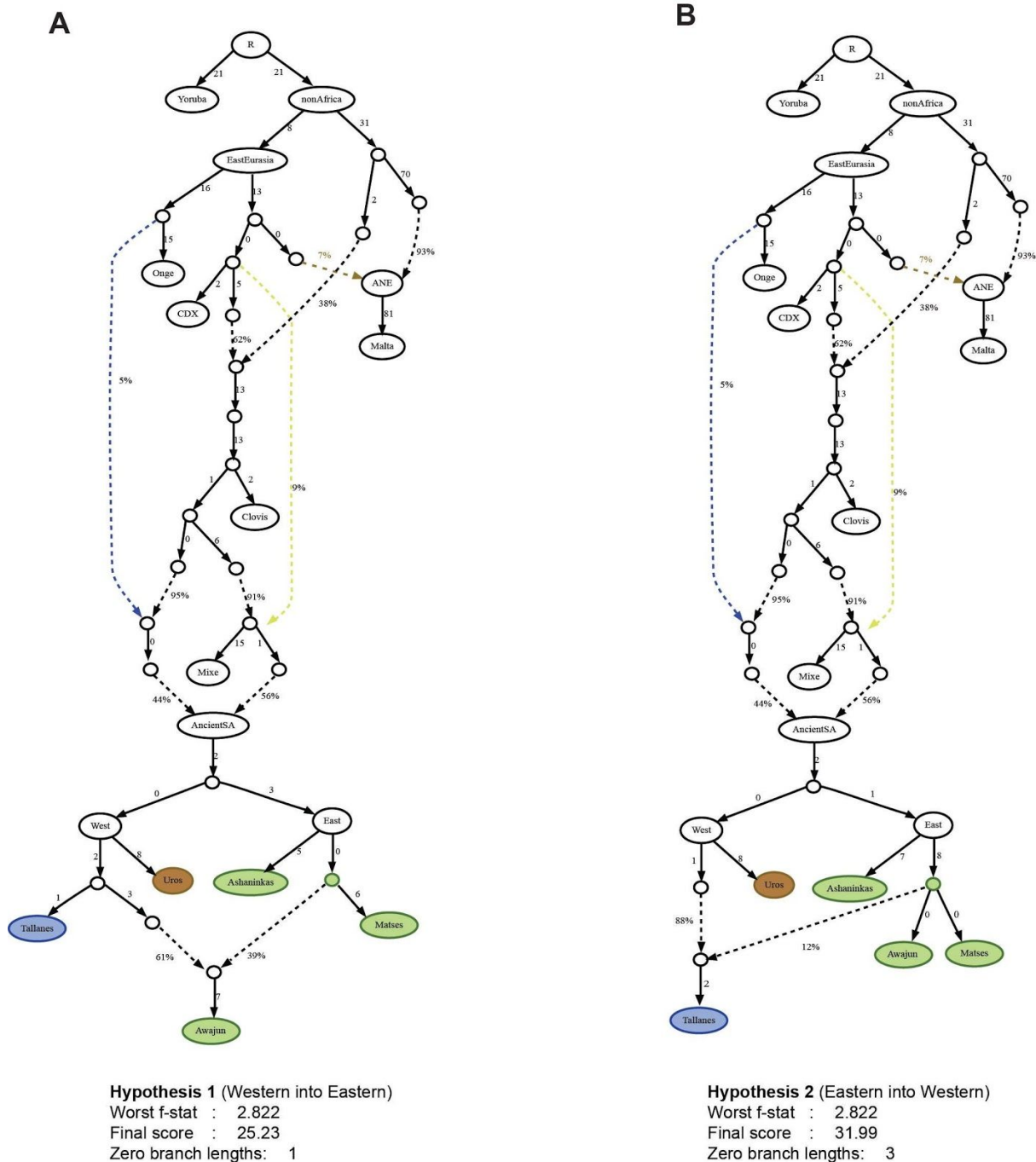


Figure S16. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes. We explore the relationship of the Tallanes-Awajun and the direction of the gene flow. White balls in the intermediate nodes represent hypothetical ancestors for each divergence event. **A)**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

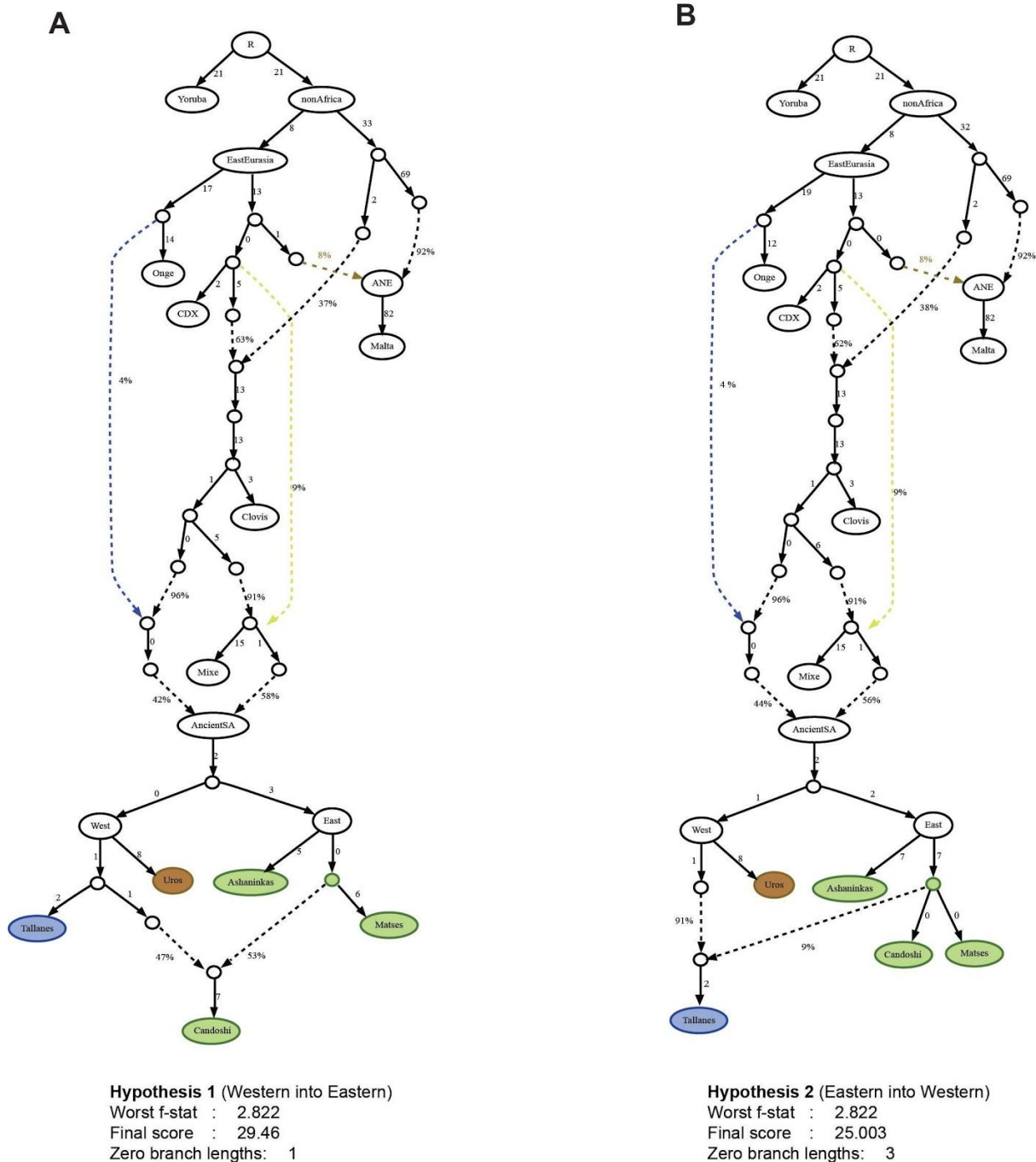


Figure S17. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes. We explore the relationship of the Tallanes-Candoshi and the direction of the gene flow. White balls in the intermediate nodes represent hypothetical ancestors for each divergence event. **A)**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

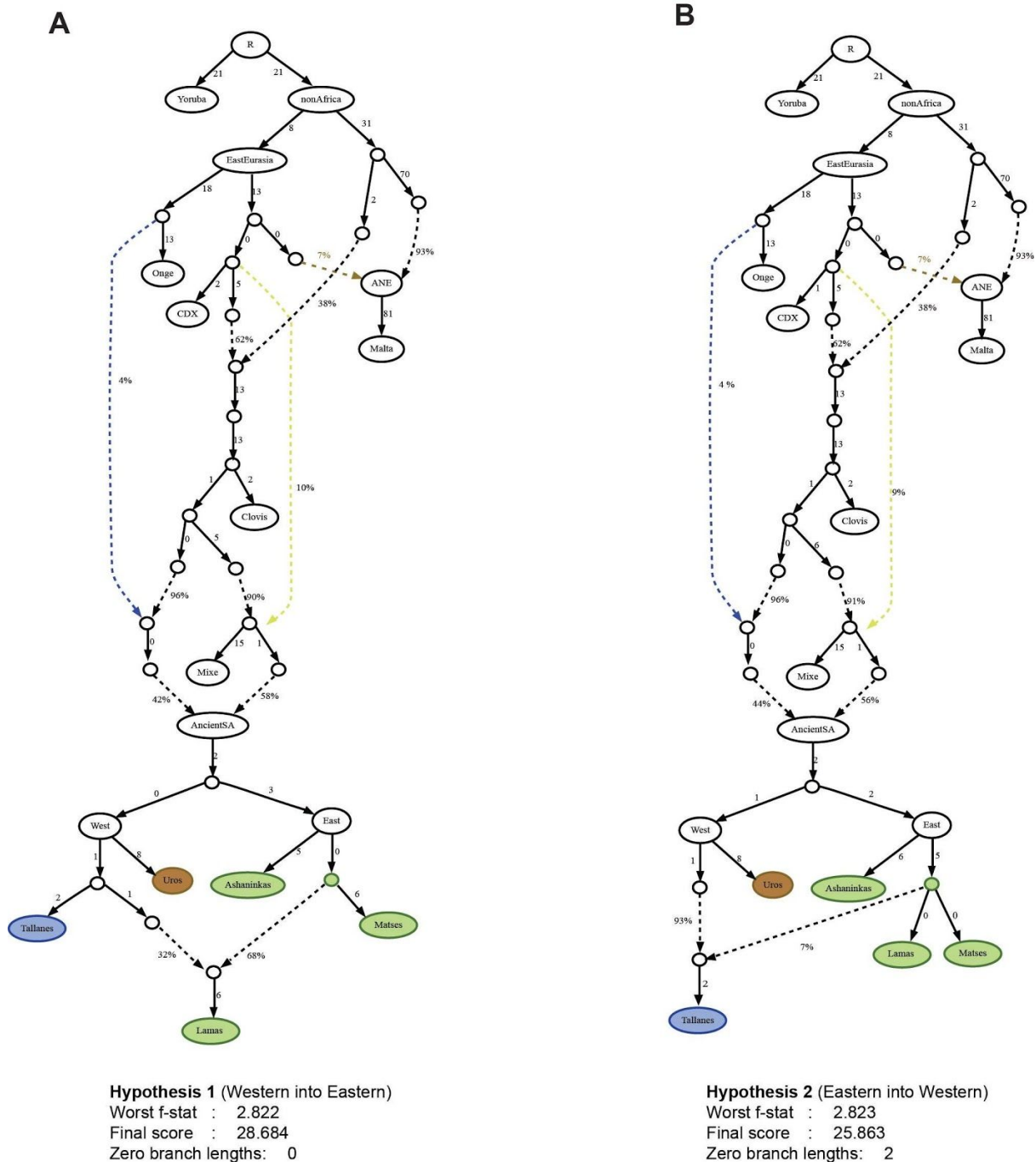


Figure S18. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes. We explore the relationship of the Tallanes-Lamas and the direction of the gene flow. White balls in the intermediate nodes represent hypothetical ancestors for each divergence event. **A)**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

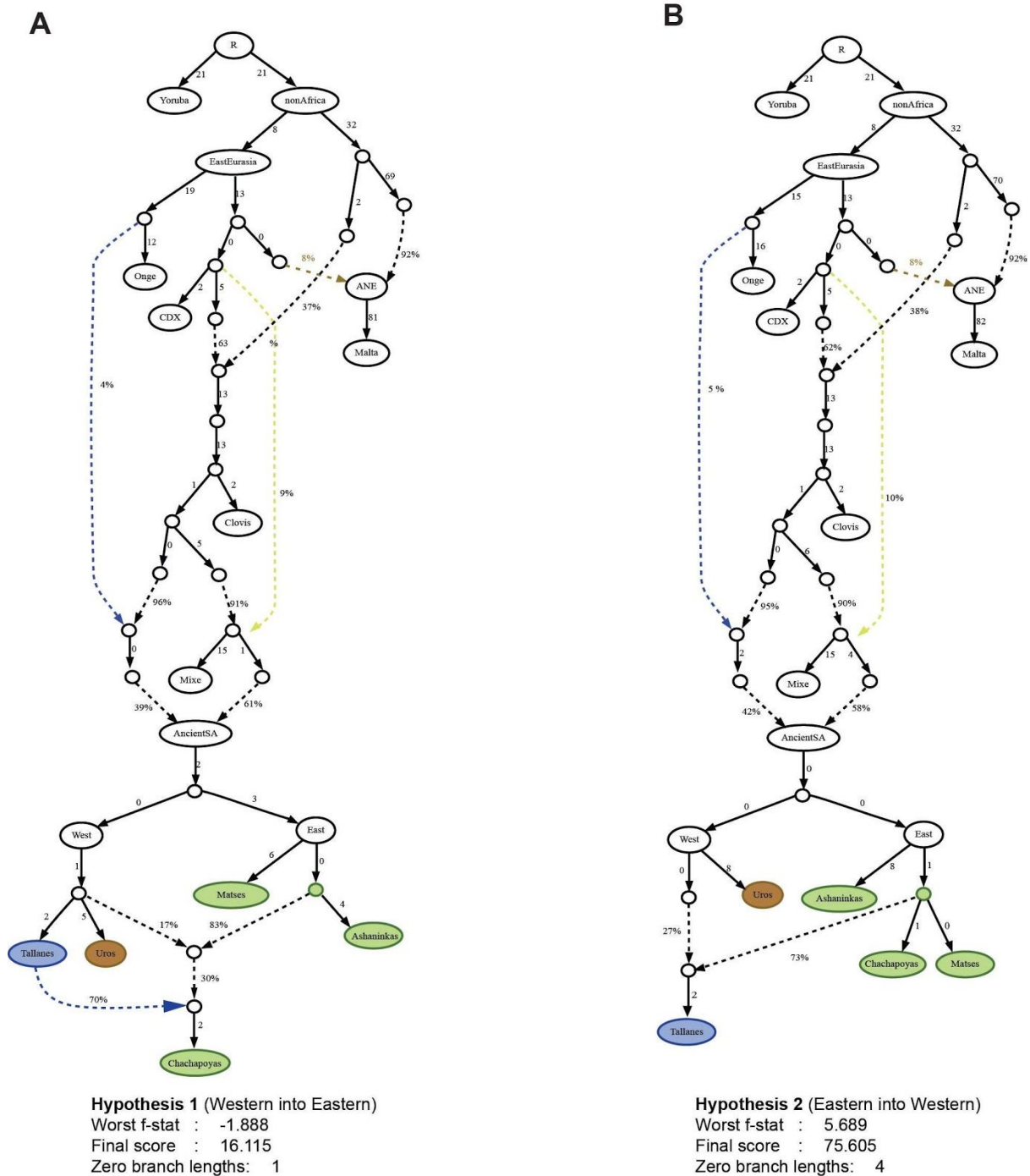


Figure S19. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes. We explore the relationship of the Tallanes-Chachapoyas and the direction of the gene flow. White balls in the intermediate nodes represent hypothetical ancestors for each divergence event.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

