

# UC Davis

## UC Davis Previously Published Works

### Title

Uncovering Signals of Positive Selection in Peruvian Populations from Three Ecological Regions

### Permalink

<https://escholarship.org/uc/item/5345997r>

### Journal

Molecular Biology and Evolution, 39(8)

### ISSN

0737-4038

### Authors

Caro-Consuegra, Rocio

Nieves-Colón, Maria A

Rawls, Erin

et al.

### Publication Date

2022-08-03

### DOI


10.1093/molbev/msac158

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at <https://creativecommons.org/licenses/by-nc/4.0/>

Peer reviewed

# Uncovering Signals of Positive Selection in Peruvian Populations from Three Ecological Regions

Rocio Caro-Consuegra,<sup>1,†</sup> Maria A. Nieves-Colón,<sup>2,3,4,†</sup> Erin Rawls,<sup>3</sup> Verónica Rubin-de-Celis,<sup>5</sup> Beatriz Lizárraga,<sup>6</sup> Tatiana Vidaurre,<sup>7</sup> Karla Sandoval,<sup>2</sup> Laura Fejerman,<sup>8</sup> Anne C. Stone,<sup>3,9</sup> Andrés Moreno-Estrada,<sup>2,\*</sup> and Elena Bosch <sup>1,10,\*</sup>

<sup>1</sup>Institute of Evolutionary Biology (UPF-CSIC), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona 08003, Catalonia, Spain

<sup>2</sup>Laboratorio Nacional de Genómica para la Biodiversidad, Unidad de Genómica Avanzada (UGA-LANGEBIO), CINVESTAV, Irapuato, Guanajuato 36821, Mexico

<sup>3</sup>School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85287, USA

<sup>4</sup>Department of Anthropology, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA

<sup>5</sup>Laboratorio de Genómica Molecular Evolutiva, Instituto de Ciencia y Tecnología, Universidad Ricardo Palma, Lima 33, Ap 1801, Peru

<sup>6</sup>Emeritus Professor, Facultad de Ciencias Biológicas, Universidad Nacional Mayor de San Marcos, 15081 Lima, Peru

<sup>7</sup>Instituto de Enfermedades Neoplásicas, 15038 Surquillo, Lima, Peru

<sup>8</sup>Department of Public Health Sciences, University of California Davis, Davis, CA 95616, USA

<sup>9</sup>Center for Evolution and Medicine, Arizona State University, Tempe, AZ 85287-4501, USA

<sup>10</sup>Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), 43206 Reus, Spain

\*Corresponding authors: E-mails: andres.moreno@cinvestav.mx; elena.bosch@upf.edu.

†Co-first authors.

Associate editor: Connie Mulligan

## Abstract

Peru hosts extremely diverse ecosystems which can be broadly classified into the following three major ecoregions: the Pacific desert coast, the Andean highlands, and the Amazon rainforest. Since its initial peopling approximately 12,000 years ago, the populations inhabiting such ecoregions might have differentially adapted to their contrasting environmental pressures. Previous studies have described several candidate genes underlying adaptation to hypobaric hypoxia among Andean highlanders. However, the adaptive genetic diversity of coastal and rainforest populations has been less studied. Here, we gathered genome-wide single-nucleotide polymorphism-array data from 286 Peruvians living across the three ecoregions and analyzed signals of recent positive selection through population differentiation and haplotype-based selection scans. Among highland populations, we identify candidate genes related to cardiovascular function (*TLL1*, *DUSP27*, *TBX5*, *PLXNA4*, *SGCD*), to the Hypoxia-Inducible Factor pathway (*TGFA*, *APIP*), to skin pigmentation (*MITF*), as well as to glucose (*GLIS3*) and glycogen metabolism (*PPP1R3C*, *GANC*). In contrast, most signatures of adaptation in coastal and rainforest populations comprise candidate genes related to the immune system (including *SIGLEC8*, *TRIM21*, *CD44*, and *ICAM1* in the coast; *CBLB* and *PRDM1* in the rainforest; and *BRD2*, *HLA-DOA*, *HLA-DPA1* regions in both), possibly as a result of strong pathogen-driven selection. This study identifies candidate genes related to human adaptation to the diverse environments of South America.

**Key words:** Peruvian populations, high-altitude adaptation, human adaptation.

## Introduction

Since the Out-of-Africa event, humans have spread across the world inhabiting regions with a wide variety of environments. Thus, human populations have been exposed to very different selective pressures, which have left traceable marks in the human genome (Nielsen *et al.* 2017). Peru hosts extremely diverse ecosystems which can be broadly classified into three major ecological regions (ecoregions): the arid Pacific coast, the Andean highlands, and

the Amazon rainforest (Ponce de León Bardalez 1994). This diversity provides a unique opportunity to identify those genomic regions and functional variants that could have favored human adaptation to the contrasting environmental pressures across the Americas.

Ancestral Indigenous American populations are estimated to have diverged from east Asians around 23,000 years ago (ya) and entered the Americas after an approximately 8,000 year period of isolation in Beringia (Raghavan

© The Author(s) 2022. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

Open Access

*et al.* 2015). Although this was likely followed by a divergence into northern and southern branches between 17,500 and 14,600 ya in south-eastern Beringia, the specifics of the posterior population dispersal across the Americas are still under debate (Reich *et al.* 2012; Raghavan *et al.* 2015; Moreno-Mayar *et al.* 2018; Posth *et al.* 2018). Archaeological findings—such as the site of Monte Verde in southern Chile, dating to approximately 14,000 ya (Dillehay *et al.* 2008)—and genetic evidence, suggest that after the initial entry into the Americas humans dispersed rapidly, probably moving along a coastal route and reaching South America within a 1,500 year time span (Brandini *et al.* 2018; Goldberg *et al.* 2016; Llamas *et al.* 2016).

As for the three Peruvian ecoregions under study in this work, no consensus has been reached on how they were populated, but two main hypotheses have been suggested. The first hypothesis posits that there was a split migration through both sides of the Andes, followed by a divergence of highland and coastal populations (Skoglund and Reich 2016; Gómez-Carballa *et al.* 2018; Harris *et al.* 2018). The second hypothesis suggests that the three ecoregions were populated in parallel 12,000 ya by populations descending from Ancestral South Americans (Rothhammer and Dillehay 2009; Harris *et al.* 2018). Genetic studies have identified differential genetic patterns when comparing populations from the Andes and Amazonia. These patterns suggest that Andean communities were large and connected by gene flow, whereas Amazonian groups were probably small and isolated (Tarazona-Santos *et al.* 2001; Fuselli *et al.* 2003; Lewis *et al.* 2007). However, more recent archaeological modeling challenges that view by pointing to material evidence suggestive of the existence of large-scale Amazonian societies (de Souza *et al.* 2018). Moreover, there is evidence for extensive past migrations and gene flow across the three ecoregions probably fostered by trade in products unique to each (Rodríguez-Delfin *et al.* 2001; Tarazona-Santos *et al.* 2001; Lewis *et al.* 2005; Sandoval *et al.* 2013; Barbieri *et al.* 2019; Castro e Silva *et al.* 2022). Many of these population movements occurred in the pre-Columbian period, when state-level civilizations such as the Inca Empire forced population movements across the region (Quilter 2013; D'Altroy 2014; Barberena *et al.* 2020). Recent genetic research also suggests that some of these migrations may have occurred in response to climatic fluctuations (Fehren-Schmitz *et al.* 2014). Subsequent movements also occurred during the Spanish conquest in the 16th century, which not only induced migration, but also caused important bottlenecks as a result of wars and of newly imported pathogens and epidemics (Merbs 1992; Patterson and Runge 2002; Livi-Bacci 2006; Riley 2010; O'Fallon and Fehren-Schmitz 2011; Lindo *et al.* 2016; Llamas *et al.* 2016). Furthermore, during the colonial period a number of admixture events occurred among Indigenous South Americans, European settlers, and populations of African ancestry who were forcibly brought to Peru by the Transatlantic Slave Trade (Gravel *et al.* 2013; Homburger

*et al.* 2015; Chacón-Duque *et al.* 2018; Harris *et al.* 2018; Barbieri *et al.* 2019).

The diverse environments found across Peru have also influenced patterns of human settlement, migration, interaction, and subsistence (Escobar and Beall 1982; Cárdenas-Arroyo and Bray 1998; Murra 2002). As a global biodiversity hotspot, Peru hosts multiple ecosystems with high species endemism (de Queiroz *et al.* 2014; Garcia-Longoria *et al.* 2022). This biodiversity can be classified into ecological regions; defined land areas which contain distinct natural communities and differ from each other due to factors such as topography, climate, and vegetation (Olson *et al.* 2001). Following this framework, three distinctive and sharply contrasting ecoregions have been identified in Peru: the hyper-arid desert Pacific coast, the high-altitude Andean mountain range, and the tropical rainforest lowlands of the Amazon (de Queiroz *et al.* 2014). These three ecoregions differ in elevation, vegetation, faunal communities, and overall climate (although see Vidal 2014 and Britto 2017 for alternate classifications). We hypothesize that given the contrasting pressures posed by differences in environment, geography, and historical population sizes between coastal, highland and lowland rainforest environments, patterns of adaptation may differ in native populations from each Peruvian ecoregion.

Several studies have focused on identifying human adaptation to hypobaric hypoxia at high altitudes, usually defined as altitudes >2,500 m above sea level (Moore 2001; Bigham *et al.* 2013; Jacovas *et al.* 2018). These high-altitude regions comprise populations living in the Andean Altiplano, as well as in the Qinghai–Tibetan plateau in China and the Semien plateau in Ethiopia (reviewed in Bigham and Lee 2014 and Moore 2017). Lowlanders experience a set of short-term physiological responses (or acclimatization) when exposed to lower atmospheric oxygen concentrations, such as an increase in ventilation rate, a reduction in plasma volume, and an increase in erythrocyte production, among others (Heath and Williams 1995; Siebenmann *et al.* 2015). These responses aim to compensate for the arterial oxygen saturation decline, collectively contributing to an increase in hemoglobin. However, high concentrations of hemoglobin increase blood viscosity, exerting an additional stress on the cardiovascular system and hindering appropriate blood flow to the tissues (Guyton and Richardson 1961). In populations permanently inhabiting high-altitude environments, several physiological adaptations have arisen to avoid or somehow compensate for the putative negative effects of such physiological responses. Andeans, unlike other long-term highland populations, show larger hemoglobin concentrations (Beall 2007) and thus higher blood viscosity. However, to counteract its effects, they also present an additional layer of muscles in the pulmonary artery (Penaloza and Arias-Stella 2007) and higher pulmonary vasoconstriction (Rupert and Hochachka 2001; Beall 2007). These specific adaptive phenotypes suggest that local genetic adaptations to hypoxia must have arisen in Andean highlanders.

Genome scans of positive selection focused on highland populations have described candidate genes for adaptation to hypoxia in the Hypoxia-Inducible transcription Factor (HIF) pathway (Bigham *et al.* 2009, 2010). Among these, *EGLN1* plays a critical role in oxygen homeostasis in mammals and has been found to be under selection in both Andean and Tibetan populations (Bigham *et al.* 2009; Yi *et al.* 2010). *ENDRA*, *PRKAA1*, and *NOS2A* (Bigham *et al.* 2009, 2010) are involved in vasoregulation and have been associated with reproductive success (Moore 2010). Crawford *et al.* (2017) generated low-coverage whole genome sequencing data from a sample of high-altitude resident Andeans and identified genes such as *BRINP3*, *NOS2*, and *TBX5*, which play an important functional role in the cardiovascular system, among the strongest signals of positive selection. Thus, a genetic adaptive strategy to permanent hypoxia related to cardiovascular phenotypes was specifically suggested for Andeans. In agreement with that, when exploring signatures of positive selection in Andeans, Borda *et al.* (2020) identified *HAND2-AS1*, which is involved in the modulation of cardiogenesis (Anderson *et al.* 2016; Cheng and Jiang 2019), but also *DUOX2*, which plays a role in the synthesis of thyroid hormones and in innate immunity (De Deken *et al.* 2014; Maruo *et al.* 2016; van der Vliet *et al.* 2018).

Deserts are typically characterized not only by aridity and water scarcity but also by extreme temperature changes and intense UV radiation. Studies of human adaptation to desert conditions are quite scarce (reviewed in Rocha *et al.* 2021). Moreover, in the case of South American deserts, most selection scans have focused on populations from northern Chile (Apata *et al.* 2017; Vicuña *et al.* 2019) and Argentina (Eichstaedt *et al.* 2015), which have been historically exposed to inorganic arsenic. As for the Amazon rainforest, characterized by low light incidence, tropical climate, and high pathogen diversity (Guernier *et al.* 2004), Borda *et al.* (2020) recently described the *PTPRC* gene, which regulates T- and B-cell antigen receptor interactions and plays a major role in the innate immune system (Anand and Ganju 2006; Dawes *et al.* 2006; Caignard *et al.* 2013; Windheim *et al.* 2013), as one of the top signals for positive selection. These results are in concordance with previous genome scans for positive selection performed in other rainforest populations, where a number of immune-related candidate genes have been identified (Amorim *et al.* 2015; reviewed in Fan *et al.* 2016).

In this study, genome-wide single-nucleotide polymorphism (SNP) data (Illumina Infinium® MEGA array) was gathered for 286 individuals distributed across Peru, and then analyzed to investigate signals of recent positive selection specific to populations living in the high-altitude environments of the Andes, the arid Pacific coast, and the Amazon rainforest. To this aim, we computed the Population Branch Statistic (PBS) and performed haplotype-based selection scans with the integrated Haplotype Score (iHS) and the cross-population Extended Haplotype Homozygosity (XP-EHH) statistics. Besides identifying the strongest signals of recent positive

selection specific to each ecoregion, we used gene-set and trait-associated SNP over-representation approaches to unravel more subtle trends of adaptation in specific biological functions. Furthermore, we also explored genotype frequency changes correlated with elevation to identify putative adaptive changes directly related to environmental selective pressures across ecoregions.

## Results

### Dataset

We genotyped 189 samples collected from populations distributed across Peru with the Illumina Infinium® Multi-Ethnic Global Array (MEGA). The resulting dataset was merged with 119 additional Peruvian individual samples genotyped with the same array (Wojcik *et al.* 2019, Ioannidis *et al.* 2020; see Materials and Methods), as well as to the publicly available genotype data from four populations from the 1000 Genomes Project (1 KGP): Yoruba from Ibadan, Nigeria (YRI), Utah residents with Northern and Western European Ancestry (CEU), Han Chinese from Beijing (CHB), and Peruvians from Lima (PEL) (Auton *et al.* 2015). After quality control (QC), the intersected dataset included 610,576 autosomal SNPs for 681 individuals, including 85 PEL and 286 newly compiled Peruvian individuals (supplementary figs. S1 and S2, Supplementary Material online). The latter were arranged into 15 departments across three differentiated ecoregions: Andean highlands ( $n=94$ , after downsampling Puno individuals to 25 individuals to avoid over-

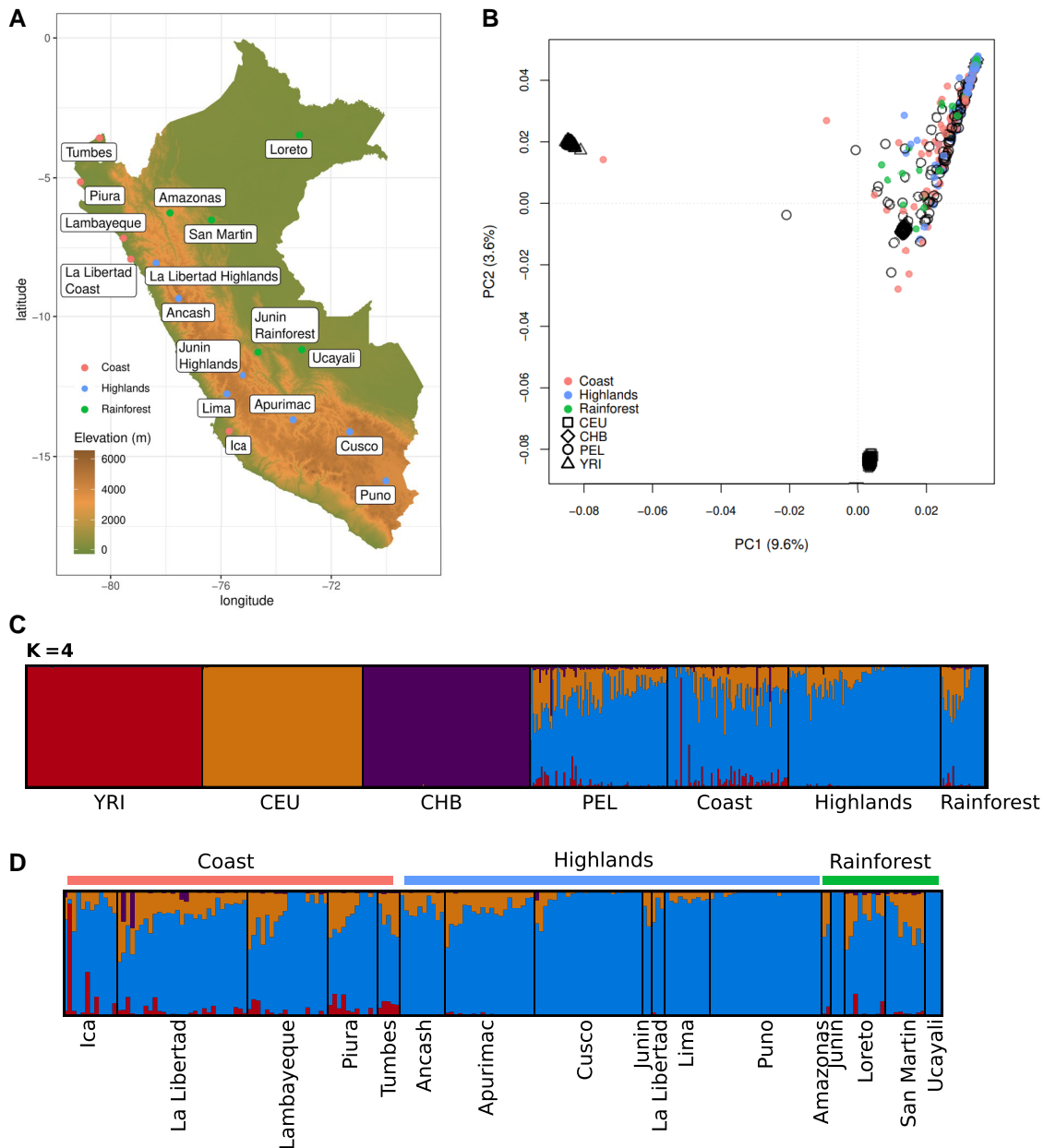
**Table 1.** Individual Samples Grouped by Department and Ecoregion

Ecoregion	Department	Altitude range (mamsl)	N
Coast	Ica	15–585	11 (+1) <sup>a</sup>
	La Libertad	28	28 (+1) <sup>a</sup>
	Lambayeque	5–43	18
	Piura	5–60	11
	Tumbes	6–12	5
	Highlands	Apurimac	2,760–3,665
	Cusco	3,345–3,913	24
	Ancash	2,965–3,281	10
	La Libertad	2,641–3,099	3
	Junin	3,249	2
	Lima	2,836	10
	Puno	3,827	25 (+75) <sup>b</sup>
Rainforest	Amazonas	1,022–1,630	2
	Loreto	100–111	9
	San Martin	207–860	9
	Junin	631	3
	Ucayali	280	4

Samples were assigned to each department and ecoregion according to the population collection site and/or additional origin information available (for details, see supplementary table S1, Supplementary Material online). La Libertad and Junin departments comprise two differentiated ecoregions.

<sup>a</sup>One individual from La Libertad and one from Ica were excluded from the selection analyses due to the high proportions of European and African ancestries detected in them, respectively.

<sup>b</sup>Puno samples were randomly downsampled to avoid over-representation in the genetic structure and selection analyses.  
mamsl, meters above mean sea level.