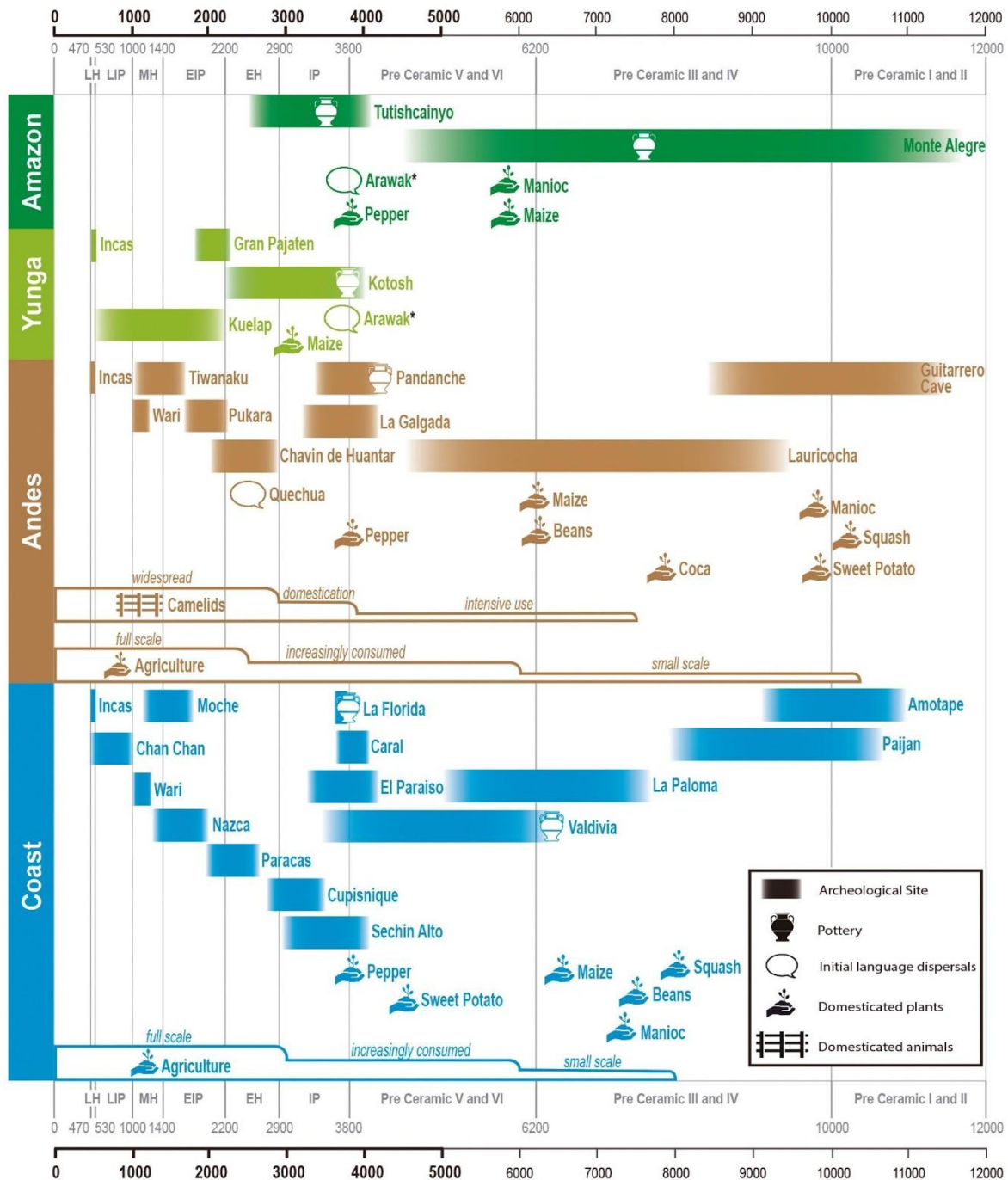


Figure S20. Key historical events of Peruvian prehistory in four longitudinal regions: Peruvian Coast, Andes, Amazon Yunga and Amazonia. Pottery and cultivars symbols represent the earliest archaeological record for the region. To account for time uncertainties, This figure showed the events in the chronology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

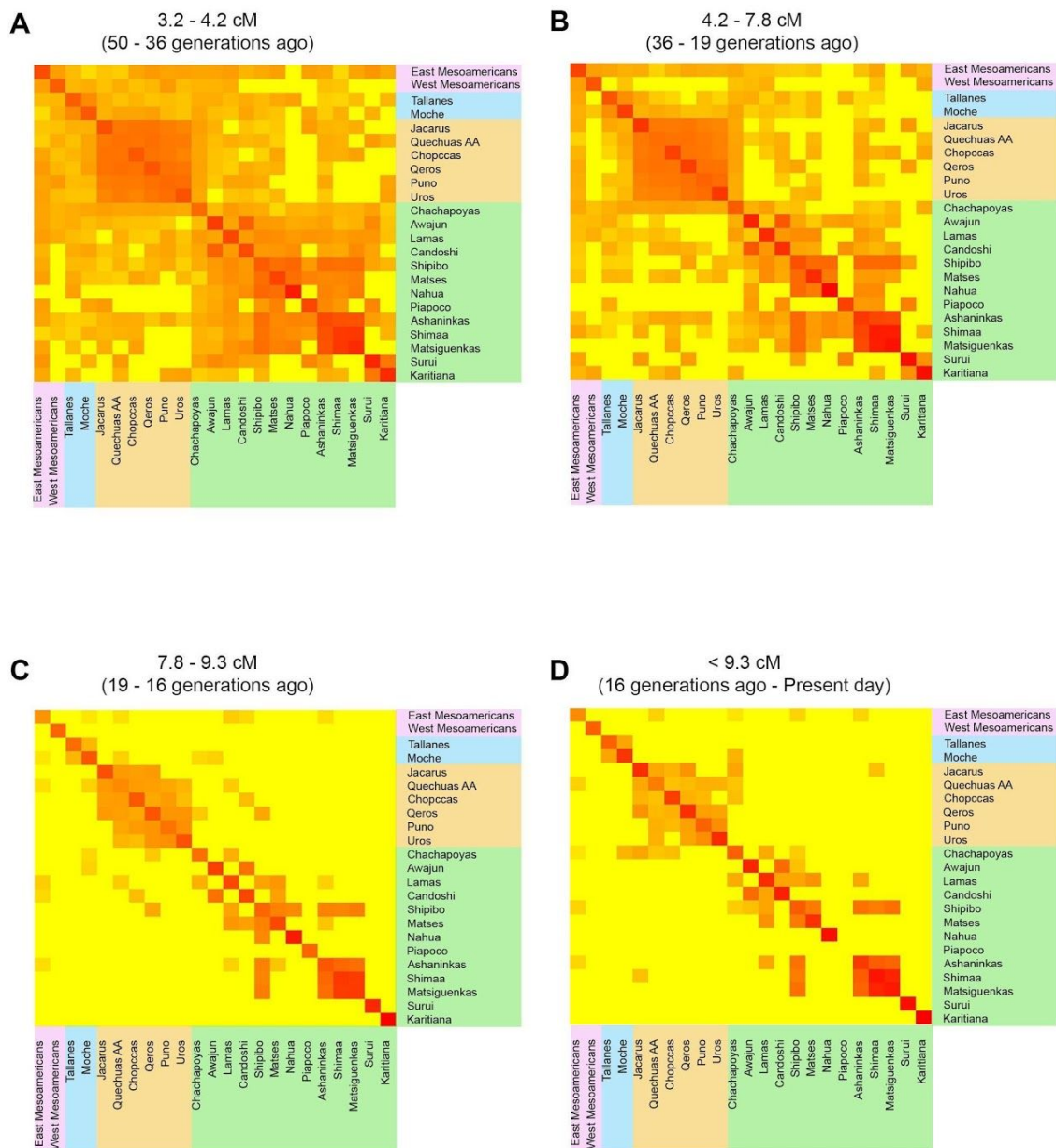


Figure S21. Heatmap representation of the shared Identical by descent (IBD) segments among Native Americans of the Natives 1.9M dataset. Each heatmap represents an interval of segments size and is correlated with time generation for the most common recent ancestor. A) An interval from 3.2 to 4.2 cM correlated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

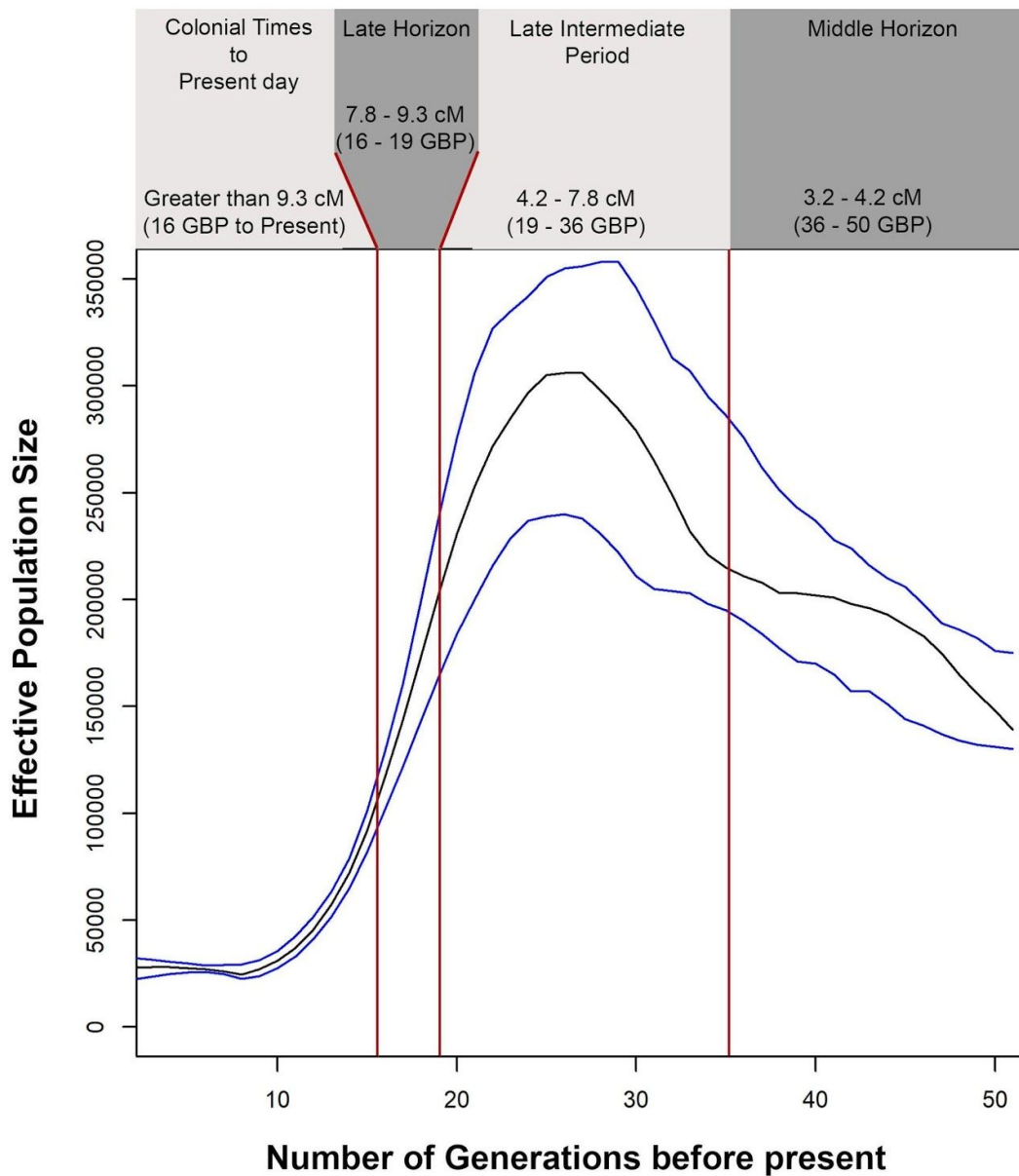


Figure S22. IBDNe analysis to infer the dynamic of the effective population size (N_e) from 4 generations ago to the last 50 generations for the Andean populations (Quechuas_AA, Aymaras_P, Chopccas, Qeros and Uros) as a whole. We used the Natives 1.9M dataset. The x axis represents the number of generations 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.