**Fig. 1.** (A) Map of Peru indicating 15 departments where sampled populations are located. Data points colored by ecoregion. Elevation is represented in meters above sea level using the colorbar scale. (B) PCA of Peruvian individuals colored by ecoregion, including populations from the 1000 Genomes Project (1 KGP\*). (C) ADMIXTURE plot at  $K=4$  with the newly analyzed Peruvian samples classified per ecoregion. (D) Zoom in on the ADMIXTURE plot at  $K=4$  for the analyzed Peruvian samples further divided by department. \*YRI, Yoruba from Ibadan, Nigeria; CEU, Utah residents with Northern and Western European Ancestry; CHB, Han Chinese from Beijing; PEL, Peruvians from Lima.

representation), arid coast ( $n=75$ ), and Amazon rainforest ( $n=27$ ) (table 1 and fig. 1A; for details about collection sites and excluded individuals see Materials and Methods and supplementary table S1, Supplementary Material online).

### Population Structure and Admixture

Global population structure and admixture patterns across the three Peruvian ecoregions were investigated by means of Principal Component Analysis (PCA) and ADMIXTURE analyses including the YRI, CEU, CHB, and PEL reference populations from the 1 KGP (fig. 1B–D

and supplementary fig. S3, Supplementary Material online). The first principal component (PC1) clearly divides African from non-African ancestries, while PC2 separates European ancestries from the rest. Most Peruvian individuals cluster near the CHB individuals but form a rather dispersed cloud around them as a result of their diversity of African and European ancestry proportions. Individuals from the highlands have the smallest proportions of continental admixture relative to the rest of our sample.

In the ADMIXTURE analysis, the lowest cross-validation error is found when four ancestral components ( $K=4$ ) are considered. These represent African, European, and East

Asian ancestries, in addition to Native American ancestry (shown in blue in [fig. 1C and D](#)). The Peruvian individuals from Lima (PEL) included in the 1 KGP and the Peruvian populations included in our analyses show varying proportions of admixture with European and African ancestries, as expected given the PCA. Specifically, individuals from the coast carry the highest proportions of African admixture with an average African ancestry component of 13.4% in the Ica department, while the maximum average African ancestry component for highlanders is only 0.5% in the Apurimac department. For individuals from the rainforest ecoregion, the African component represents an average of 1.9% of their autosomal ancestry. European ancestry is more evenly represented throughout all Peruvian populations sampled, except for the highland population of Puno where it is practically absent (0.6% in average). For the coastal groups, proportions of European ancestry range from 11.1% in the Ica department to 24.3% in Tumbes, whereas in rainforest populations, European ancestry represents 16.8%. At  $K=5$ , an additional ancestry component (shown in black in [supplementary fig. S3, Supplementary Material](#) online) appears in most Peruvian populations, displaying a North–South cline, as it is predominant in the coast and rainforest ecoregions, intermediate in the central Ancash and Apurimac highland departments, and marginal in Cusco and Puno (2.8% and 0.9%, respectively). At  $K=7$ , this Peruvian component is further divided, slightly separating the coastal individuals (shown in black) from the rainforest individuals (shown in light green).

Next, we repeated the PCA including only Peruvian populations, and the PCA and ADMIXTURE analyses masking all non-Native American components (see details in [supplementary note, Supplementary Material](#) online). The PCA on the unmasked Peruvian samples does not show a clear differentiation pattern across ecoregions. However, after the masking procedure, the following two patterns emerge: 1) the first component differentiates the rainforest from the coast and highland samples and 2) a gradient from the coast to the highlands is apparent in the second component ([supplementary note, Supplementary Material](#) online; [fig. 1B](#)). The ADMIXTURE analysis on the masked dataset is in concordance with these findings, even if some sharing of the coast component is also detected in the Loreto and San Martin rainforest departments at  $K=7$  ([supplementary note, Supplementary Material](#) online; [fig. 1C](#)). On the contrary, an IBD network analysis did not show any clear clustering by ecoregion ([supplementary fig. S4, Supplementary Material](#) online), in agreement with the Analysis of Molecular Variance (AMOVA), which showed that the fraction of variation among ecoregions was similar to that found among departments within each ecoregion (1.33% vs. 1.93%, both  $P$ -values = 0.001). Thus, even if ecoregions are not a clear distinctive unit of differentiation, a subtle degree of genetic substructure corresponding to the ecoregions can be recognized.

## Candidate Genes for Positive Selection

Since differential environmental pressures exist between the three Peruvian ecoregions under study, we grouped all newly compiled Peruvian samples by ecoregion and performed three complementary tests of recent positive selection: the PBS, based on allele frequency differences between two closely related population groups with respect to an outgroup; the iHS based on identifying unusual extended haplotypes within a population to detect incomplete or ongoing sweeps ([Voight et al. 2006](#)) and the XP-EHH, also based on haplotypes but better suited for detecting complete (or near complete) selective sweeps when comparing populations ([Sabeti et al. 2007](#)). For each selection test, the top 50 signals for positive selection were identified following an empirical outlier approach (see details in Materials and Methods). We then annotated all genes present in these candidate regions with ANNOVAR ([Wang et al. 2010](#)), and we used GeneCards ([Stelzer et al. 2016](#)) and UniProt ([The UniProt Consortium 2019](#)) to characterize their functions. We only considered a signal of positive selection to be specific for each ecoregion when it was not present in the top 50 candidate regions of the other ecoregions. From a total of 118 candidate regions identified within the top 10 signals of each selection statistic and ecoregion, 22 are specific for highland populations, 21 for the coast, and 23 for the rainforest ([tables 2–4; fig. 2](#)). The remaining candidate regions are shared between at least two of the ecoregions ([supplementary table S3, Supplementary Material](#) online; for detailed annotations on the SNPs within each top region—including position, CADD-score, allele frequencies, overlapping genes, and local ancestry deviations—see [supplementary tables S4–S15, Supplementary Material](#) online). We note, however, that since each selection statistic captures different features of the selective sweep and since different pairwise ecoregion comparisons have been used, the candidate regions displaying outlier values for more than one statistic and consistently across comparisons should be considered the more reliable.

Within the top 10 signals of positive selection identified exclusively in Andean highlanders, we detect genes related to heart development (*TLL1*) and cardiac function (*SGCD*), glucose (*GLIS3*) and glycogen metabolism (*PPP1R3C*), as well as a gene belonging to the HIF pathway (*TGFA*), among others ([table 2](#) and [fig. 2A](#)). However, additional candidate genes related to the same biological functions were further identified among the top 50 strongest signals of positive selection shared between highlanders and populations from other ecoregions ([fig. 2D](#) and [supplementary table S3, Supplementary Material](#) online). Among these, we identify *DUSP27* and *TBX5*, which are involved in heart development and detected in the rainforest ecoregion with different strengths; and *GANC*, which codifies a key enzyme in glycogen metabolism and is detected within the top 50 signals in the coast. Other interesting genes with strong signals of positive selection in highlands but also detected as top 50 candidates in the

**Table 2.** Top 10 Selection Signals per Statistic and Population Comparison Exclusively Found in the Highland Ecoregion

Candidate region	Candidate genes	PBS	iHS	XP-EHH	Nat Am LAP (%)	Nat Am LAD (SDs)	RS id— Ref. allele	Allele freq. in H	Allele freq. in C	Allele freq. in R
1:111740082– 111944073	<i>CHIA, PIFO, PGCP1</i>	—	H5*	—	88.03	0.53	rs2275254-T	0.89	0.84	0.76
1:178563987– 178671750	<i>C1orf220, MIR4424, RALGPS2</i>	HC8	—	—	93.62	1.68	rs1122579-T	0.15	0.42	0.41
1:184029993– 184283674	Intergenic— <i>TSEN15, C1orf21</i>	HC5 HR	—	—	94.15	1.90	rs12119930-A	0.21	0.5	0.43
2:146944278– 147020927	Intergenic— <i>PABPC1P2</i>	HC1	—	—	90.43	0.35	rs2016340-G	0.72	0.58	0.65
2:70187173– 72306479	<i>FAM136A, TGFA, TGFA-IT1</i>	HR5 HC	—	H (HR)	92.02	1.02	rs6714409-G	0.24	0.49	0.61
4:166597189– 167143385	<i>TLL1</i>	—	—	H3 (HC)	92.02	1.02	rs1995126-A	0.94	0.74	0.94
4:178992130– 179614998	Intergenic	—	H	H1 (HR)* H2 (HC)	88.83	0.31	rs2702432-G	0.54	0.73	0.57
4:61626653– 61950476	<i>MIR548AG1</i>	—	—	H8	90.96	0.58	rs7699903-G	0.18	0.28	0.35
5:11641039– 11742668	<i>CTNND2</i>	HC2	—	—	90.43	0.35	rs4702813-G	0.95	0.72	0.93
5:154745227– 155568469	Intergenic— <i>SGCD</i>	HC9 HR1*	H	H8 (HR)	91.49–90.43	0.8–0.35	rs1432723-G	0.11	0.34	0.48
5:29886971– 30137291	Intergenic— <i>LOC105374704</i>	HC6	—	—	94.68	2.12	rs10940848-C	0.14	0.44	0.19
5:86932232– 87197535	Intergenic— <i>CCNH, TMEM161B</i>	—	—	H7 (HC)	90.96	0.58	rs710375-T	0.84	0.69	0.72
6:153600685– 153955693	<i>MIR5641-2</i>	HR6 HC	—	—	90.43	0.35	rs1221930-G	0.22	0.49	0.63
9:18046709– 18349993	Intergenic— <i>SH3GL2, ADAMTSL1</i>	HR10 HC	—	—	95.74	2.56	rs10810942-A	0.97	0.84	0.78
9:3962727– 4529671	<i>GLIS3, SLC1A1</i>	HC4 HR	H6	H1 (HC)* H6 (HR)	94.15	1.90	rs7024944-A	0.90	0.78	0.85
10:93369096– 93542186	<i>PPP1R3C, TNKS2-AS1</i>	HR2	—	—	91.49	0.80	rs150183914-T	0.06	0.16	0.28
11:134196849– 134622517	Intergenic— <i>LOC283177, LOC100507548</i>	—	—	H10 (HC)	91.49	0.80	rs10750576-A	0.90	0.75	0.81
12:47312454– 47985899	<i>PCED1B, LOC105369747, MIR4494, VDR</i>	HC HR	—	H4 (HR) H (HC)	90.96	0.58	rs855185-G	0.25	0.49	0.56
14:21924207– 22160553	<i>RAB2B, METTL3, SALL2, OR10G3, OR10G2</i>	HC3	—	—	93.09	0.90	rs1263807-T	0.14	0.36	0.30
18:72073738– 72108787	<i>C18orf63, LINC01922, FAM69C</i>	HR4	—	—	94.15	1.90	rs377380065-T	0.04	0.16	0.22
21:38251558– 39104180	<i>DSCR3, DYRK1A</i>	—	—	H9 (HC) H (HR)	88.83	0.31	rs11911146-T	0.36	0.55	0.48
22:45421536– 45527815	<i>PHF21B, NUP50-AS1</i>	—	—	H8 (HC)	92.55	1.24	rs738548-A	0.80	0.53	0.54

For each candidate region, the Native American local ancestry proportion (LAP) is included together with the corresponding local ancestry deviation (LAD). The allele frequency per ecoregion (H, C, and R for highlands, coast and rainforest, respectively) for the reference allele of a top outlier SNP within each candidate region is also indicated. When the region is identified by multiple tests, the top SNP is taken from the highest ranked and marked with \*. Genes in intergenic candidate regions are only shown when their distance to the peak is <500 kbp. In bold, top 10 selection signals with the number representing the ranking of that signal; otherwise additional hits detected within the top 50 signals in the highlands.