Demographic Model

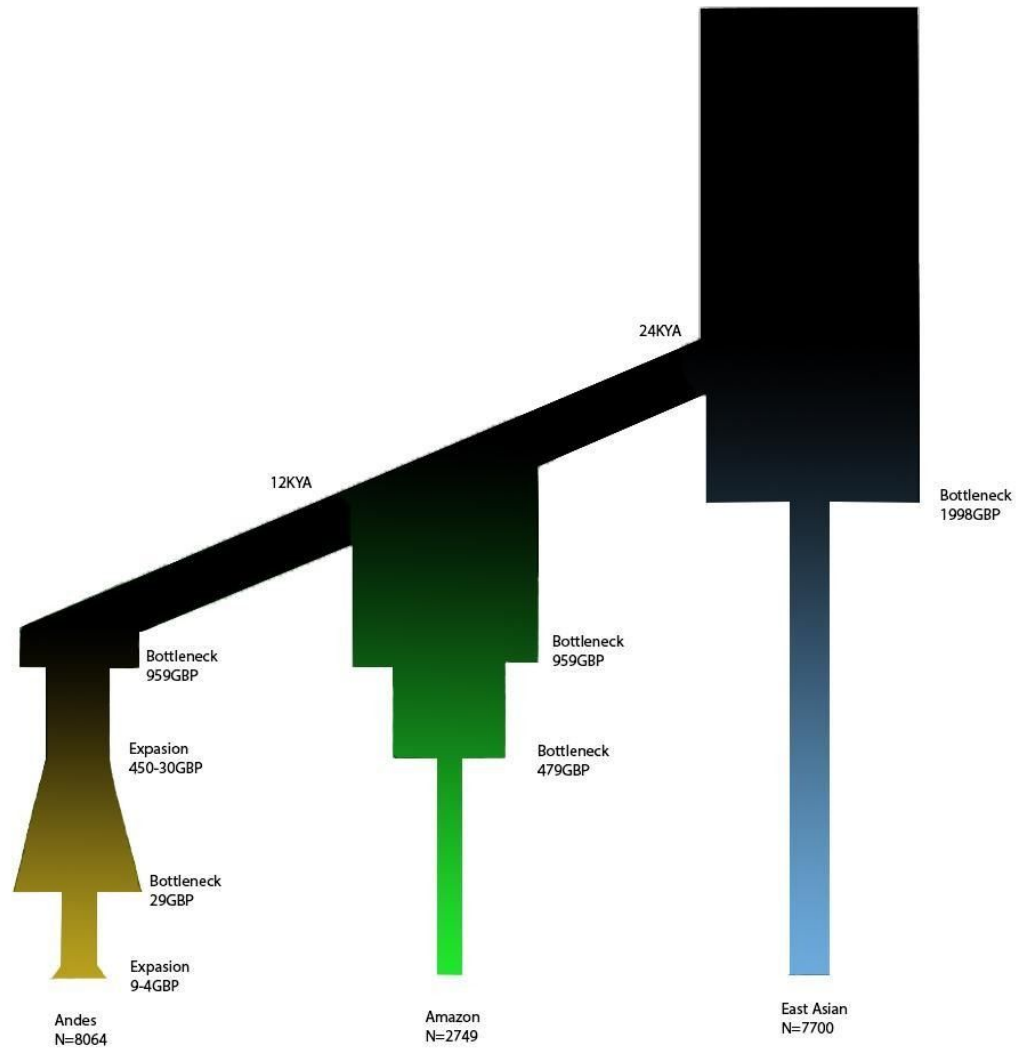


Figure S23. Demographic model of the Andean, Amazonian and East Asian populations. This model was used for the simulations made to calculate the p-value of the obtained PBSn values.

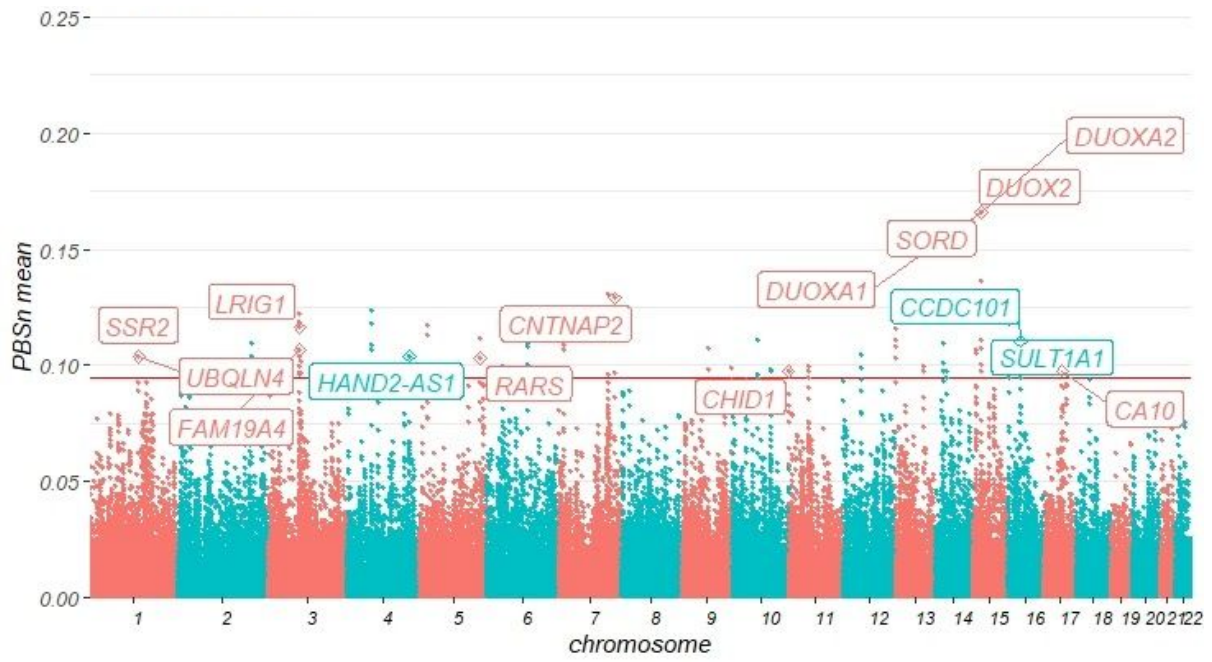


Figure S24. PBSn mean values Andean populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of PBSn mean (red line).

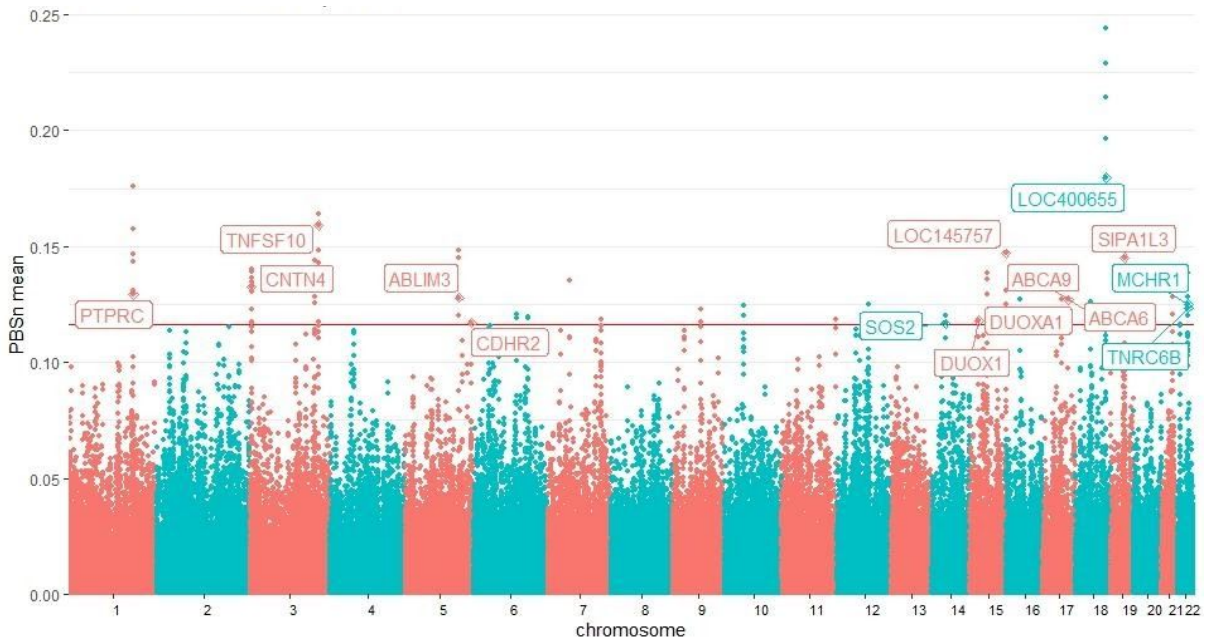


Figure S25. PBSn mean values Amazon populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of PBSn mean (red line).

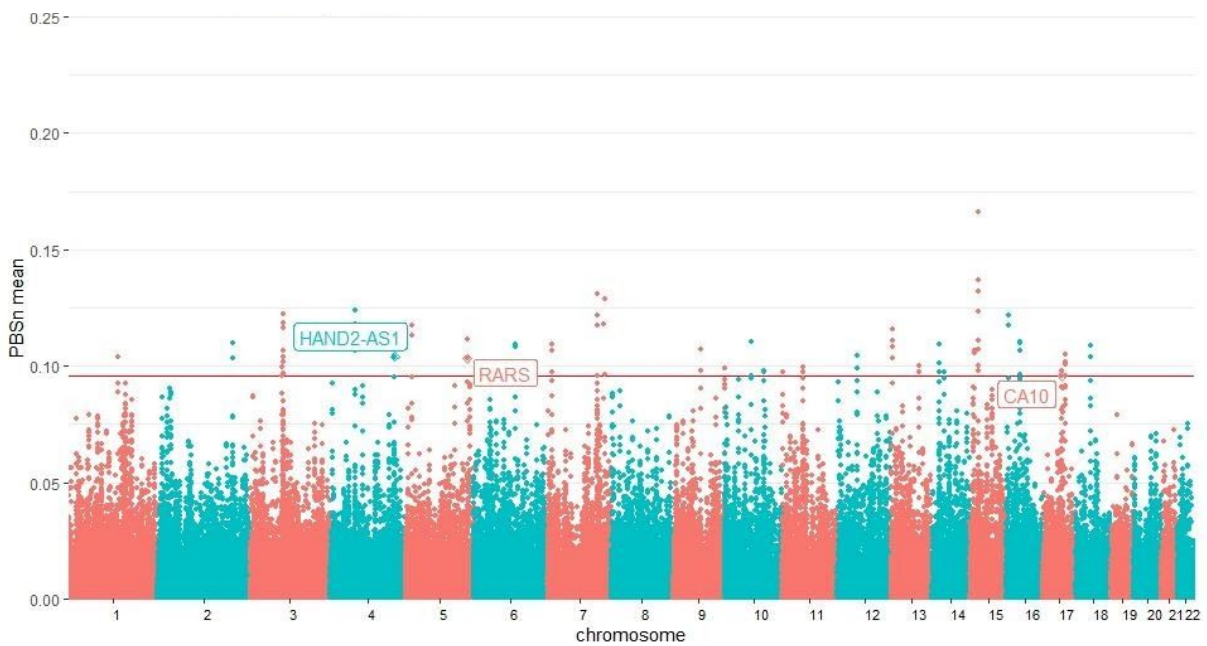


Figure S26. PBSn mean values for windows of 20 SNPs with five SNPs of overlap in Andean populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of windows PBSn mean (red line) that also present high values for xpEHH are labeled.

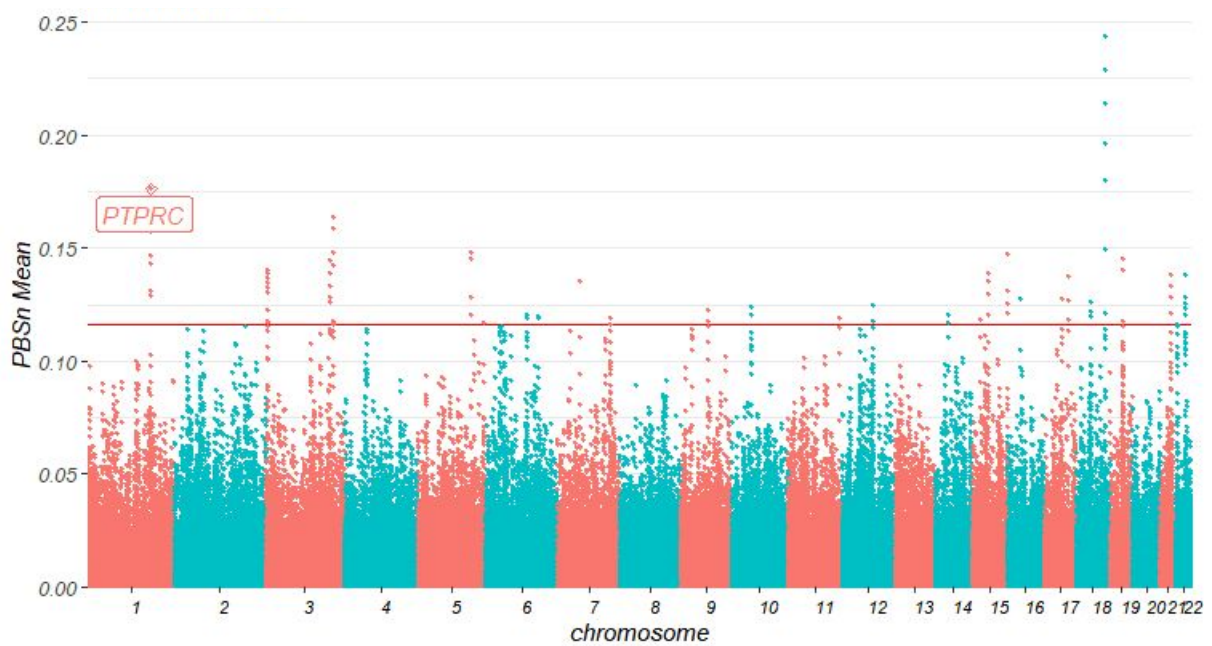


Figure S27. PBSn mean values for windows of 20 SNPs with 5 SNPs of overlap in Amazon populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of windows PBSn mean (red line) that also present high values for xpEHH are labeled.

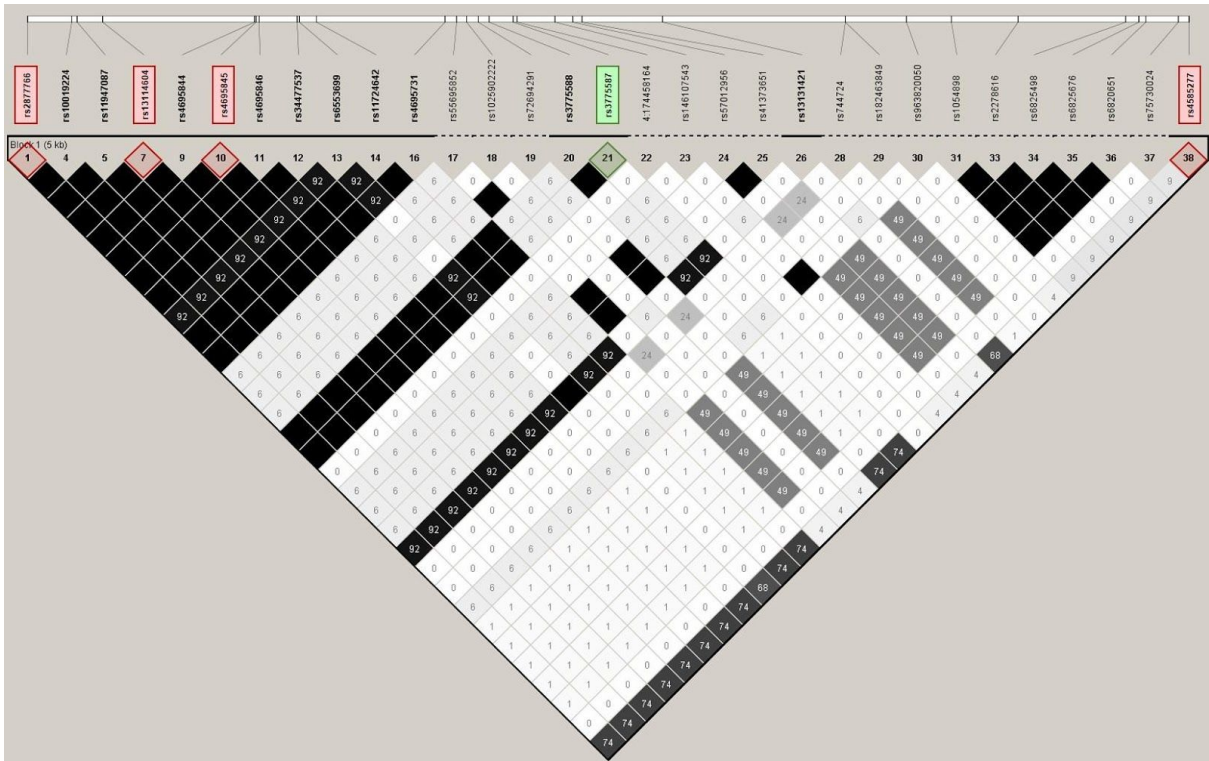


Figure S28. Linkage Disequilibrium between rs3775587 and the SNPs found to be under selection in the gene HAND2-AS1. SNPs with signals in PBS and xpEHH analysis are in red and SNP rs3775587, mapped within the putative enhancer GH04J173536 is in green.