PBS, Population Branch Statistic; iHS, integrated Haplotype Score; XP-EHH, cross-population Extended Haplotype Homozygosity; HC, highlands to coast comparison; HR, highlands to rainforest comparison; H (HC), Highlands signal for the XP-EHH highlands to coast comparison; H (HR), Highlands signal for the XP-EHH highlands to rainforest comparison; H, highlands.

other ecoregions are related to thrombophilia (*SERPINE1*), cardiovascular development (*PLXNA4*), skin pigmentation (*MITF*), and the immune system (*CD40*) (supplementary table S3, Supplementary Material online).

Among the top 10 strongest signals of positive selection exclusively identified in coastal or rainforest populations, we mostly identified genes related to the immune system. These include different *SIGLEC* genes and the *CD44* and *ICAM1* genes in the coast, as well as the *ALCAM-CBLB* genomic region and the *CARD8* gene in the rainforest ecoregion (tables 3 and 4 and fig. 2B and C). An additional strong selection signal around the *PRDM1* gene, probably

related to the immune response, was detected among the top 10 signals in the rainforest but also within the top 50 in highlands (supplementary table S3, Supplementary Material online). Moreover, at least three further candidate genes related to the immune response were identified as top 10 signals in both rainforest and coastal populations (*FBXO40-HCLS1*, *BDR2*, *HLA-DOA*, *HLA-DPA1*, and *RNF220*).

Other top 10 selection signals were shared across regions (fig. 2D and supplementary table S3, Supplementary Material online). One of the strongest signals detected in the three ecoregions with the iHS statistic includes two candidate genes involved in lipid metabolism (*CPT2* and *LRP8*).

**Table 3.** Top 10 Selection Signals per Statistic and Population Comparison Exclusively Found in the Coast Ecoregion

Candidate region	Candidate genes	PBS	iHS	XP-EHH	Nat Am LAP (%)	Nat Am LAD (SDs)	RS id— Ref. allele	Allele freq. in H	Allele freq. in C	Allele freq. in R
2:227205102–227428810	Intergenic— <i>LOC646736, MIR5702</i>	CR	—	<b>C8 (RC)</b>	78.77	0.81	rs9789638-A	0.80	0.84	0.65
2:231044200–231429220	<i>SP110, SP140, SP140L, SP100</i>	—	—	<b>C6 (HC)</b>	81.51	1.63	rs58941251-C	0.63	0.88	0.81
2:98629966–100380563	<i>TSGA10, LIPT1</i>	CH8	—	—	72.60	1.05	rs2632277-C	0.76	0.48	0.65
3:195477791–196737295	<i>PIGZ, MELTF</i>	CH3	—	—	76.03	0.02	rs544688-G	0.48	0.72	0.74
5:1875037–1948519	<i>CTD-2194D22.4</i>	CH10	—	—	82.19	1.84	rs200756822-G	0.53	0.33	0.46
5:63309892–64417806	<i>RNF180, RGS7BP</i>	CR3	—	—	78.77	0.81	rs16892721-A	0.51	0.53	0.85
7:101060777–101449071	<i>COL26A1</i>	—	<b>C3</b>	—	76.03	0.02	rs28759973-T	0.81	0.77	0.74
8:10390452–10485154	<i>PRSS55, RP1L1</i>	—	—	<b>C4 (HC)</b>	73.29	0.84	rs150931842-G	0.85	0.97	0.94
8:82335354–82849452	<i>ZFAND1, CHMP4C, SNX16</i>	CH5	—	—	73.29	0.84	rs11991098-A	0.51	0.32	0.46
8:96200507–96223793	<i>PLEKHF2, LINC01298</i>	—	<b>C7</b>	—	69.18	2.08	rs77609822-C	0.95	0.89	0.91
9:109802363–109982588	Intergenic— <i>LOC340512, RAD23B</i>	CR	—	<b>C4 (RC)</b>	75.34	0.22	rs12551497-A	0.71	0.87	0.69
10:63248358–63583070	<i>TMEM26-AS1, C10orf107</i>	CR10	—	—	78.77	0.81	rs1456279-A	0.22	0.12	0.30
11:35123195–35360016	<i>CD44</i>	CH6 CR	—	—	81.51	1.63	rs4756196-G	0.49	0.79	0.83
11:4434519–4460261	<i>TRIM21, OR52K2</i>	CH1	—	—	82.19	1.84	rs10633520-A	0.32	0.14	0.15
13:24656407–24789809	<i>SPATA13</i>	CH4	—	—	77.40	0.39	rs60187376-T	0.18	0.11	0.19
15:63290705–63372180	<i>TLN2, TPM1, LOC100128979</i>	—	—	<b>C5 (HC)</b>	84.25	2.46	rs72741190-A	0.87	0.94	0.70
18:6667606–6776759	Intergenic— <i>LINC01387, LOC101927168</i>	CH9	—	—	71.92	1.25	rs12456358-T	0.69	0.91	0.91
19:10285806–13466988	<i>MRPL4, ICAM1</i>	CR2	—	—	71.23	1.46	rs74257295-G	0.90	0.94	0.59
19:51892016–52250216	<i>SIGLEC10, LOC100129083, SIGLEC8, CEACAM18, SIGLEC12, SIGLEC6</i>	CH	—	<b>C3 (HC)</b>	80.14	1.22	rs39711-T	0.68	0.83	0.81
19:52009560–52246157	<i>ZNF175, LINC01530, SIGLECS, SIGLEC14, SPACA6P-AS</i>	CH7	—	<b>C (CH)</b>	80.82	1.42	rs10500308-T	0.61	0.37	0.54
20:37013181–37291377	<i>ARHGAP40</i>	—	—	<b>C6 (RC)</b>	79.45	1.01	rs74983286-T	0.96	0.95	0.81
22:45116127–45367385	<i>PRRS, PRR5-ARHGAP8, ARHGAP8</i>	—	—	<b>C2 (HC)</b>	77.40	0.39	rs132410-A	0.66	0.78	0.69

For each candidate region, the Native American local ancestry proportion (LAP) is included together with the corresponding local ancestry deviation (LAD). The allele frequency per ecoregion (H, C and R for highlands, coast and rainforest, respectively) for the reference allele of a top outlier SNP within each candidate region is also indicated. Genes in intergenic candidate regions are only shown when their distance to the peak is <500 kbp. In bold, top 10 selection signals with the number representing the ranking of that signal; otherwise additional hits detected within the top 50 signals in the coast.

PBS, Population Branch Statistic; iHS, integrated Haplotype Score; XP-EHH, cross-population Extended Haplotype Homozygosity. CR, coast to rainforest comparison; CH, coast to highlands comparison; C (HC), Coast signals from the XP-EHH highlands to coast comparison; C (RC), Coast signals from the XP-EHH rainforest to coast comparison; C, coast.

Additional strong signals shared between the coast and highlands include three candidate genes, detected mainly with the PBS statistic, which encode dual oxidases (*DUOX2, DUOXA1, and DUOX1*) that are involved in thyroid hormone synthesis and the immune system; as well as the *TLR4* gene, detected with the iHS and XP-EHH statistics, and with an important recognized role in pathogen recognition and the innate immune response. As for those strong signatures shared between highlands and rainforest populations, we identified a candidate gene related to the negative regulation of hypoxic injury (*APIP*) and the aforementioned *TBX5* gene related to heart development.

Since previous studies have reported candidate genes for high-altitude adaptation in Andeans, we also analyzed the overlap with the top 50 signatures described here for the highland ecoregion (supplementary tables S16 and S17, Supplementary Material online). Out of 217 genes compiled from the literature, 24 were replicated within the top 50 signals found in highlands, with the *TBX5, TGFA* and *DUOX2* genes being among the top 10 (supplementary table S17, Supplementary Material online). Other replicated previously known candidate genes for high-altitude adaptation include the *BRINP3* and the *HAND2-AS1* genes, associated with cardiac function and related phenotypes, and the *EGLN3* gene, from the HIF pathway.