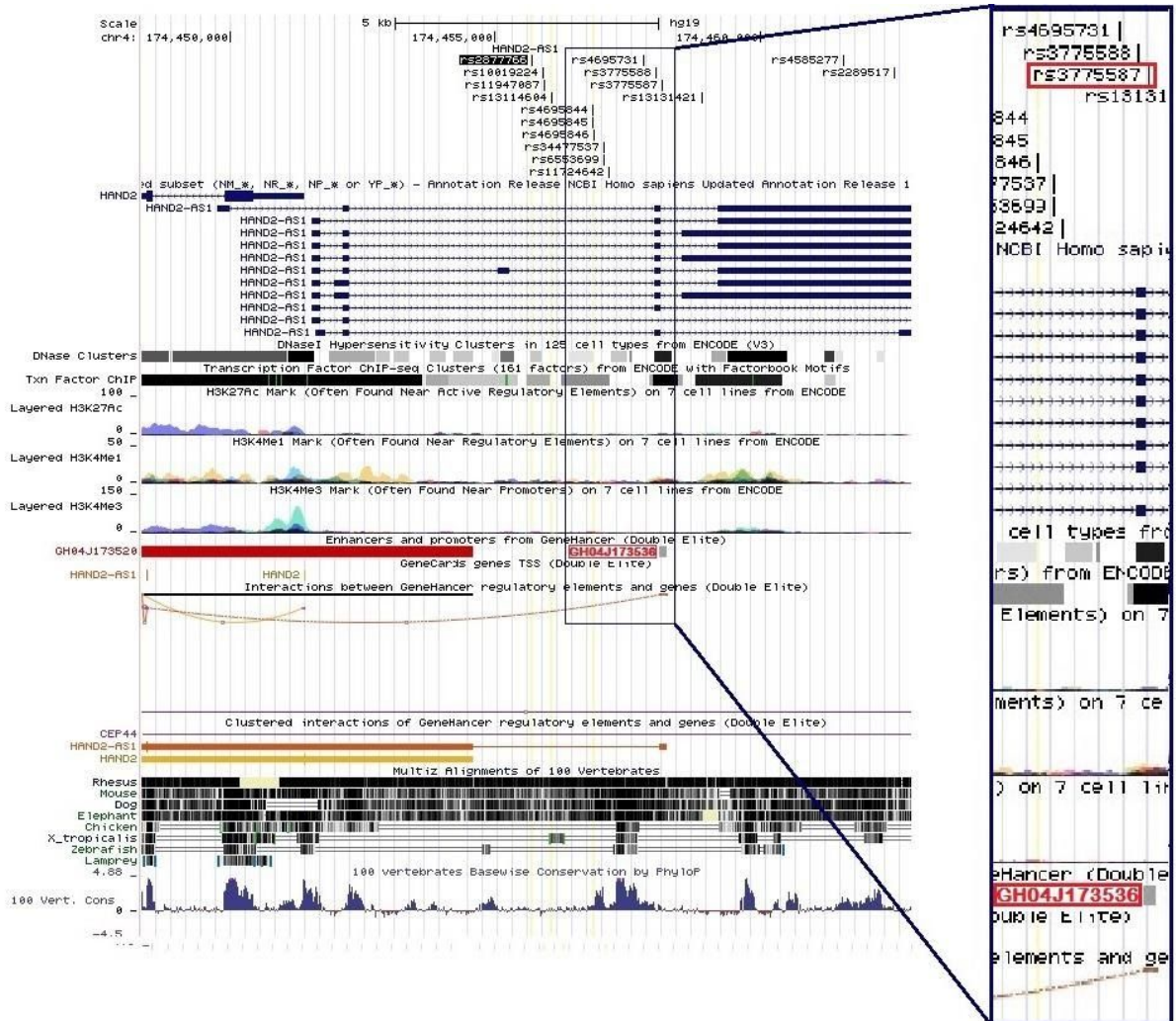


Figure S29. UCSC Genome Browser view of HAND2-AS1 locus with the SNPs located in regions found to be under selection, DNase I hypersensitivity clusters, Transcription Factor ChIP-seq binding sites, and the histone modifications H3K27ac (Often Found Near Active Regulatory Elements), H3K4me1 (Often Found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

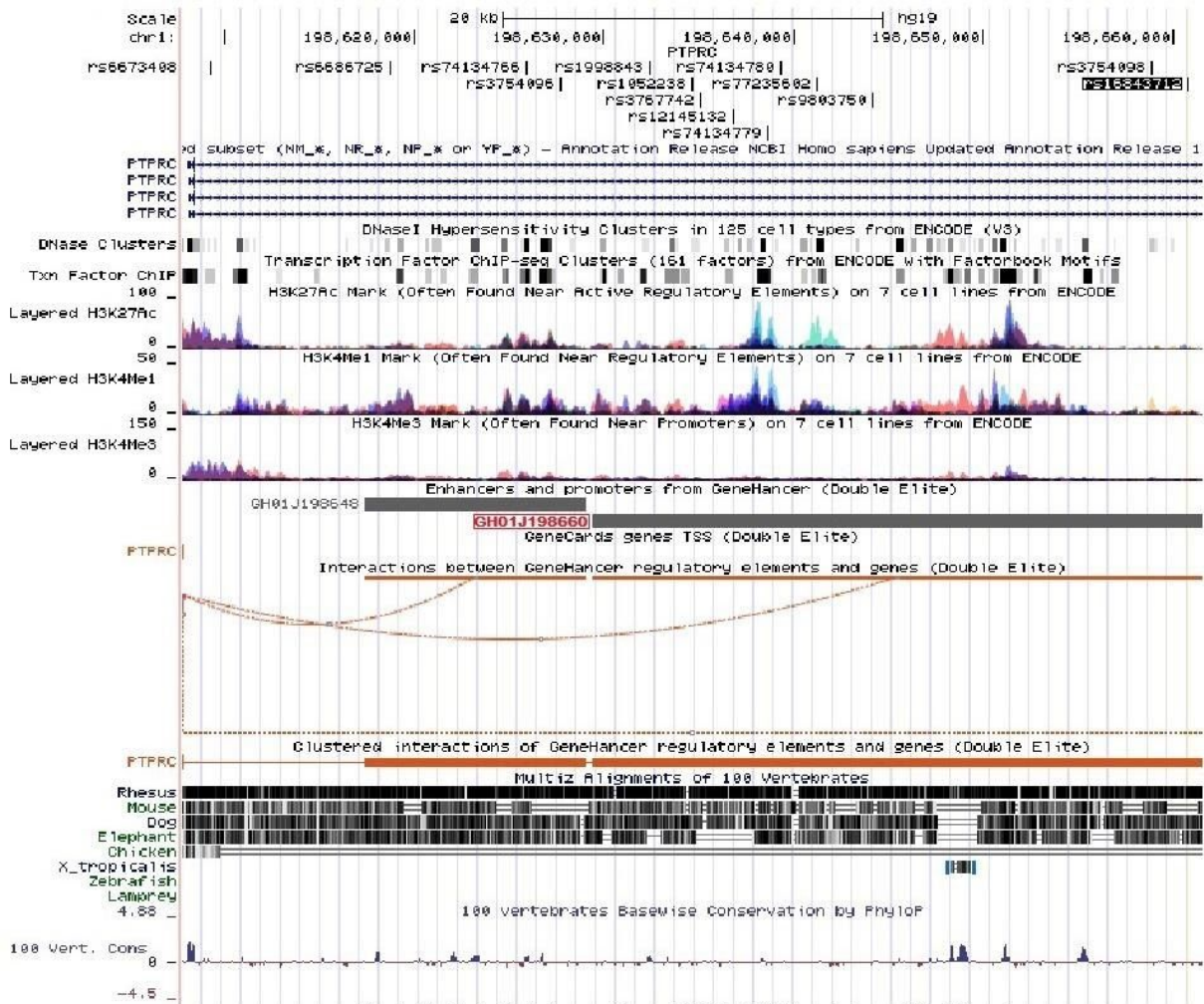


Figure S30. UCSC Genome Browser view of PTPRC locus with the SNPs located in regions found to be under selection, DNase I hypersensitivity clusters, Transcription Factor ChIP-seq binding sites, and the histone modifications H3K27ac (Often Found Near Active Regulatory Elements), H3K4me1 (Often Found Near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

Legends for Datasets S1 to S3

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

Dataset 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 Andean Cloud Forest Crossroads" in *The Handbook of South American Archaeology*, H. Silverman, W.

- H. Isbell, Eds. (Springer New York, 2008), pp. 903–926.
2. I. Schellerup, Wayko-Lamas: a Quechua community in the Selva Alta of North Peru under change. *Geografisk Tidsskrift*, 199–208 (1999).
 3. G. Seitz, *Cultural Discontinuity: The New Social Face of the Awajun* (Amakella Publishing, 2017).
 4. J. M. Guallart, *La tierra de los cinco ríos* (Pontificia Universidad Católica del Perú, Instituto Riva Agüero, 1997).
 5. L. Campbell, “Language isolates and their history” in *Language Isolates*, (Routledge, 2017), pp. 1–18.
 6. S. Purcell, *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
 7. W. C. S. Magalhães, *et al.*, EPIGEN-Brazil Initiative resources: a Latin American imputation panel and the Scientific Workflow. *Genome Res.* **28**, 1090–1095 (2018).
 8. A. L. Price, N. A. Zaitlen, D. Reich, N. Patterson, New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
 9. F. S. G. Kehdy, *et al.*, Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8696–8701 (2015).
 10. 1000 Genomes Project Consortium, *et al.*, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
 11. J. Z. Li, *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
 12. D. Reich, *et al.*, Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
 13. S. Mallick, *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
 14. M. Raghavan, *et al.*, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
 15. R. E. Green, *et al.*, A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
 16. N. Patterson, *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
 17. B. K. Maples, S. Gravel, E. E. Kenny, C. D. Bustamante, RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
 18. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

19. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
20. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
21. O. Delaneau, J. Marchini, J.-F. Zagury, A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
22. G. Hellenthal, *et al.*, A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
23. S. Leslie, *et al.*, The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
24. L. van Dorp, *et al.*, Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. *PLoS Genet.* **11**, e1005397 (2015).
25. J.-C. Chacón-Duque, *et al.*, Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat. Commun.* **9**, 5388 (2018).
26. G. A. 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).

Complementary Discussion

Some of the results of the analyses of natural selection and genetic differentiation were presented only in the supplementary tables of the article, so they will be discussed in more detail in this section.

Natural Selection Analyses

In the intersection between PBS and xpEHH analyses, other 2 to genes were pointed as strong candidates in Andean populations (Table S6, Figure S26): RARS (Arginyl-tRNA synthetase; PBSn=0.151 and p-value=0.01, xpEHH=2.98 and p-value<0.0025) and CA10 (carbonic anhydrase 10; PBSn=0.163 and p-value=0.008, xpEHH=3.09 and p-value<0.001). The Arginyl-TRNA Synthetase (RARS) is an essential protein for RNA translation that links amino acids to the appropriate tRNAs. RARS proteins also play a role in hypoxia resistance. Lori L. et al (2009)⁷¹ performed a genetic screen that identified the Arginyl-TRNA Synthetase gene of *C. elegans* (*rrt-1*) as a relevant gene implicated in the control of hypoxia sensitivity. In the same study, they demonstrated that the knockdown of Arginyl-TRNA Synthetase gene through RNAi before or after hypoxia injury prevented animal death. Further studies conducted in rats demonstrated significant alterations in Arginyl-TRNA Synthetase expression and activity in models of induced cerebral ⁷² and retinal ischemia ⁷³. Aminoacyl-tRNA synthetase genes were pointed as those with the most prominent altered expression in animals submitted to induced ischemia when comparing control groups to animals submitted to protective ischemic preconditioning ⁷³. These studies indicate that RARS protein is important to hypoxia response and resistance. Although the mechanisms underlying the role of RARS in the response to hypoxia are still not well established, variants that alter its expression may have been favored in high-altitude populations for helping to prevent cell injury due to the low O₂ concentration.

The CA10 is a gene highly conserved within vertebrates that encodes a protein member of the carbonic anhydrase family. Carbonic anhydrases catalyze the conversion of CO₂ to HCO₃⁻ and H⁺, participating in the pH regulation in several tissues and in the transport of ions across the membrane by transporter proteins. However, CA10 proteins are catalytically inactive ⁷⁴. This protein has remained many years since its discovery without an established function, mainly because it does not have the characteristic catalytic function of the carbonic anhydrase protein family. Recent studies have identified its role in the central

nervous system, where they are mainly expressed, facilitating the transport of neurexins in transsynaptic interaction networks ⁷⁵. The USCS Genome browser ⁷⁰ shows that SNPs rs16951028 (nPBS = 0.163 and p-value = 0.008, xpEHH = 3.097 and p-value = 0.002) and rs5010295 (nPBS = 0.160 and p-value = 0.008, xpEHH = 3.354 and p-value = 0.001) lie to a region with H3K4Me1 marks (often found near regulatory elements) (Attachment 1) in Human Lung Fibroblasts analysed for the the ENCODE project. However, according to GTEx Portal (ref) CA10 is not expressed in the lungs, and none of the SNPs around CA10 region under selection according to our analyses are eQTLs.

In a Genome-wide association study about asthma in Asian children, Perin and Potočnik ⁷⁶ identified a correlation between one variant in gene CA10 (rs967676-G, not present in our data) and greater pulmonary obstruction, as well as with a better response to glucocorticoid therapy. Other GWAS also claim that variants in this gene are related to increased risk of metabolic syndrome in African ancestry populations ⁷⁷, and osteoarthritis ⁷⁸ and osteoporosis ⁷⁹ in Asian populations. However, there is no stronger evidence on the role of CA10 in phenotypic characteristics whose function can be directly related to the individual's fitness. The lack of information makes it difficult to hypothesise on the possible reason why this allele has been selected in the Andes. This is an example of how the current high availability of genomic data, that allows us to explore the genome in many ways, can be limited by the lack of in vivo and in vitro studies that do not develop at the same pace. The same can be said for the long non-coding RNA on chromosome 18 (LOC400655) with extreme PBS values (nPBS=0.30 p-value=0.002) for the Amazon populations (discussed in the article), which has not yet been functionally characterized.

Genetic Differentiation Analyses

Regarding the genetic differentiation analyses, The annotation in PharmGKB database returned information on variant-drug associations for 6 highly differentiated SNPs ($F_{ST} > 0.318$) (Table S8). Of these, only one is classified in PharmGKB Clinical Annotation (level 3 - significant association in multiple studies but lacking clear evidence): the genotypes CC and CT for rs3114020 in gene ABCG2 (CC+CT frequency: Andes=0.74, Amazon=1) are associated with higher blood levels of the antiepileptic lamotrigine. According to the GTEx Portal ⁸⁰, this SNP is an eQTL for gene ABCG2 in whole blood samples. Although the associations regarding the variants in table S8 still need do be replicated, two of the genes on the list are important pharmacogenes with substantial evidence supporting other clinically

relevant polymorphisms ^{81,82}, ABCG2 and ABCB1 (also known as multidrug resistance protein 1 - MDRP1), both from the superfamily of ATP-binding cassette (ABC) transporters. A further investigation on these genes revealed 2 SNPs with PharmGKB Clinical Annotation level 2A (Annotation for a variant-drug association with moderate evidence where the variant is within a VIP - Very Important Pharmacogene) with F_{ST} values that do not pass our threshold (top 0.1%) but are within the top 0.5% higher values: rs2231142 (F_{ST} = 0.22, Allele T frequency: Andes=0.14, Amazon=0.47) in gene ABCG2, and rs1045642 (F_{ST} = 0.27, Allele T frequency: Andes=0.27, Amazon=0.67) in gene ABCB1 (Figure 1).

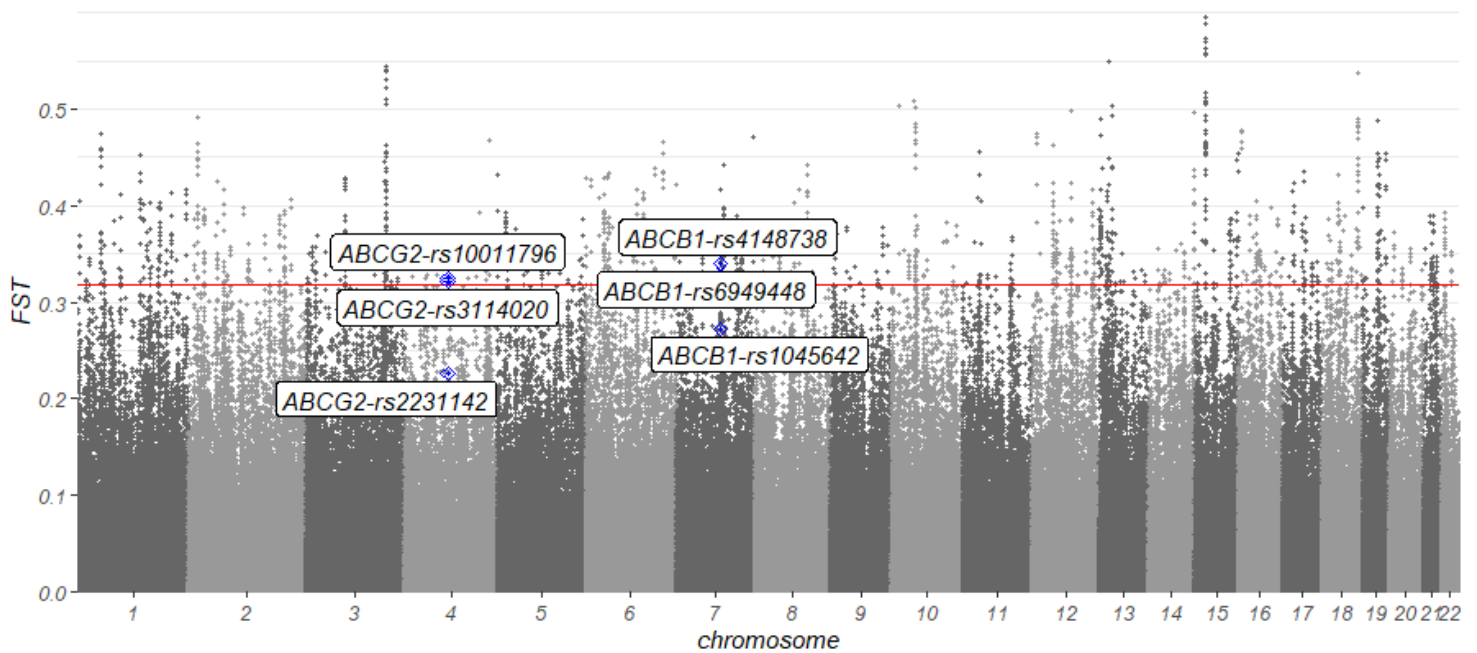


Figure 1. Manhattan plot showing F_{ST} values between Andean and Amazon populations. The red line delimits the 0,01% highest values ($F_{ST}>0.318$). Only SNPs undifferentiated within each group ($F_{sc}<0.15$) are shown. Labels indicate SNPs with high F_{ST} values (>0.22) within the pharmacogenes ABGG2 and ABCB1.

The variant rs2231142 (ABCG2) influences the metabolism of rosuvastatin, used to reduce cholesterol levels in patients with hypercholesterolemia, and allopurinol, used to treat gout reducing uric acid synthesis (PharmGKB Clinical Annotation level 2A). The variant rs2231142-T is associated with higher plasma concentrations of rosuvastatin and a better response regarding the reduction in LDL-C ⁸³⁻⁸⁷. The label of rosuvastatin from the company Swissmedic advertises that people with the genotype rs2231142-TT may have increased exposure (blood levels of the drug after the administration) to the medication ⁸⁸. On the other hand, the same allele is associated with worse response to allopurinol treatment ⁸⁹⁻⁹². This indicates that a larger fraction of this population may need a dose adjustment when treated

with rosuvastatin or allopurinol, which are medications regularly prescribed for hypercholesterolemia and gout^{93 94}.

The polymorphism rs1045642 (ABCB1) is in LD with the highly differentiated SNPs (Table S8) rs6949448 ($F_{ST} = 0.34$, Allele T frequency: Andes=0.24, Amazon=0.69) and rs4148738 ($F_{ST} = 0.34$, Allele C frequency: Andes=0.24, Amazon=0.69) in Andean ($r^2 > 0.69$) and Amazonian ($r^2 > 0.85$) populations (Attachment 2, Table 1). rs1045642 has been associated with the dosage, efficiency and toxicity of several drugs. The ones with higher levels of evidence (PharmGKB Clinical Annotation level 2A) are listed in table 2. The association with methotrexate is particularly important because it is used in the treatment of Acute Lymphoblastic Leukemia (ALL)⁹⁵. This is the most common cancer subtype in children and around 20% of them present severe toxicological reactions to methotrexate. ALL incidence is associated with Native American ancestry⁹⁶, as well as ALL relapse after treatment, in part due to pharmaco-alleles highly prevalent in most if not all Native American populations^{97,98}. A recent study showed that 9 markers in pharmacogenes linked to the metabolism of methotrexate and 6-mercaptopurine (also used in ALL treatment) have discrepant frequencies in Native Americans from the Brazilian Amazon compared to European, Asian, African and North American populations⁹⁸. The gene ABCB1 was not evaluated in this study that limited the analyses to variants with known functional consequences such as nucleotide changes or alternative splicing promoter regions. These data indicate that further studies to assess the diversity of these markers, including the ABCB1 gene, in different Native American populations are important to determining the treatment to Acute Lymphoblastic Leukemia in children with Native American ancestry. Moreover, differences in biomarkers frequencies in Native Americans are important not only with respect to other populations, but also between different Native American populations.

Table 1. Frequency of SNPs in LD in gene ABCB1. A1: minor allele, PEL: Peruvians from Lima, EAS: East asian populations, EUR: European populations, WAFR: West African populations.

CHR	SNP	Gene	F_{ST}	A1	Andes	Amazon	PEL	EAS	EUR	WAFR
7	rs4148738	ABCB1	0.340	T	0.76	0.307	0.706	0.594	0.565	0.838
7	rs6949448	ABCB1	0.339	C	0.76	0.313	0.682	0.609	0.584	0.842
7	rs1045642	ABCB1	0.272	C	0.7255	0.3311	0.6235	0.6022	0.4821	0.8543

Table 2. PharmGKB annotations level 2A* for SNP rs1045642 in gene ABCB1. Adapted from tables in ⁹⁹ and ¹⁰⁰. *Level 2A: Annotation for a variant-drug combination with moderate evidence of an association where the variant is within a VIP (Very Important Pharmacogene) as defined by PharmGKB.

Molecule	Effect	Phenotype	Pgkb drug sentence
Methotrexate	Toxicity/ADR	Burkitt Lymphoma, Drug Toxicity, Lymphoma, T-Cell, Precursor Cell Lymphoblastic Leukemia-Lymphoma, Toxic liver disease	Patients with the AA genotype and lymphoma or leukemia who are treated with methotrexate may have increased concentrations of the drug and may have an increased risk of toxicity as compared to patients with the GG genotype, although this is contradicted in some studies. Other genetic and clinical factors may also influence a patient's risk of methotrexate-induced toxicities.
Nevirapine	Toxicity/ADR	HIV Infections	Patients with the GG genotype and HIV-1 infection who are treated with nevirapine may have an increased risk for nevirapine hepatotoxicity as compared to patients with the AA genotype. Other genetic and clinical factors may also influence a patient's risk for hepatotoxicity with nevirapine treatment.
Ondansetron	Efficacy	Nausea and vomiting	Patients with genotype GG may have increased likelihood of nausea and vomiting shortly after being treated with ondansetron as compared to patients with genotype AA. Other genetic and clinical factors may also influence a patient's response to ondansetron.
Digoxin	Other	-	Patients with GG genotype may have increased metabolism and decreased serum concentration of digoxin as compared to patients with the AA genotype. Other genetic and clinical factors may also impact the metabolism of digoxin.
Fentanyl	Dosage	PainPain, Postoperative	Patients with the GG genotype may have increased fentanyl opioid dose requirements as compared to patients with the AA or AG genotypes. However, one study did not find an association between this variant and fentanyl dosing. Other genetic and clinical factors may also affect a patient's fentanyl dose requirements.

Perspectives

Genetic differentiation in Native South Americans

Population genetics studies are more frequently conducted in populations of European descent, leaving other ethnic groups underrepresented. Considering the diversity in genetic structure among different populations, this unbalance in representation creates a bias when applying new discoveries for these underrepresented groups ^{26,101} as illustrated in the previous section. The large dataset presented in Borda et al. (2020) ² represents a great opportunity to evaluate how the biomedical discoveries made in other ethnic groups affect Native American populations. With that in mind, we are developing an analysis of genetic differentiation based on F-statistics, similar to that held between Andean and Amazonian groups, to determine highly differentiated variants between Native Americans and populations that are better represented in genetic studies (of European and Asian ancestry) ²⁶. In this way, we will be

able to identify genetic regions previously reported in biomedical studies and analyse the possible impacts of frequency differences for Native American populations. In this context, the Native American frequencies from the database described in the article ² were incorporated to the software DANCE ⁶⁹. This tool integrates information about complex traits and genetic variability with a network-based approach, making it easier to interpret and contextualize the data.

Functional characterization of natural selection candidate variants

Functional studies to verify the effect of natural selection candidate variants found in genetic analyses are often left behind due to the high cost and the need of robust candidate targets, as well as the need to use the appropriate cells with the appropriate genotypes. This limits the knowledge about the process of adaptation to inferences based on functions of a few well known genes, and leaves the contribution of several signals in less studied genes and noncoding regions unexplained. In order to evaluate the functional effects of the variants for which we found signs of selection, I wrote a project in cooperation with Dr. Irene Gallego Romero (specialist in functional human genomics studies) at the University of Melbourne, Australia, that was contemplated in the CAPES-PRINT program. The aim of the project was to evaluate the levels of expression of genes under selection in the Andes in cell cultures subjected to different concentrations of oxygen. In Dr. Irene's laboratory I had access to cells collected from individuals in Lima, Peru, in the context of the 1000 Genomes Project. I identified cell lines with the selected and alternative genotypes for the DUOX2 and RARS genes and submitted them to Oxygen concentrations of 2.5%, 5% and 20% (atmospheric concentration) for 4hrs, 24hrs and 48hrs. Unfortunately, due to the COVID-19 pandemic, the University of Melbourne was closed and I had to return to Brazil before starting RNA extraction for the analysis of gene expression. I am in contact with Dr. Irene to define a strategy for the experiment to be completed when possible.

Continuing the initiative to functionally characterize the selection signals, we wrote a project submitted to the Collaborative Research Program (CRP) - ICGEB (International Center for Genetic Engineering and Biotechnology) and to the Leakey Foundation Research Grants, with the following objectives:

- (1) DUOX2 encodes a trans-membrane component of an NADPH oxidase, which produces the hydrogen peroxide (H₂O₂) essential for the synthesis of the thyroid hormone ^{102,103} and for the production and the microbicidal hypothiocyanite anion

(OSCN-) during mucosal innate immunity response against bacterial and viral infections in the airways^{103,104}. The non-synonymous SNP rs269868 (C>T, Ser1067Leu) is located in a region of the molecule that interacts with its coactivator DUOX2, and has two alleles rs269868-C (common and positively selected in the Andes [frequency: 0.53] and rare in the Amazon [frequency: 0.01]) and rs269868-T (common in the Amazon). To test the hypothesis that alternative alleles of DUOX2 are associated with functional differences, we will perform CRISPR/cas9 gene editing on human primary thyroid follicular epithelial cells (Nthy-ori 3-1 - Sigma-Aldrich), that express DUOX2, to create two sub-lineages, rs269868-CC (positively selected in the Andes) and rs269868-TT, and compare the level of gene and protein expression and production of H₂O₂ between the two lineages.