**Table 4.** Top 10 Selection Signals per Statistic and Population Comparison Exclusively Found in the Rainforest Ecoregion

Candidate region	Candidate genes	PBS	iHS	XP-EHH	Nat Am LAP (%)	Nat Am LAD (SDs)	RS id—Ref. allele	Allele freq. in H	Allele freq. in C	Allele freq. in R
1:198957867–199707262	Intergenic— <i>LINC01221</i>	—	<b>R5*</b>	<b>R5 (RC)</b> <b>R8 (HR)</b>	88.89	2.03	rs1325187-C	0.43	0.51	0.31
2:12821349–13089226	<i>TRIB2</i> , <i>LOC100506474</i>	<b>RC10</b> RH	—	—	81.48	0.42	rs973977-T	0.61	0.64	0.22
2:164399671–164879136	<i>FIGN</i>	<b>RH8 RC</b>	—	—	81.48	0.42	rs13003002-T	0.51	0.52	0.31
2:41894321–42063475	Intergenic— <i>LINC01913</i>	—	<b>R4</b>	—	79.63	0.02	rs4952511-C	0.60	0.66	0.63
2:54506070–55183537	<i>EML6</i>	—	<b>R8</b>	<b>R (RC)</b>	70.37	1.99	rs17046413-A	0.54	0.65	0.54
3:105334386–105790348	<i>ALCAM</i> , <i>CBLB</i>	—	<b>R</b>	<b>R2 (RC)</b>	87.04	1.63	rs61138958-A	0.40	0.55	0.31
3:192279892–192338861	<i>FGF12</i>	—	—	<b>R6 (RC)</b>	81.48	0.42	rs781417-C	0.88	0.78	0.89
3:59667385–59821071	<i>FHIT</i>	<b>RH6 RC</b>	—	—	83.33	0.82	rs1683366-G	0.71	0.66	0.41
6:119335865–119655145	<i>FAM184A</i>	—	<b>R6</b>	—	72.22	1.59	rs612607-G	0.74	0.82	0.91
7:116617532–117003695	<i>ST7</i> , <i>WNT2</i>	<b>RH7</b>	—	<b>R7 (HR)*</b> <b>R (RC)</b>	83.33	0.82	rs4612282-T	0.60	0.73	0.81
8:59017474–59194917	<i>FAM110B</i>	—	—	<b>R1 (RC) R (RH)</b>	87.04	1.63	rs7843838-A	0.49	0.56	0.33
9:73915883–74044936	Intergenic— <i>TRPM3</i> , <i>TMEM2</i>	<b>RH5*</b> <b>RC7</b>	—	—	85.19	1.22	rs147846688-G	0.84	0.84	0.52
10:31860345–32054109	Intergenic— <i>ZEB1</i> , <i>ARHGAP12</i>	<b>RC</b>	—	<b>R10 (RC)</b>	85.19	1.22	rs796159-T	0.46	0.51	0.74
12:97635817–97737126	Intergenic— <i>NEDD1</i> , <i>RMST</i>	<b>RC9</b>	—	—	74.07	1.19	rs6538791-A	0.61	0.68	0.46
13:29222286–29411957	Intergenic— <i>SLC46A3</i> , <i>MTUS2</i>	<b>RC4*</b> <b>RH4</b>	—	—	79.63	0.02	rs17561728-A	0.69	0.80	0.44
13:97027560–97323933	<i>HS6ST3</i>	—	—	<b>R10 (HR)</b>	85.19	1.22	rs643765-T	0.57	0.60	0.76
14:32389856–32488986	<i>LINC02313</i>	<b>RH10</b> <b>RC</b>	<b>R</b>	—	79.63	0.02	rs4640079-G	0.62	0.59	0.39
15:92467322–92717179	<i>SLCO3A1</i>	<b>RH1*</b> <b>RC1</b>	—	—	77.78	0.38	rs371469142-C	0.17	0.12	0.04
16:50361170–50505628	<i>BRD7</i> , <i>LINC02178</i>	<b>RC5 RH</b>	—	—	81.48	0.42	rs142711401-G	0.78	0.86	0.50
17:9536109–9724255	<i>USP43</i> , <i>DHRS7C</i>	—	—	<b>R7 (RC)</b>	83.33	0.82	rs380596-G	0.79	0.87	0.98
18:65485121–65528209	<i>LOC643542</i>	<b>RC2</b>	—	—	85.19	1.22	rs12454152-G	0.72	0.84	0.70
19:48622545–48935653	<i>CARD8</i> , <i>CARD8-AS1</i> , <i>ZNF114</i>	—	—	<b>R9 (HR)</b>	77.78	0.38	rs4801753-G	0.59	0.72	0.76
21:43444731–43454614	<i>ZNF295-AS1</i>	<b>RH2*</b> <b>RC3</b>	—	—	79.63	0.02	rs2839444-C	0.84	0.80	0.57

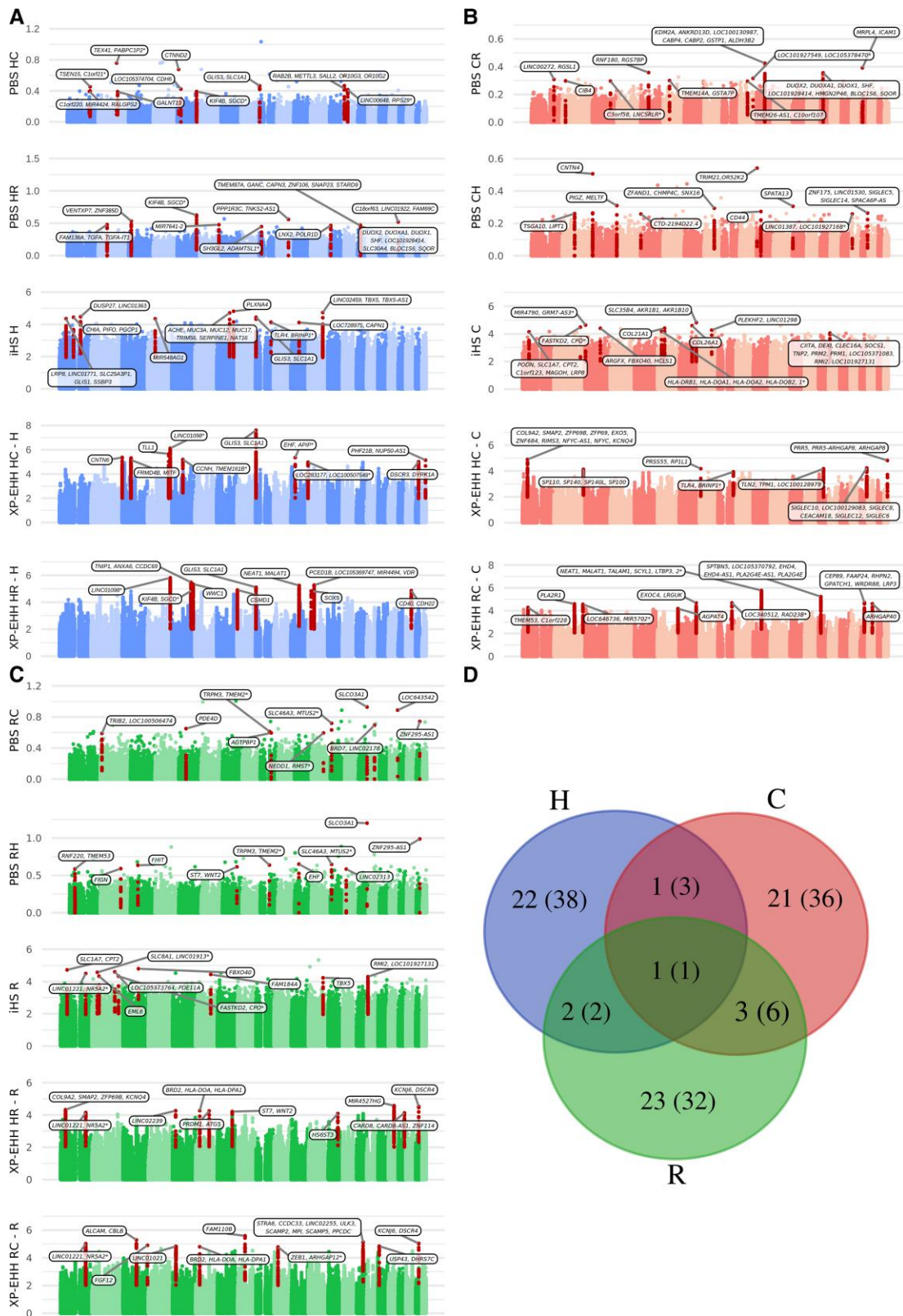
For each candidate region, the Native American local ancestry proportion (LAP) is included together with the corresponding local ancestry deviation (LAD). The allele frequency per ecoregion (H, C and R for highlands, coast and rainforest, respectively) for the reference allele of a top outlier SNP within each candidate region is also indicated. When the region is identified by multiple tests, the top SNP is taken from the highest ranked and marked with \*. Genes in intergenic candidate regions are only shown when their distance to the peak is <500 kbp. In bold, top 10 selection signals with the number representing the ranking of that signal; otherwise additional hits detected within the top 50 signals in the rainforest ecoregion.

PBS, Population Branch Statistic; iHS, integrated Haplotype Score; XP-EHH, cross-population Extended Haplotype Homozygosity. RH, rainforest to highlands comparison; RC, rainforest to coast comparison; R (HR), Rainforest signal from the XP-EHH highlands to rainforest comparison; R (RC), Rainforest signal from the XP-EHH rainforest to coast comparison; R, rainforest.

### Over-representation Analysis in Candidate Regions for Positive Selection

For each test and ecoregion, we also conducted gene-set (supplementary table S18, Supplementary Material online) and trait-associated SNP (supplementary table S19, Supplementary Material online) over-representation analyses on the top 50 candidate regions for positive selection, using the DAVID and TraseR tools, respectively.