(2) The haplotype in gene PTPRC that flanks the rs16843712-A allele (frequencies: Amazon: 0.81, Andes: 0.32), identified as under selection in Amazon populations in the xpEHH analyses, lies within the putative human intronic enhancer GH01J198660 (in sensu Genehancer). It comprises the A (Thr193) allele of the non-synonymous SNP rs4915154 (A>G: Thr193Ala) that affect alternative splicing and alter a potential O- and N-linked glycosylation site. The positively selected allele A (Thr193) has been associated in functional studies¹⁰⁵ 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. Because the functional effect of rs4915154 has been already established, we will perform a gene reporter assay to verify if the alternative haplotypes (positively selected in the Amazon vs. alternative haplotypes) that comprise the putative enhancer GH01J198660 are also associated with different transcription rates.

Chapter 3 - Collaborations in other projects on American populations

Introduction

During my PhD I had the opportunity to participate in other projects of the LDGH group. In this process I was able to apply the knowledge acquired during the development of my main project (described in the previous chapters) to new datasets with different objectives, and to acquire new skills that contributed to my training in population genetics. In this chapter I will describe my contributions in three articles on the following topics: the genetic history of the African diaspora on the American continent ³⁰; the detection of variants associated with BMI through an admixture mapping analysis in Brazilian populations ³¹; and the genetic variability of genes related to SARS in Native South Americans ³².

The first two articles include data from the EPIGEN-Brasil Initiative ¹⁰⁶, a project by the Ministry of Health that aims to study the genomic diversity of Brazilian populations and their effects on complex diseases. Socioeconomic data were collected and 2.5 million SNPs were genotyped from 6,487 Brazilians from the three largest population-based cohorts in the country, representing three regions: Salvador - BA (Northeast), Bambuí - MG (Southeast) and Pelotas - RS (South). These data allowed not only the advancement of epidemiological genetic studies in Brazil, but also the study of the history of these admixed populations. One of the several products of the EPIGEN initiative, that includes many research groups around the country, is the analysis of an admixture mapping for BMI performed in these cohorts, reported in an article published in the International Journal of Obesity ³¹. The EPIGEN data was also combined with those of 23 other populations to conduct a study on the origins of American populations in the context of the African diaspora, published in Molecular Biology and Evolution ³⁰.

The last article presented here analyzes data from South American Natives in the current context of the COVID-19 pandemic. Even with the great current increase in the availability of genomic data, non-European populations are still underrepresented in genetic studies. In the midst of a pandemic, the lack of knowledge about genetic diversity in pathways related to the disease is a gap that harms the populations that are left out, but can also affect global initiatives to combat the virus. In an effort to contribute to this issue, we

conducted analyses on the genetic diversity of Native South Americans regarding genes important to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. This article was preprinted in Authorea ³² and submitted to Genetics and Molecular Biology.

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

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

Mateus H Gouveia, Victor Borda, Thiago P Leal, Rennan G Moreira, Andrew W Bergen, Fernanda S G Kehdy, Isabela Alvim, Marla M Aquino, Gilderlanio S Araujo, Nathalia M Araujo ... Show more

Molecular Biology and Evolution, Volume 37, Issue 6, June 2020, Pages 1647–1656,

<https://doi.org/10.1093/molbev/msaa033>

Published: 03 March 2020

“ Cite 🔑 Permissions ➦ Share ▼

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 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.

Keywords: African diaspora, Transatlantic Slave Trade, admixture dynamics, *mestizaje*

Issue Section: Discoveries

Associate Editor: Rasmus Nielsen

This paper is focused on three main topics: (i) the relation between the origin and destination of African populations; (ii) the association between the admixture process in the Americas and the dynamics of arrivals of Africans in the continent; and (iii) the between-populations genetic differentiation regarding the African gene pool in the Americas. I participated in the analyses on the third topic by calculating the genetic differentiation (F_{ST}^{67}) between African populations and between American populations shown in Figure 2C (figure 3C in the original paper). The purpose of this analysis was to evaluate the diversity of variants with African origin between American populations and compare it to the diversity between African populations. For this, we selected SNPs with >90% probability a posteriori for African ancestry (estimated by the RFMix method ¹⁰⁷). To avoid problems related to sample size, only the SNPs present in at least 20 haplotypes from each of the American populations were included in the analysis (21,078 SNPs). The F_{ST} for each SNP was calculated between African populations and between American populations with the following formula (ref):

$$F_{st} = \frac{\text{var}(p)}{\bar{p}(1 - \bar{p})}$$

Where the p is a vector with the minor-allele frequencies of a SNP i for all the considered populations, $\text{var}(p)$ is p variance, and \bar{p} is the average of p .

The comparison of the F_{ST} distributions for populations from each continent (Kolmogorov-Smirnov $D = 0.30$ and $p\text{-value} < 10^{-16}$) shows that the average differentiation between American populations ($F_{ST} = 0.02$) is two-thirds of the average between African populations ($F_{ST} = 0.03$). This result corroborates with the analysis of African-Specific Genetic Distance (ASGD) that shows a greater genetic distance between African populations (mean: 0.057), followed by the African populations in relation to the American ones (mean: 0.043), and a smaller distance between the American populations (mean : 0.018, 32% of the ASGD between African populations) (Figure 2A). Regarding diversity within populations, the measurement of average heterozygosity for SNPs in African fragments reveals similar patterns for populations on both continents (Figure 2C). Together, these results indicate that, although the intra-population diversity in African genomic fragments of American populations is similar to that seen in the ancestral populations, there was a homogenization of fragments from different African ancestral populations in America represented by the lower level of differentiation between these populations.

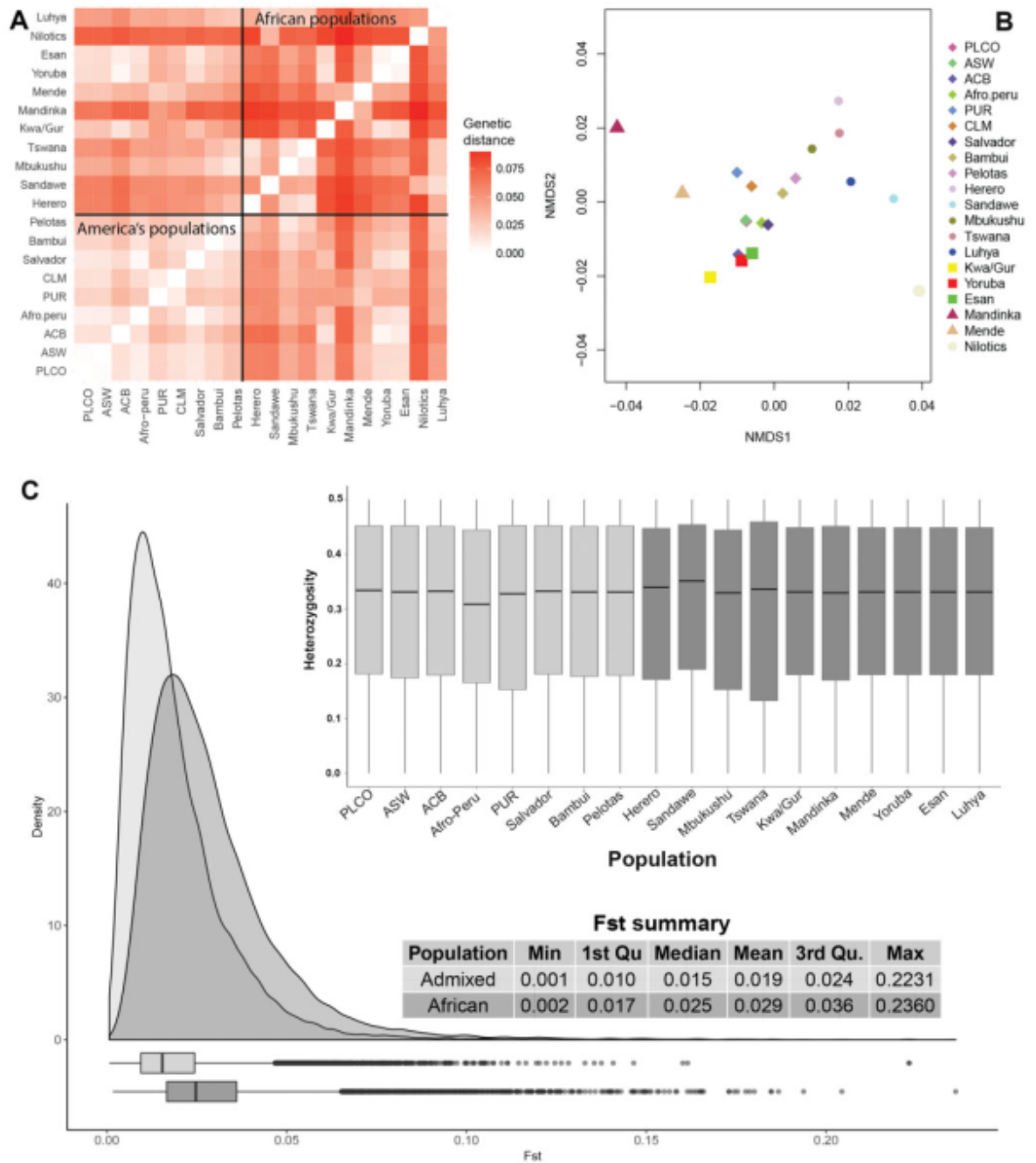


Figure 2. Pairwise genetic distances of the African gene pool between populations of the American continent and Africa. (A) Heatmap Matrix and (B) multidimensional scaling of the African gene pool genetic distances. We used solid squares, triangles, and circles to represent populations associated with 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.

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

Abstract

Admixed populations are a resource to study the global genetic architecture of complex phenotypes, which is critical, considering that non-European populations are severely under-represented in genomic studies. Leveraging admixture in Brazilians, whose chromosomes are mosaics of fragments of Native American, European and African origins, we used genome-wide data to perform admixture mapping/fine-mapping of Body Mass Index (BMI) in three population-based cohorts from Northeast (Salvador), Southeast (Bambuí) and South (Pelotas) of the country. We found significant associations with African-associated alleles in children from Salvador (PALD1 and ZMIZ1 genes), and in young adults from Pelotas (NOD2 and MTUS2 genes). More importantly, in Pelotas, rs114066381, mapped in a potential regulatory region, is significantly associated only in females ($p= 2.76 \times 10^{-6}$). This variant is very rare in Europeans but with frequencies of ~3% in West Africa, and has a strong female-specific effect (95%CI: 2.32-5.65 kg/m² per each A allele). We confirmed this sex-specific association and replicated its strong effect for an adjusted fat-mass index in the same Pelotas cohort, and for BMI in another Brazilian cohort from São Paulo (Southeast Brazil). A meta-analysis confirmed the significant association. Remarkably, we observed that while the frequency of rs114066381-A allele ranges from 0.8 to 2.1% in the studied populations, it attains ~9% among morbidly obese women from Pelotas, São Paulo, and Bambuí. The effect size of rs114066381 is at least five-times the effect size of the FTO SNPs rs9939609 and rs1558902, already emblematic for their high effects, and for which we replicated associations in Pelotas. We demonstrate how, after a decade of GWAS mostly performed in European-ancestry populations, non-European and admixed populations are a source of new relevant phenotype-associated genetic variants.

The Admixture mapping is a method that analyzes whether or not there is an association between a specific phenotype and the ancestry of a small chromosomal fragment. Despite being an excellent tool for the study of complex traits in admixed populations, one of the issues of this analysis is that the method implies that the association found refers to a region rather than a causal variant. GWAS share the same issue because they are usually performed with array data, which means that the hits are tag-SNPs that may not be the causal variant, but may be in linkage disequilibrium (LD) with it. To arrive at a more accurate result it is important to: do the association analysis for the region where the signal was found with more dense data (fine mapping); perform LD analyzes that will indicate if multiple significative SNPs in the same region represent the same signal (they are in LD) or two different hits ¹⁰⁸.

LD analyzes also serve to identify functional relevant variants around the signal, and to assess whether there is a relationship between the hits found in the study in question and in previous studies carried out in different populations. In this article we replicated 28 hits for BMI from GWAS Catalog ²⁹ and identified six new hits. My contribution was to conduct LD analysis (r^2), with the software Haploview ⁶⁸, between close significant SNPs identified in the fine mapping analysis, and between these SNPs and variants reported in GWAS Catalog that have been associated with BMI. We found that the SNP with the strongest effect on BMI in females (beta = 3.99 ± 0.84 6kg/m² per allele, 95% CI: 2.32-5.65, p = 2.76×10^{-6}), rs114066381, is not in LD ($r^2 < 0.001$) with rs113214936 (beta = 2,48, p = 1.51×10^{-5}), which is located in the same region, indicating that they are independent signals. In addition, we found that none of the six new hits identified in this study are in LD with the 389 BMI GWAS-Catalog-hits ($r^2 < 0.022$) in the 3 Brazilian cohorts, mining that the association between these regions and BMI were not identified before.

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

Abstract

The current search for host-susceptibility variants for COVID-19 contrasts with the fact that the study of the genetic architecture of Severe Acute Respiratory Syndrome (SARS) has been neglected. For human/SARS-CoV-2 interactome genes ACE2, TMPRSS2 and BSG, there is only one 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.

In the context of the worldwide scientific effort to understand all aspects of the infection caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), this article highlights the importance of considering the genetic diversity of human populations illustrated by the patterns found for the SNP rs2070788-G in the TMPRSS2 gene. This variant is a tag-SNP for rs383510, which is uncommon in arrays and is an eQTL for the TMPRSS2 gene ⁸⁰. The rs383510 has been associated with severe lung damage caused by influenza (A(H7N9) and A(H1N1)) in Asian populations ¹⁰⁹. A regression analysis performed with several world populations showed a strong association between rs2070788-G and Native American ancestry. My role in this article was to carry out the pairwise F_{ST} analysis between Native South American and East Asian populations (EAS from 1000 Genomes ¹¹⁰) with the hierfstat R package ¹¹¹. We found that rs2070788 is among the 5% most differentiated SNPs between these populations ($F_{ST} = 0.30$) (Figure 1, figure 1C of the paper). The rs2070788-G frequency in Asians is 30-40% while in Native Americans is 76-94%. Interestingly, when analyzing sequencing data from Native South Americans to access rs383510, we found that this SNP is not in LD with rs2070788 in this population. Therefore, the first impression that South American Natives could be more susceptible to SARS due to the high frequency of the tag-SNP rs2070788-G is mistaken and it is not a tag-SNP for rs383510 in Native Americans, so it is essential to do a control for ancestry in studies focused on these SNPs.