In highland populations, the top 50 highest scoring signals for positive selection present a clear over-representation of genes related to sensory perception (PBS highlands vs. coast), fatty acid metabolism (PBS highlands vs. rainforest), xenobiotic metabolic process and drug metabolism (PBS highlands vs. rainforest and XP-EHH highlands vs. rainforest), as well as many immune-related terms, including immunoglobulin production involved in



**Fig. 2.** Manhattan plots representing the scores of each statistic for selection tests performed on highland (A), coastal (B), and rainforest (C) populations. In red, the top 10 significant candidate regions identified in each test. H, Highlands; C, Coast; R, Rainforest. (D) Number of top signals of positive selection detected with the PBS, his, and XP-EHH statistics per each ecoregion. Genomic regions are considered under selection exclusively in an ecoregion when they are not found within the top 50 signals obtained by tests performed on other ecoregions. In brackets, total number of top 10 signals detected without filtering out those also within the top 50 of the remaining ecoregions. \*Indicates that the highest scoring SNPs within a peak are in an intergenic region. \*1 in (B): HLA-DRB1, HLA-DQA1, HLA-DQA2, HLA-DQB2, HLA-DOB, TAP2, PSMB8, PSMB8-AS1, TAP1, PSMB9, LOC100294145, HLA-DMB, HLA-DMA, BRD2, HLA-DOA, HLA-DPA1. \*2 in (B): SSSCA1-AS1, EHBP1L1, KCNK7, MAP3K11, PCNX3, RELA, KATS, RNASEH2C, AP5B1, MIR1234, OVOL1-AS1, OVOL1, SNX32, CFL1, MUS81, EFEMP2, FIBP.

immunoglobulin-mediated immune response, antigen processing and presentation of peptide or polysaccharide antigen via MHC class II, innate immune response and anti-bacterial humoral response (iHS and XP-EHH highlands vs. rainforest). Moreover, we identify a significant over-representation of SNPs associated with traits related to body mass index and body weights and measures (mostly with PBS), to stroke and cardiovascular diseases (with PBS highlands vs. rainforest and iHS), to diabetes mellitus type 1 and metabolic syndrome X, to nutritional and metabolic diseases (with iHS), to gout (with XP-EHH highlands vs. rainforest), and also to immune system disease and bacterial infections and mycoses (with iHS).

In coastal populations, the gene-set over-representation identifies groups of terms associated with fatty acid transport and triglyceride catabolism (PBS coast vs. highlands), thyroid synthesis and response to oxidative stress (PBS coast vs. rainforest), as well as to several immune-related terms (antigen processing and presentation of peptide or polysaccharide antigen via MHC class II with iHS, adaptive immune response with iHS and XP-EHH highlands vs. coast, and antibacterial humoral response and innate immune response with XP-EHH rainforest vs. coast). As for the identified over-represented traits, they include body height, body weights and measures and hypertrophy, left ventricular hypertrophy (with PBS coast vs. highlands) and diabetes mellitus type 1, digestive system disease, respiratory tract disease, immune system and cardiovascular disease (with iHS).

In the rainforest ecoregion, we detect a similar over-representation of terms associated with the immune system (such as antigen processing and presentation of peptide or polysaccharide antigen via MHC class II with PBS rainforest vs. coast and with iHS), but also with xenobiotics and fatty acids (PBS rainforest vs. highlands), as well as with arrhythmogenic right ventricular cardiomyopathy and hypertrophic cardiomyopathy (with iHS), and with steroid hormone biosynthesis and metabolism of lipids (with XP-EHH rainforest vs. coast). Similarly, we identified enriched traits related to lipoproteins (PBS rainforest vs. highlands), but also to diabetes mellitus type 1 and metabolic syndrome, nutritional and metabolic diseases, cardiovascular diseases and myocardial infarction, immune system diseases, and bacterial infections and mycoses (mostly with PBS rainforest vs. coast and with iHS).

### Adaptation to Environmental Pressures

Next, we used the Samβada software (Stucki *et al.* 2017) to identify genotype frequency changes correlating with elevation as a proxy for any associated environmental pressure, while controlling for population structure. After Bonferroni correction for multiple testing on the *P*-values obtained from the Wald score, we obtained 1,937 SNPs, assigned to 1,118 genes as annotated using Variant Effect Predictor (VEP) (McLaren *et al.* 2016), whose genotypes significantly correlated with elevation (supplementary tables S20 and S21, Supplementary Material online).

A gene-set over-representation analysis on these genes revealed several terms related to xenobiotic metabolism in the highest scoring cluster, but also to dilated cardiomyopathy, cardiac muscle contractions, platelet homeostasis, and type II diabetes mellitus and regulation of insulin secretion, among others (supplementary table S22, Supplementary Material online). This result suggests the existence of concerted shifts of allele frequencies correlated with elevation in genes associated with xenobiotic metabolism as well as with cardiovascular and metabolic traits, and is consistent with the over-represented clusters previously identified for the top candidate regions for selection in the highlands ecoregion. Some of the genes detected in the Samβada analysis are also found among the top 10 signals of selection identified in each ecoregion, including 18 candidate genes for highlands, 12 for coast, and 14 for rainforest ecoregions (supplementary table S23, Supplementary Material online). Among them, we find some of the aforementioned candidate genes for positive selection, such as *SGCD*, *DUOX1-DUOX1A1*, *RNF220*, *CBLB*, *APIP*, *DUSP27*, and *GLIS3*. These results suggest elevation (or any correlated environmental variable) as the primary selective pressure behind the genomic signatures of positive selection detected. For example, the TT genotype at an intronic SNP in *GLIS3* (rs4567095) is highly correlated with elevation. This genotype is found at low frequencies in coastal populations and conversely, at higher frequencies in the highlands (supplementary fig. S5, Supplementary Material online).

### Impact of Admixture on Positive Selection

We also investigated whether the detected signals of positive selection could have been facilitated by the recent admixture from European and African continental populations. After computing the local ancestry proportions of the Indigenous American, European, African, and East Asian ancestral components per SNP in each ecoregion, we did not identify any significant local standard deviations from the global ancestry proportions, meaning that none of the identified selection signals can be directly attributed to past admixture events with European or African populations (see details in supplementary tables S3–S15 and S24, Supplementary Material online).

To ensure the robustness of our selection scan results to potential biases resulting from external (African or European) recent admixture, we also recalculated the three selection tests after masking all haplotypes with non-Native American ancestry from the dataset (see details in supplementary note, Supplementary Material online). The masking procedure significantly decreased the power to detect signals of selection because of the reduction of the number of available individuals, particularly in rainforest populations, and of marker density after QC. Despite that, an important proportion of SNPs within the top 10 signals identified for the original dataset were also found at least within the top 5% SNPs identified for the masked version, indicating a consistent overlap of

selection signatures ([supplementary note](#), [Supplementary Material](#) online; [table 2](#)).

## Discussion

We have gathered and analyzed SNP genome-wide data for 286 individuals distributed across three differentiated ecoregions in Peru to investigate signals of recent positive selection specific to populations from the Andean highlands, the Pacific coast, and the Amazon rainforest. The compilation of SNP data from 1 KGP reference populations and subsequent PCA and ADMIXTURE analyses indicated varying proportions of European and African ancestral components across the three ecoregions, consistent with historical records of the Spanish colonization of Peru in the 16th century and the posterior arrival of African peoples through the Transatlantic Slave Trade ([Aguirre 2005](#); [D'Altroy 2014](#)). Furthermore, we detect evidence for a North–South gradient of genetic differentiation between the northern coast and rainforest populations into the southern Andes (mostly including Cusco and Puno) that is consistent with previous analyses on Peruvian populations ([Barbieri et al. 2019](#), [Nakatsuka et al. 2020](#), [Borda et al. 2020](#), [Castro e Silva et al. 2022](#)). Additionally, when masking the non-native American genetic components, we also find some sharing of the coastal genetic component in northwestern rainforest individuals, which is consistent with the mixed ancestry profiles recently described by [Borda et al. \(2020\)](#) and [Castro e Silva et al. \(2022\)](#).

Within the top 10 signals of positive selection found in Andean highlanders, we identified multiple genes involved in heart function, including *TLL1*, known for its role in heart septation and positioning ([Clark et al. 1999](#); [Sieron et al. 2019](#)); *SGCD*, highly expressed in skeletal and cardiac muscle and associated to dilated cardiomyopathy; *PLXNA4* that encodes for plexin receptors which in turn interact with semaphorins that participate in heart and vascular morphogenesis ([Epstein et al. 2015](#)); *DUSP27*, involved in muscle and heart development ([Li 2011](#)); and *TBX5* which is annotated in the Gene Ontology (GO) under terms such as ‘heart development’, ‘bundle of His development’ and ‘atrial septum morphogenesis’, among others. Variants contained within the *TBX5* gene signal have previously been associated with multiple heart-related traits in genome-wide association studies (GWASs) including atrial fibrillation ([Christophersen et al. 2017](#)), QRS complex ([Prins et al. 2018](#)) and other electrocardiographic traits ([Holm et al. 2010](#)). Moreover, *TBX5* is one of the three candidate genes related to cardiovascular health previously identified by [Crawford et al. \(2017\)](#) in high-altitude Andeans. Similarly, *PLXNA4* has been identified as a candidate gene for adaptation to hypoxia in Tibetan dogs ([Witt and Huerta-Sánchez 2019](#)). It has been suggested that the putative adaptive variants driving the selection signatures detected on genes related to heart function traits could have a protective role against an additional lifelong stress to the cardiovascular system given the higher

concentration of red blood cells present in the bloodstreams of Andeans, and consequently, the higher blood viscosity ([Crawford et al. 2017](#)). According to our selection analyses, *TLL1* and *SGCD* are candidate genes exclusively identified in Andean highlander populations, whereas *DUSP27*, *PLXNA4*, and *TBX5* were also detected within the top 50 iHS signals in either coastal or rainforest populations. Such a pattern could have easily resulted from gene flow from the highland ecoregion towards the coastal and rainforest populations, and would be consistent with previous genetic research ([Gravel et al. 2013](#); [Chacón-Duque et al. 2018](#), [Castro e Silva et al. 2022](#)). Alternatively, we could be detecting selection signatures already present in the ancestral Indigenous South American population before the divergence of the three population groups under study.

In Andean highlanders, we also identified candidate genes for positive selection related to hypoxia such as *APIP*, which has been described to act as a negative regulator of ischemic injury ([Cho et al. 2004](#)) and *TGFA*, already described as a candidate gene for positive selection in highland populations by [Bigham et al. \(2009\)](#) and [Bigham et al. \(2010\)](#). Other top 10 signals of selection in Andeans are related to pigmentation and to the vitamin D receptor. *MITF*, a gene involved in melanocyte development, was identified within the top 10 signals for positive selection with the XP-EHH statistic when comparing highlands versus coast, and in the top 50 signals in highlands with the remaining tests of selection. As for the *VDR* gene, it was detected within a large candidate region in the top 10 signals of selection exclusively detected in highlands with the XP-EHH statistic when using the rainforest ecoregion as reference. Such a selection signal at *VDR* could result from a co-adaptation to reach optimal levels of vitamin D in an environment where strong UV radiation could have also favored darker skin pigmentation ([Missaggia et al. 2020](#)). Incident ultraviolet (UV) radiation, which is correlated with skin pigmentation ([Jablonski and Chaplin 2000](#)), is the most likely environmental variable responsible for these selection signals in Andeans, since UV radiation increases with altitude ([Blumthaler et al. 1997](#)).

Other candidate genes for selection identified in highland populations are related to glucose (*GLIS3*) and glycogen (*PPP1R3C* and *GANC*) metabolism, which could result from differential fasting glycemia compared with populations living at sea level (reviewed in [Woolcott et al. 2015](#)). In particular, *GLIS3* was identified in the top 10 signals with the PBS statistic when comparing highlands to coast, with the iHS statistic in highlands, and in highlands versus coast and highlands versus rainforest with the XP-EHH statistic. Notably, an intronic SNP in *GLIS3* (rs10974438), linked to the detected signatures, presents an elevated CADD-score (20.2) and has been associated with type 2 diabetes in multiple GWASs ([Mahajan et al. 2018](#); [Xue et al. 2018](#); [Vujkovic et al. 2020](#)). Moreover, *GLIS3* has been previously identified as a candidate for adaptation to highland environments in Tibetan populations ([Wang et al. 2011](#)). Additional candidate regions

for positive selection in Andeans include genes such as *CD40*, *TNIP1*, and *CHIA*, which play a role in the immune system and innate response (Blotta *et al.* 1996; Mikolajczak *et al.* 2004). Further research may reveal whether these adaptations occurred due to long-term exposure to the selective pressures posed by pathogens present in the Andes during the pre-Columbian period, or if they occurred more recently after exposure to novel diseases brought by Europeans (Merbs 1992; Patterson and Runge 2002; Riley 2010; O’Fallon and Fehren-Schmitz 2011; Lindo *et al.* 2016).

In the case of populations from the arid Pacific coast, many of the identified candidate genes for positive selection are involved in the regulation of the immune system. For instance, one of the top 10 signals includes several sialic acid-binding immunoglobulin-like lectin (*SIGLEC*) genes, which encode important cell surface receptors expressed on innate immune cells involved in modulating the host response (Cao and Crocker 2011). Among these, we identified *SIGLEC5*, whose encoded receptor has been shown to be targeted by group B *Streptococcus* to promote immune evasion (Carlin *et al.* 2009). Other examples of immune-related candidate genes include *ICAM1*, which participates in several pathways annotated in Reactome as ‘adaptive immune system’, ‘signaling by interleukins’ and ‘cytokine signaling in immune system’, among others; and *CD44*, encoding for a widely-expressed adhesion receptor known to be upregulated after activation of naive T lymphocytes during their responses against invading microbes (Baaten *et al.* 2010). To our knowledge, as of this writing there are no genome-wide positive selection studies conducted among populations from South American desert areas, except for those focused on the effects of arsenic as a selective pressure (Eichstaedt *et al.* 2015; Apata *et al.* 2017; Vicuña *et al.* 2019; Rocha *et al.* 2021).

In populations from the Amazon rainforest, as at the coast, most of the identified selection signals contain candidate genes with functional roles in the immune system. Among those specifically found within the top 10 signatures in rainforest populations, we detected the *ALCAM-CBLB* genomic region and the *CARD8* gene with the XP-EHH statistic when comparing rainforest with coastal and rainforest with highland populations, respectively. In particular, while the *CBLB* gene encodes a negative regulator of adaptive immune responses (Sanna *et al.* 2010), *CARD8* mediates inflammasome activation in response to pathogens and other signals (Johnson *et al.* 2020).

Despite initially focusing on candidate genes specific for each ecoregion, a significant number of them were detected in more than one ecoregion. Similarly, when extending the selection analysis to the top 50 signals, both the gene-set and the trait-associated SNP overrepresentation analyses displayed several common terms significantly enriched across ecoregions. In most cases, a candidate gene is clearly found within a top 10 signal of positive selection in a given ecoregion but it is also detected among the top 50 signals in others (mostly with

the iHS and the PBS statistics). In other instances, two or more ecoregions share strong signatures (i.e., share the same candidate within the top 10 signals of selection detected by any selection statistic). In such cases, it is difficult to discern whether the selection signal and corresponding selective pressures are shared across ecoregions or whether the shared selection signal results from gene flow across them. Moreover, it could also be the case that the driving selective pressure was already present before the divergence of these populations from a common ancestral population group, and that we are thus detecting the remnants of past selection signatures. The same scenario applies for those gene-sets and trait-associated SNPs found enriched in the top 50 candidate regions for positive selection in more than one ecoregion. Some of the top 10 shared signatures across ecoregions include the *TBX5* gene, detected with the iHS statistic in rainforest and highlands populations; the *APIP* gene also detected in highlands (with XP-EHH) and rainforest (with PBS); the *DUOX2-DUOX1-DUOX1* family of genes, that were found with PBS in both coast and highlands when compared with rainforest; the *BRD2*, *HLA-DOA*, and *HLA-DPA1* genes, detected in rainforest and coastal populations with the iHS and XP-EHH statistics, respectively; and the *CPT2* and *LRP8* genes, involved in lipid metabolism, which were detected within the top 10 iHS signals in all three ecoregions. The *DUOX2-DUOX1-DUOX1* family encodes for dual oxidases which are required for the synthesis of the thyroid hormone (and hence could be related to thermoregulation), but also play a role in antimicrobial defence at mucosal surfaces (De Deken *et al.* 2014; van der Vliet *et al.* 2018). *DUOX2* has already been reported under positive selection in Andean populations (Zhou *et al.* 2013; Jacovas *et al.* 2018; Borda *et al.* 2020). Furthermore, adaptations related to lipid metabolism have been suggested to have occurred before the dispersals of the first Indigenous Americans. Amorim *et al.* (2017) showed that the fatty acid desaturases (*FADS*) genes present signals of positive selection in Indigenous populations throughout the Americas, probably resulting from an adaptation event that took place in the common ancestral population previous to the first peopling events. Note that although we searched for deviations in local ancestry proportions, we did not find evidence of post-admixture selection in any of the signatures detected across ecoregions.

Finally, we also used the Samβada software to investigate genotype frequency changes correlating with elevation while correcting for population structure. Such a strategy not only allowed us to explore whether the top signals of positive selection detected across ecoregions significantly correlated with elevation, but also to interrogate for concerted shifts of allele frequencies on genes involved in particular biological functions as expected under a model of polygenic selection (Pritchard *et al.* 2010; Stephan 2016; Novembre and Barton 2018). Although we used elevation as an environmental variable in this analysis, we also note that the particular selective pressure driving such

allele frequency correlations could be any correlated environmental variable such as temperature (annual mean, maximum in the warmest month or minimum in the coldest month), solar radiation, and other related environmental factors (such as dryness or vegetation). Several of the top candidate genes for adaptation across ecoregions were found to match with those genes displaying significant genotype frequency correlations with elevation ([supplementary table S23, Supplementary Material](#) online), meaning that elevation (or any highly correlated environmental variable) might have been acting as a selective pressure. Moreover, several of the common gene-set terms found enriched among the top 50 signals across ecoregions were also found to be enriched among those identified by Samβada.

A limitation of this study is the use of genotype data. While we can reliably identify regions of the genome with the patterns of variation expected under positive selection, and thus propose putative candidate genes for adaptation, the use of SNP-array data does not allow the direct identification of the causal variant driving the selection signals. Therefore, for most of the top signals of positive selection detected we have focused on identifying plausible candidate genes, even if no functional variant in the suggested candidate region was interrogated in our array. Because of that, our results are limited by the current available knowledge on gene functions and biased towards particular selective pressures hypothesized as adaptive in each ecoregion. Furthermore, by comparing signals of selection across populations grouped into three ecoregions, our analysis may also overlook the impact of other localized environmental variation on human adaptation patterns. Future studies may overcome this limitation by comparing signals of selection between populations within each ecoregion, or by employing alternative frameworks to guide comparative analyses such as the 8, 11, or 13 ecoregion models proposed by geography and ecology scholars ([Vidal 2014](#); [Britto 2017](#)). It should as well be noted, that the retrieved data does not have sufficient geographic coverage to be representative of all the genetic variation present in each ecoregion. This is especially the case for the rainforest ecoregion, for which we only analyzed 27 individuals, thus limiting the interpretation of our results. A further limitation is our heterogeneous sampling strategy which combined samples collected from several studies and thus contained varying levels of ethnographic and family origin information. Future efforts could include focused sampling designs that prioritize the detailed collection of family background information of the participants, including place of birth and long-term residence, at each sampling location. Despite these limitations, the study of signatures of positive natural selection in the human genome is valuable as it allows deciphering past and ongoing selective pressures which continue to impact present-day populations. Since natural selection operates on functional variants, the adaptive events resulting from local selective pressures could in turn have impacts on functional and medically relevant traits.

In conclusion, our study investigated genetic adaptation among Peruvian populations from three distinct ecological regions: the desert Pacific coast, the Andean highlands and the Amazon rainforest lowlands. We found that among highlanders most top candidate genes for positive selection are involved in oxygen homeostasis, cardiovascular function, skin pigmentation, and lipid or glucose metabolism. In contrast, the top selection signals specific for populations living in the coast and rainforest ecoregions are mostly associated with the innate and adaptive immune response. These patterns suggest dissimilar selective pressures across populations from different ecoregions, and identify chronic altitude-associated hypoxia exposure and localized pathogen diversity as two of the main drivers of human adaptation in South America. Finally, we also found several signals of selection shared across populations from the three ecoregions. We infer that this pattern might result from historical gene flow between highland, coastal and rainforest communities, or from common adaptations present among ancestral Indigenous populations during the first dispersals into the Americas. Further detailed knowledge on the past and recent demographic events experienced by the populations inhabiting the three ecoregions studied here is required to be implemented in the appropriate selection and demography models, and to ultimately help to distinguish between selective pressures acting on more than a particular region or, conversely, shared signals originating from selection before divergence or posterior gene flow. By characterizing signals of selection among Peruvian populations from diverse ecological contexts, this work contributes to our understanding of the genetic changes underlying human adaptations to the diverse environments of the Americas.