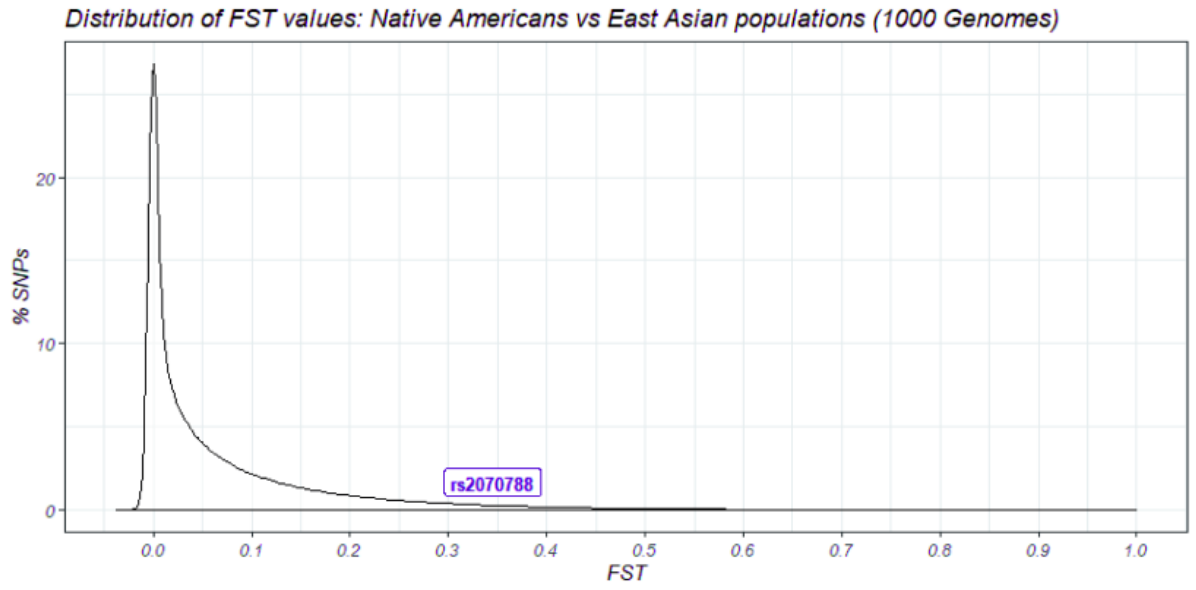


Figure 1. F_{ST} values distribution of Native Americans vs East Asian populations for 71 SNPs of TMPRSS2 gene

Final remarks

This thesis focused on the study of American populations that are underrepresented in genetic studies. In the first chapter, I 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 ¹. 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. In chapters 2 and 3 I present articles that exemplify the importance of this diversity by revealing new information on human history and health through the analysis of large datasets of native and admixed American populations.

The merging of datasets from the Peruvian Genome Project and the LDGH into a large dataset of Native South Americans enable us to increase the knowledge about the demographic history of these populations and their adaptation to the Andean and the Amazonian environment. Even though adaptation to the high altitude is a classical topic of evolutionary anthropology, we identify new genes with signals of natural selection, in addition to replicating a previously reported result (DUOX2) ¹⁷. On the other hand, studies focused on adaptation in the Amazon are very scarce and here we were able to contribute to fill this gap by identifying an important gene for immune response under selection in populations from this environment. However, as illustrated by the literature review and the lack of information available for some genes reported in chapter 2, there is a lack of functional studies to characterize the selection signals found in genome scans, and of polygenic studies that detect more subtle signals in gene networks ^{112,113}. Therefore, the work presented in this thesis led to the development of two new projects addressing these topics. The first with the objective of developing a polygenic selection test (submitted to CNPq as a postdoctoral proposal), and the second to perform functional studies for the genes HAND2-AS1, DUOX2 and PTPRC (submitted to The Leakey Foundation to apply for a grant).

The identification of loci that evolved under the effect of natural selection is an efficient method of identifying functional genomic regions that have played an essential role in survival, and possibly have consequences for human health ^{114,115}. However, other

evolutionary factors generate genetic diversity that is associated with different outcomes in medical treatments and diseases development ¹¹⁶. The lack of diversity in genetic studies makes it difficult to understand the real effect of new discoveries ¹⁰¹, and how to project this knowledge to different populations ²⁷. We exemplified that with the survey and annotation of the highly differentiated variants between Andean and Amazonian populations, reported in chapter 2. We were able to identify discrepancies in the frequency of variants of biomedical relevance that may affect the health of these populations differently. The same issue is addressed in the article presented in chapter 3, where we show that, due to differences in the pattern of linkage disequilibrium, an Asian tag-SNP for a SARS-related variant ¹⁰⁹ (the genetically closest continental group of Native Americans) cannot be used as a tag-SNP in Native South American populations.

Exploring the diversity of human genetics is not only important for the neglected populations, it is also essential for increasing the capacity to make new discoveries ¹⁰¹. The first two articles ^{30,31} in chapter 3 are examples of how the genomic mosaic of American populations is a rich source of information. Performing an admixture mapping in Brazilian cohorts we were able to make a contribution to the issue of missing heritability of complex traits identifying a non-European variant with a large effect size on BMI. Moreover, by mapping and analysing the patterns behind the genomic fragments from different source populations we were able to put more details on an important part of human history: the African diaspora to the American continent.

The work presented here shows in several ways how the effort to generate data from underrepresented populations, as done by the Peruvian Genome Project and the EPIGEN initiative ¹⁰⁶, bring important contributions to science. These results are a practical example that reinforce the importance of diversity in genetic studies, which in recent years has been declared by population geneticists in several journals ^{26,101,117-120}, 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 the locus related to BMI discovered in Brazilian miscegenated populations reported in Scliar et al. ³¹, and the clinical application of results of genetic studies to other populations due to the lack of knowledge of their allelic frequencies and population structure. We have 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/North American populations, not only bringing new

insights into the history of American populations and the genetics of complex traits, but by (which may be our greatest contribution) making public available the largest database of non-admixed South American native populations that we have today, which will allow our and other groups around the world to develop more inclusive studies.

References

1. Mendes, M., Alvim, I., Borda, V. & Tarazona-Santos, E. The history behind the mosaic of the Americas. *Curr. Opin. Genet. Dev.* **62**, 72–77 (2020).
2. Borda, V. *et al.* The genetic structure and adaptation of Andean highlanders and Amazonian dwellers is influenced by the interplay between geography and culture.
doi:10.1101/2020.01.30.916270.
3. Salzano, F. M. The role of natural selection in human evolution - insights from Latin America. *Genet. Mol. Biol.* **39**, 302–311 (2016).
4. Werren, E. A., Garcia, O. & Bigham, A. W. Identifying adaptive alleles in the human genome: from selection mapping to functional validation. *Hum. Genet.* (2020)
doi:10.1007/s00439-020-02206-7.
5. Lamason, R. L. *et al.* SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310**, 1782–1786 (2005).
6. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
7. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
8. Ohashi, J. *et al.* Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. *Am. J. Hum. Genet.* **74**, 1198–1208 (2004).
9. Fan, S., Hansen, M. E. B., Lo, Y. & Tishkoff, S. A. Going global by adapting local: A review of recent human adaptation. *Science* vol. 354 54–59 (2016).
10. Julian, C. G. & Moore, L. G. Human Genetic Adaptation to High Altitude: Evidence from the Andes. *Genes* **10**, (2019).
11. 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.* **111**, 165–176 (2000).

12. Frisancho, A. R. Developmental functional adaptation to high altitude: review. *Am. J. Hum. Biol.* **25**, 151–168 (2013).
13. Brutsaert, T. D., Araoz, M., Soria, R., Spielvogel, H. & Haas, J. D. Higher arterial oxygen saturation during submaximal exercise in Bolivian Aymara compared to European sojourners and Europeans born and raised at high altitude. *Am. J. Phys. Anthropol.* **113**, 169–181 (2000).
14. Bigham, A. *et al.* Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet.* **6**, e1001116 (2010).
15. Moore, L. G. *et al.* Maternal adaptation to high-altitude pregnancy: an experiment of nature--a review. *Placenta* **25 Suppl A**, S60–71 (2004).
16. Valverde, G. *et al.* A novel candidate region for genetic adaptation to high altitude in Andean populations. *PLoS One* **10**, e0125444 (2015).
17. Jacovas, V. C. *et al.* Selection scan reveals three new loci related to high altitude adaptation in Native Andeans. *Sci. Rep.* **8**, 12733 (2018).
18. Crawford, J. E. *et al.* Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans. *Am. J. Hum. Genet.* **101**, 752–767 (2017).
19. Eichstaedt, C. A. *et al.* The Andean adaptive toolkit to counteract high altitude maladaptation: genome-wide and phenotypic analysis of the Collas. *PLoS One* **9**, e93314 (2014).
20. Gouy, A., Daub, J. T. & Excoffier, L. Detecting gene subnetworks under selection in biological pathways. *Nucleic Acids Res.* **45**, e149 (2017).
21. Amorim, C. E. G., Daub, J. T., Salzano, F. M., Foll, M. & Excoffier, L. Detection of convergent genome-wide signals of adaptation to tropical forests in humans. *PLoS One* **10**, e0121557 (2015).
22. Fan, S. *et al.* African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol.* **20**, 82 (2019).
23. Perry, G. H. *et al.* Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E3596–603 (2014).
24. Bergey, C. M. *et al.* Polygenic adaptation and convergent evolution on growth and cardiac genetic pathways in African and Asian rainforest hunter-gatherers. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11256–E11263 (2018).

25. Hart, T. B. & Hart, J. A. The Ecological Basis of Hunter-Gatherer Subsistence in African Rain Forests: The Mbuti of Eastern Zaire. *Case Studies in Human Ecology* 55–83 (1996)
doi:10.1007/978-1-4757-9584-4_3.
26. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 1080 (2019).
27. Martin, A. R. *et al.* Current clinical use of polygenic scores will risk exacerbating health disparities. doi:10.1101/441261.
28. Whirl-Carrillo, M. *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92**, 414–417 (2012).
29. 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**, D1005–D1012 (2019).
30. Gouveia, M. H. *et al.* Origins, Admixture Dynamics, and Homogenization of the African Gene Pool in the Americas. *Mol. Biol. Evol.* **37**, 1647–1656 (2020).
31. Scliar, M. 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. *Int. J. Obes.* 1–13 (2021).
32. Kehdy, F. *et al.* Human-SARS-CoV-2 interactome and human genetic diversity: TMPRSS2-rs2070788, associated with severe influenza, and its population genetics caveats in Native Americans. doi:10.22541/au.159410553.31991801.
33. Deng, L., Ruiz-Linares, A., Xu, S. & Wang, S. Ancestry variation and footprints of natural selection along the genome in Latin American populations. *Sci. Rep.* **6**, 21766 (2016).
34. Fehren-Schmitz, L. & Georges, L. Ancient DNA reveals selection acting on genes associated with hypoxia response in pre-Columbian Peruvian Highlanders in the last 8500 years. *Sci. Rep.* **6**, 23485 (2016).
35. Salzano, F. M. The role of natural selection in human evolution – insights from Latin America. *Genetics and Molecular Biology* vol. 39 302–311 (2016).
36. Amorim, C. E. *et al.* Genetic signature of natural selection in first Americans. *Proc. Natl. Acad.*

- Sci. U. S. A.* **114**, 2195–2199 (2017).
37. Apata, M., Arriaza, B., Llop, E. & Moraga, M. Human adaptation to arsenic in Andean populations of the Atacama Desert. *Am. J. Phys. Anthropol.* **163**, 192–199 (2017).
 38. Eichstaedt, C. A. *et al.* Evidence of Early-Stage Selection on EPAS1 and GPR126 Genes in Andean High Altitude Populations. *Sci. Rep.* **7**, 13042 (2017).
 39. Massey, S. E. Strong Amerindian Mitonuclear Discordance in Puerto Rican Genomes Suggests Amerindian Mitochondrial Benefit. *Ann. Hum. Genet.* **81**, 59–77 (2017).
 40. Moore, L. G. Measuring high-altitude adaptation. *J. Appl. Physiol.* **123**, 1371–1385 (2017).
 41. Mychaleckyj, J. C. *et al.* Genome-Wide Analysis in Brazilians Reveals Highly Differentiated Native American Genome Regions. *Mol. Biol. Evol.* **34**, 559–574 (2017).
 42. Oettlé, A. C., Demeter, F. P. & L'abbé, E. N. Ancestral Variations in the Shape and Size of the Zygoma. *Anat. Rec.* **300**, 196–208 (2017).
 43. Seguchi, N., Quintyn, C. B., Yonemoto, S. & Takamuku, H. An assessment of postcranial indices, ratios, and body mass versus eco-geographical variables of prehistoric Jomon, Yayoi agriculturalists, and Kumejima Islanders of Japan. *Am. J. Hum. Biol.* **29**, (2017).
 44. Stobdan, T. *et al.* New Insights into the Genetic Basis of Monge's Disease and Adaptation to High-Altitude. *Molecular Biology and Evolution* vol. 34 3154–3168 (2017).
 45. Auerbach, B. M., King, K. A., Campbell, R. M., Campbell, M. L. & Sylvester, A. D. Variation in obstetric dimensions of the human bony pelvis in relation to age-at-death and latitude. *Am. J. Phys. Anthropol.* **167**, 628–643 (2018).
 46. Posth, C. *et al.* Reconstructing the Deep Population History of Central and South America. *Cell* **175**, 1185–1197.e22 (2018).
 47. Guimarães, L. O. *et al.* Genetic ancestry effects on the distribution of toll-like receptors (TLRs) gene polymorphisms in a population of the Atlantic Forest, São Paulo, Brazil. *Human Immunology* vol. 79 101–108 (2018).
 48. Hlusko, L. J. *et al.* Environmental selection during the last ice age on the mother-to-infant transmission of vitamin D and fatty acids through breast milk. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4426–E4432 (2018).

49. Hsueh, W.-C. *et al.* Analysis of type 2 diabetes and obesity genetic variants in Mexican Pima Indians: Marked allelic differentiation among Amerindians at HLA. *Annals of Human Genetics* vol. 82 287–299 (2018).
50. Kano, F. S. *et al.* Susceptibility to *Plasmodium vivax* malaria associated with DARC (Duffy antigen) polymorphisms is influenced by the time of exposure to malaria. *Sci. Rep.* **8**, 13851 (2018).
51. Li, J. *et al.* Natural Selection Has Differentiated the Progesterone Receptor among Human Populations. *Am. J. Hum. Genet.* **103**, 45–57 (2018).
52. Mathieson, S. & Mathieson, I. FADS1 and the Timing of Human Adaptation to Agriculture. *Mol. Biol. Evol.* **35**, 2957–2970 (2018).
53. Norris, E. T. *et al.* Genetic ancestry, admixture and health determinants in Latin America. *BMC Genomics* **19**, 861 (2018).
54. Oliveira, M. L. G. de *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. *Human Immunology* vol. 79 790–799 (2018).
55. Racimo, F., Berg, J. J. & Pickrell, J. K. Detecting Polygenic Adaptation in Admixture Graphs. *Genetics* **208**, 1565–1584 (2018).
56. Rodrigues, J. C. G. *et al.* Polymorphisms of ADME-related genes and their implications for drug safety and efficacy in Amazonian Amerindians. *Sci. Rep.* **9**, 7201 (2019).
57. Adhikari, K. *et al.* A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia. *Nat. Commun.* **10**, 358 (2019).
58. Ávila-Arcos, M. C. *et al.* Population History and Gene Divergence in Native Mexicans Inferred from 76 Human Exomes. *Mol. Biol. Evol.* **37**, 994–1006 (2020).
59. Calonga-Solís, V. *et al.* 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. *Frontiers in Immunology* vol. 10 (2019).
60. Gazal, S. *et al.* The Genetic Architecture of Chronic Mountain Sickness in Peru. *Front. Genet.* **10**, 690 (2019).

61. Harris, D. N. *et al.* Evolution of Hominin Polyunsaturated Fatty Acid Metabolism: From Africa to the New World. *Genome Biol. Evol.* **11**, 1417–1430 (2019).
62. Reynolds, A. W. *et al.* Comparing signals of natural selection between three Indigenous North American populations. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 9312–9317 (2019).
63. Ross, A. H. & Ubelaker, D. H. Complex Nature of Hominin Dispersals: Ecogeographical and Climatic Evidence for Pre-Contact Craniofacial Variation. *Sci. Rep.* **9**, 11743 (2019).
64. Vicuña, L. *et al.* Adaptation to Extreme Environments in an Admixed Human Population from the Atacama Desert. *Genome Biology and Evolution* vol. 11 2468–2479 (2019).
65. Zaidi, A. A. & Makova, K. D. Investigating mitonuclear interactions in human admixed populations. *Nat Ecol Evol* **3**, 213–222 (2019).
66. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
67. Cockerham, C. C. & Weir, B. S. Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics* **40**, 157–164 (1984).
68. Barrett, J. C. Haploview: Visualization and Analysis of SNP Genotype Data. *Cold Spring Harbor Protocols* vol. 2009 db.ip71–pdb.ip71 (2009).
69. DANCE. <http://www.ldgh.com.br/alpha/#/annotate>.
70. Raney, B. J. *et al.* Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**, 1003–1005 (2014).
71. Anderson, L. L., Mao, X., Scott, B. A. & Crowder, C. M. Survival from hypoxia in *C. elegans* by inactivation of aminoacyl-tRNA synthetases. *Science* **323**, 630–633 (2009).
72. Fu, R., Fan, Y.-Z., Fan, Y.-C. & Zhao, H.-Y. Expression of arginyl-tRNA synthetase in rats with focal cerebral ischemia. *Journal of Huazhong University of Science and Technology [Medical Sciences]* vol. 34 172–175 (2014).
73. Kamphuis, W., Dijk, F. & Bergen, A. A. B. Ischemic preconditioning alters the pattern of gene expression changes in response to full retinal ischemia. *Mol. Vis.* **13**, 1892–1901 (2007).
74. Aspatwar, A. *et al.* Catalytically inactive carbonic anhydrase-related proteins enhance transport of lactate by MCT1. *FEBS Open Bio* **9**, 1204–1211 (2019).

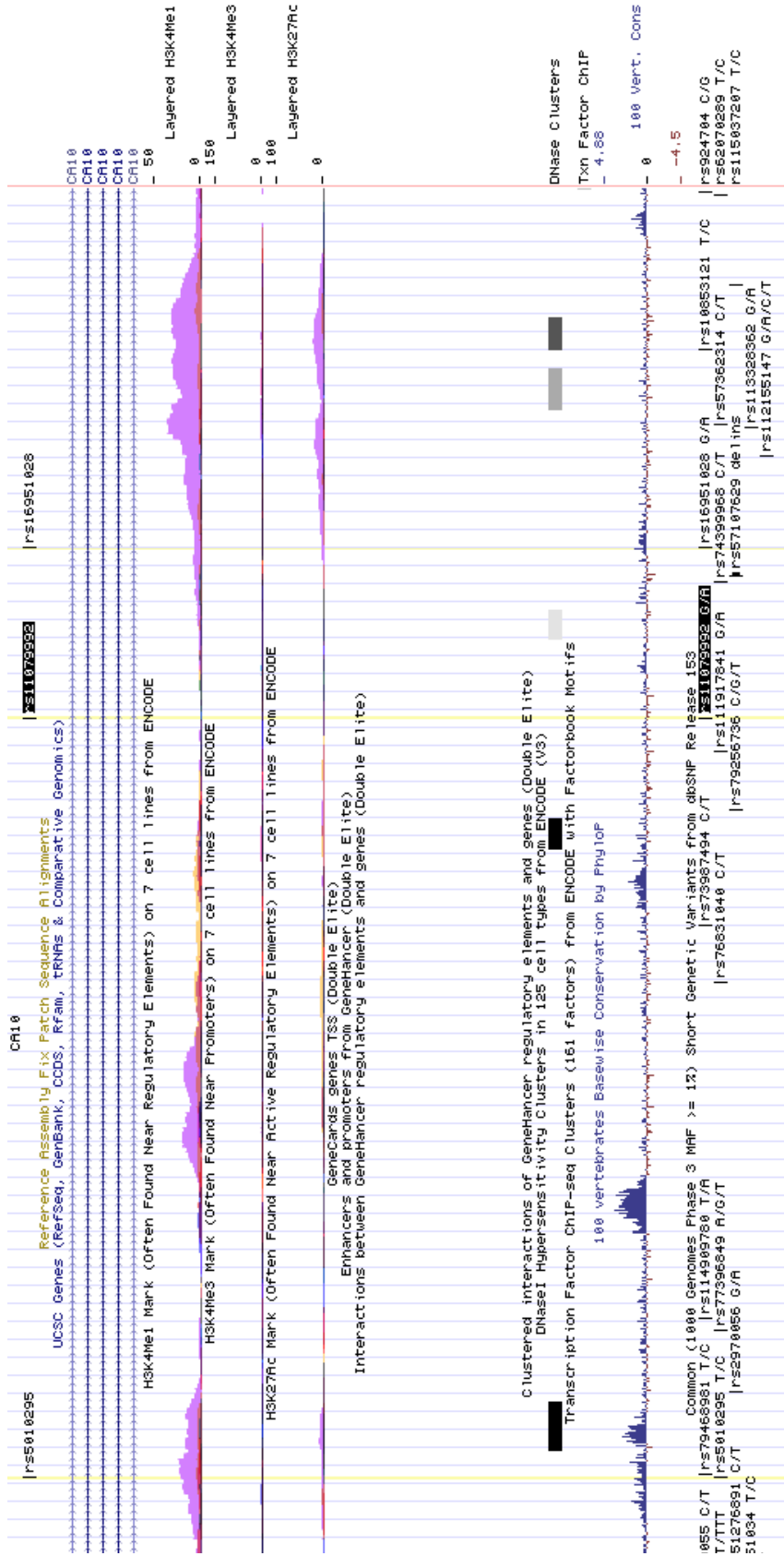
75. Sterky, F. H. *et al.* Carbonic anhydrase-related protein CA10 is an evolutionarily conserved pan-neurexin ligand. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1253 (2017).
76. Perin, P. & Potočnik, U. Polymorphisms in recent GWA identified asthma genes CA10 , SGK493 , and CTNNA3 are associated with disease severity and treatment response in childhood asthma. *Immunogenetics* **66**, 143–151 (2014).
77. Tekola-Ayele, F. *et al.* Genome-wide association study identifies African-ancestry specific variants for metabolic syndrome. *Mol. Genet. Metab.* **116**, 305 (2015).
78. Moon, S. *et al.* A genome-wide association study of copy-number variation identifies putative loci associated with osteoarthritis in Koreans. *BMC Musculoskelet. Disord.* **16**, (2015).
79. Mori, S. *et al.* Nucleotide variations in genes encoding carbonic anhydrase 8 and 10 associated with femoral bone mineral density in Japanese female with osteoporosis. *J. Bone Miner. Metab.* **27**, (2009).
80. GTEx Portal. <https://www.gtexportal.org/home/>.
81. Hodges, L. M. *et al.* Very important pharmacogene summary: ABCB1 (MDR1, P-glycoprotein). *Pharmacogenet. Genomics* **21**, 152–161 (2011).
82. Fohner, A. E., Brackman, D. J., Giacomini, K. M., Altman, R. B. & Klein, T. E. PharmGKB summary: very important pharmacogene information for ABCG2. *Pharmacogenet. Genomics* **27**, 420–427 (2017).
83. Kim, Y. *et al.* Influence of OATP1B1 and BCRP polymorphisms on the pharmacokinetics and pharmacodynamics of rosuvastatin in elderly and young Korean subjects. *Scientific Reports* vol. 9 (2019).
84. Kashiwara, Y. *et al.* Small-Dosing Clinical Study: Pharmacokinetic, Pharmacogenomic (SLCO2B1 and ABCG2), and Interaction (Atorvastatin and Grapefruit Juice) Profiles of 5 Probes for OATP2B1 and BCRP. *J. Pharm. Sci.* **106**, 2688–2694 (2017).
85. Birmingham, B. K. *et al.* Rosuvastatin pharmacokinetics and pharmacogenetics in Caucasian and Asian subjects residing in the United States. *Eur. J. Clin. Pharmacol.* **71**, 329–340 (2015).
86. DeGorter, M. K. *et al.* Clinical and pharmacogenetic predictors of circulating atorvastatin and rosuvastatin concentrations in routine clinical care. *Circ. Cardiovasc. Genet.* **6**, 400–408 (2013).

87. Keskitalo, J. E. *et al.* ABCG2 polymorphism markedly affects the pharmacokinetics of atorvastatin and rosuvastatin. *Clin. Pharmacol. Ther.* **86**, 197–203 (2009).
88. AmiKoWeb. <https://amiko.oddb.org/de/fi?gtin=66835>.
89. Wen, C. C. *et al.* Genome-wide association study identifies ABCG2 (BCRP) as an allopurinol transporter and a determinant of drug response. *Clinical Pharmacology & Therapeutics* vol. 97 518–525 (2015).
90. Roberts, R. L. *et al.* ABCG2 loss-of-function polymorphism predicts poor response to allopurinol in patients with gout. *Pharmacogenomics J.* **17**, 201–203 (2017).
91. Wright, D. F. B. *et al.* The impact of diuretic use and ABCG2 genotype on the predictive performance of a published allopurinol dosing tool. *Br. J. Clin. Pharmacol.* **84**, 937–943 (2018).
92. Brackman, D. J. *et al.* Genome-Wide Association and Functional Studies Reveal Novel Pharmacological Mechanisms for Allopurinol. *Clin. Pharmacol. Ther.* **106**, 623–631 (2019).
93. Hu, A. M. & Brown, J. N. Comparative effect of allopurinol and febuxostat on long-term renal outcomes in patients with hyperuricemia and chronic kidney disease: a systematic review. *Clin. Rheumatol.* (2020) doi:10.1007/s10067-020-05079-3.
94. Lamb, Y. N. Rosuvastatin/Ezetimibe: A Review in Hypercholesterolemia. *Am. J. Cardiovasc. Drugs* (2020) doi:10.1007/s40256-020-00421-1.
95. Hu, Y.-H. *et al.* Methotrexate Disposition in Pediatric Patients with Acute Lymphoblastic Leukemia: What Have We Learnt From the Genetic Variants of Drug Transporters. *Curr. Pharm. Des.* **25**, 627–634 (2019).
96. Miranda-Filho, A. *et al.* Epidemiological patterns of leukaemia in 184 countries: a population-based study. *Lancet Haematol* **5**, e14–e24 (2018).
97. Yang, J. J. *et al.* Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. *Nat. Genet.* **43**, 237–241 (2011).
98. de Carvalho, D. C. *et al.* Characterization of pharmacogenetic markers related to Acute Lymphoblastic Leukemia toxicity in Amazonian native Americans population. *Sci. Rep.* **10**, 10292 (2020).
99. PharmGKB. <https://www.pharmgkb.org/variant/PA166157284/clinicalAnnotation>.

100. PharmGKB. <https://www.pharmgkb.org/variant/PA166157284/clinicalAnnotation/1296599132>.
101. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
102. Maruo, Y. *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).
103. De Deken, X., Corvilain, B., Dumont, J. E. & Miot, F. Roles of DUOX-mediated hydrogen peroxide in metabolism, host defense, and signaling. *Antioxid. Redox Signal.* **20**, 2776–2793 (2014).
104. van der Vliet, A., Danyal, K. & Heppner, D. E. Dual oxidase: a novel therapeutic target in allergic disease. *Br. J. Pharmacol.* **175**, 1401–1418 (2018).
105. Stanton, T. *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).
106. Kehdy, F. S. G. *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).
107. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
108. Martin, E. R. *et al.* Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations. *Genet. Epidemiol.* **42**, 214–229 (2018).
109. Cheng, Z. *et al.* Identification of TMPRSS2as a Susceptibility Gene for Severe 2009 Pandemic A(H1N1) Influenza and A(H7N9) Influenza. *Journal of Infectious Diseases* vol. 212 1214–1221 (2015).
110. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* vol. 526 68–74 (2015).
111. de Meeûs, T. & Goudet, J. A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. *Infect. Genet. Evol.* **7**, 731–735 (2007).
112. Wellenreuther, M. & Hansson, B. Detecting Polygenic Evolution: Problems, Pitfalls, and Promises. *Trends Genet.* **32**, 155–164 (2016).

113. Barghi, N., Hermisson, J. & Schlötterer, C. Polygenic adaptation: a unifying framework to understand positive selection. *Nat. Rev. Genet.* (2020) doi:10.1038/s41576-020-0250-z.
114. Stearns, S. C., Nesse, R. M., Govindaraju, D. R. & Ellison, P. T. Evolution in health and medicine Sackler colloquium: Evolutionary perspectives on health and medicine. *Proc. Natl. Acad. Sci. U. S. A.* **107 Suppl 1**, 1691–1695 (2010).
115. Vasseur, E. & Quintana-Murci, L. The impact of natural selection on health and disease: uses of the population genetics approach in humans. *Evol. Appl.* **6**, 596–607 (2013).
116. Ridley, M. *Evolution*. (Wiley, 2003).
117. Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S. Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.* **20**, 520–535 (2019).
118. Morales, J. *et al.* A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, 21 (2018).
119. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* vol. 538 161–164 (2016).
120. Hindorff, L. A. *et al.* Diversity matters. *Nat. Rev. Genet.* **20**, 495–495 (2019).

Attachment 1. UCSC Genome Browser view of CA10 locus.



Attachment 1. UCSC Genome Browser view of the locus in chr 17 with highly differentiated SNPs rs16951028 and rs5010295 in gene CA10, DNase I hypersensitivity clusters, Transcription Factor ChIP-seq binding sites, and the histone modifications H3K27ac (Often Found Near Active Regulatory Elements), H3K4me1 (Often Found Near Regulatory Elements) and H3K4me3 (Often Found Near Promoters) on cell lines from the ENCODE Project, and GeneHancer and vertebrate conservation data.

Attachment 2. Linkage disequilibrium (r^2) between SNPs with high F_{ST} values associated with drug metabolism (>0.27) in the pharmacogene gene ABCB1 in (A) Andean and (B) Amazon populations.