## Materials and Methods

### Samples, Data Generation and QC

This study combines publicly available and newly generated SNP data from human populations from several coastal, highlands (defined as populations living >2,500 m above sea level) and rainforest areas of Peru ( $N = 308$ ) ([fig. 1, supplementary table S1, Supplementary Material](#) online). Sample collection and DNA extraction were completed over different time periods with informed consent and ethical review approval from the participating institutions (for details, see [supplementary table S1, Supplementary Material](#) online). Protocols for this study were approved by Institutional Review Boards at Arizona State University under IRB ID 0312002159 and IRB ID 0312001490 Genetic Diversity in Present-Day Populations of the South-Central Andes (including Ancash and Nomatsiguenga); and Stanford University under eProtocol 20839 Population and Functional Genomics of the Americas. Approval was also granted by the Scientific Research Ethics Committee of Universidad Ricardo Palma under stipulation No. 1889-2004 and by the Parc de Salut Mar Clinical Research Ethics Committee (project reference No. 2019/8916/I).

We generated genome-wide genotypes for over 1.2 million SNPs with the Illumina Infinium® Multi-Ethnic Global Array (MEGA) for 189 DNA samples collected in both rural and urban locations across Peru. These samples were genotyped across two batches in 2018 at the LABSERGEN Genomics Services Laboratory in the National Laboratory of Genomics for Biodiversity, Advanced Genomics Unit (UGA-LANGEBIO), CINVESTAV. Additionally, we included genotypes from 100 unrelated individuals from Puno, a highland city in south-eastern Peru, previously generated through the Population Architecture using Genomics and Epidemiology (PAGE) consortium (Wojcik *et al.* 2019); and from 19 Peruvian individuals from Magdalena de Cao (La Libertad department) previously published by Ioannidis *et al.* (2020). All datasets used in this study are summarized in [supplementary table S2, Supplementary Material](#) online.

We performed QC using Plink 1.9 (Chang *et al.* 2015) independently on each genotype batch or previously available dataset. For each dataset, we removed structural variants, SNPs with duplicate marker names and SNPs with no physical position in the GRCh37 human genome assembly (hg19). SNPs on the reverse strand were flipped to the forward strand using `snpflip` (<https://github.com/biocore-ntnu/snpflip>). SNPs with ambiguous strandedness were removed. We restricted our analyses to autosomal SNPs, removed all variants with missing call rates over 5% (with plink command `-geno 0.05`) and filtered all SNPs failing a Hardy Weinberg equilibrium (HWE) exact test with  $P$ -values below  $10^{-8}$  (`-hwe 1e-8`). We also filtered each dataset for duplicates and first-degree relatives by eliminating individuals with IBD values  $>0.5$  as described in Anderson *et al.* (2010). Finally, individual samples with missing call frequencies over 10% (`-mind 0.10`) were also removed. After QC, each dataset had between 1.2 and 1.5 million autosomal SNPs (see details in [supplementary table S2 and supplementary fig. S1, Supplementary Material](#) online).

Next, we merged all datasets at overlapping sites using Plink 1.9 and repeated the QC filtering on the obtained merged dataset, which contained 1,036,981 autosomal SNPs and 286 individuals. No SNPs or individuals were removed when filtering variants with missing call rates over 5% (`-geno 0.05`) and individuals with missing call frequencies over 10% (`-mind 0.10`). However, we removed 88 SNPs that failed the HWE filter (`-hwe 1e-8`) and excluded 110 monomorphic sites. The final merged dataset after QC included 1,036,783 SNPs and 286 individuals.

We then intersected the resulting dataset with a subset of 395 individuals from the 1000 Genomes Phase 3 populations (1 KGP) including YRI, CEU, CHB, and PEL (Auton *et al.* 2015). We restricted the intersected dataset to overlapping sites in the 1 KGP populations and again controlled for variant and sample missingness, as detailed above. Finally, we filtered SNPs with minor allele frequencies (MAFs) below 1% (`-maf 0.01`). The final intersected dataset after QC retained 681 individuals and 610,576 SNPs. The full QC process is detailed in [supplementary fig. S2, Supplementary Material](#) online.

Except for five individuals from the Cajamarca department, six from Arequipa, and four additional samples that could not be clearly assigned, the newly compiled Peruvian samples were allocated across three different ecoregions taking into account place of birth, long-term residence or family origin information as documented during the sample collection stage ([table 1 and supplementary table S1, Supplementary Material](#) online). Moreover, to avoid over-representation of highland individuals, we randomly subsampled 25 individuals from the original 100 collected Puno samples. After this procedure, our dataset comprised 591 individuals, including 395 individuals from reference populations in the 1 KGP and 196 newly compiled Peruvians (27 from the rainforest ecoregion, 94 from the highlands, and 75 from the coast).

### Global Ancestry Analyses

PCA was performed on the intersected dataset using SmartPCA from the EIGENSOFT 6.0.1 package (Patterson and Runge 2002). In addition, we also ran an unsupervised analysis with ADMIXTURE 1.3.0 (Alexander *et al.* 2009) to estimate and recognize potential differential ancestry patterns in the Peruvian departments when analyzed together with the YRI, CEU, CHB and PEL reference populations from the 1 KGP. Specifically, we ran 10 random seeds for  $K = 2$  to 10 ancestral components, and used PONG to represent the results (Behr *et al.* 2016). For these analyses, input data were pruned for linkage disequilibrium by applying the `-indep-pairwise` flag in Plink 1.9 with a window size of 200 variants, a step of 25 variants and a correlation threshold of  $r^2$  equal to 0.5. The resulting dataset contained 405,493 variants and 681 individuals. Next, we identified identity-by-descent (IBD) segments using IBDseq r1206 (Browning and Browning 2013). As in Dai *et al.* (2020), we only kept IBD segments  $\geq 3$  cM, and removed those segments overlapping with chromosomal regions above 1 Mb with no SNPs across all individuals. Then, we computed the cumulative IBD length between each pair of individuals. We built an IBD network where each vertex represents an individual and each edge weight represents the cumulative IBD length between the connected individuals. As described in Han *et al.* (2017), edges with cumulative IBD lengths below 12 cM were removed to reduce information about less recent demography and spurious identified IBD. Edges with cumulative IBD lengths above 72 cM were removed as well, to avoid clustering of extended families. In addition, we executed an AMOVA (Excoffier *et al.* 1992) at ecoregion and department levels in order to quantify the degree of population differentiation. For that, we used the `poppr` R package (Kamvar *et al.* 2015) and a permutation test to obtain the corresponding  $P$ -values.

### Analyses of Genomic Signatures of Positive Selection

Given that this study seeks to characterize adaptation amongst Peruvians with large proportions of Indigenous American ancestry, two individuals showing high proportions of African and European ancestries in the global ancestry analysis were not included in the selection

analyses. Specifically, these individuals were CM\_0240 from Ica (Coast) with an estimated 90% African ancestry proportion; and MDC025 from Magdalena de Cao (Coast), with an estimated 56% of European ancestry proportion (supplementary table S1, Supplementary Material online). Thus, a total of 194 Peruvian individuals including 73 samples from the coast, 94 from the highlands, and 27 from the rainforest (table 1) were analyzed to identify signals of positive selection in each ecoregion. Population Branch Statistic (PBS) values (Yi *et al.* 2010) were calculated using a self-implemented script in R based on  $F_{ST}$  values (Weir and Cockerham 1984) computed with VCFTools 0.1.14 (Danecek *et al.* 2011). We considered all pairwise comparisons among ecoregions (highlands, coast, and rainforest) and used CHB as an outgroup.

After phasing the dataset using SHAPEIT2 (Delaneau *et al.* 2012), the integrated Haplotype Score (iHS) (Voight *et al.* 2006) was computed on each ecoregion separately using selscan 1.2.0 (Szpiech and Hernandez 2014). For this analysis, genotype data was polarized according to the ancestral allele information provided for the 1 KGP dataset. We also used selscan to compute the cross-population Extended Haplotype Homozygosity (XP-EHH) test (Sabeti *et al.* 2007) on the phased data for the following ecoregion comparisons: highlands against coast, highlands against rainforest, and rainforest against coast. Normalization was computed for iHS and XP-EHH by minor allele frequency bins and genome-wide, respectively, using selscan default parameters.

Candidate regions for positive selection were then identified through an empirical outlier approach. For each statistic and ecoregion comparison, we initially identified as putative peaks of selection those genomic regions whose SNP scores are above the top 5% of the empirical distribution only if they also comprised at least three SNPs with scores above the top 1% within the surrounding  $\pm 100$  Kb. The SNPs within the 50 strongest peaks of signals of positive selection and all other SNPs in the top 1% of the distribution were annotated with ANNOVAR (Wang *et al.* 2010). For the top 10 peak signals, putative candidate genes were determined considering the location of the highest scoring SNPs. Intergenic signals were assigned to surrounding genes only when these were at a distance below 50 Kb. Putative functional SNPs within the top 10 candidate regions for positive selection were further explored by searching for expression quantitative trait loci (eQTLs) and splicing QTLs (sQTLs) in the GTEx database (GTEx Consortium 2013), and for SNP-trait associations in the GWAS catalog (Welter *et al.* 2014).

### Over-representation Analysis in Candidate Regions for Positive Selection

For each ecoregion comparison and statistic, a gene-set over-representation analysis was performed in the top 50 highest scoring regions with the functional annotation clustering tool DAVID 2021 (Huang *et al.* 2009), using

the whole genome as background. The queried databases included OMIM (Amberger *et al.* 2015), GO (Harris *et al.* 2004), KEGG (Kanehisa *et al.* 2017), and Reactome (Fabregat *et al.* 2016). Only enrichment scores above 1.3, equivalent to a  $P$ -value of 0.05, were considered. Additionally, we performed a trait-associated SNP over-representation analysis on the candidate regions comprising the 50 top strongest signals. For this, we used the R package TraseR (Chen and Qin 2016), which includes all traits annotated in dbGaP (Mailman *et al.* 2007) and the NHGRI GWAS catalog (Welter *et al.* 2014).

### Adaptation to Environmental Pressures

We also explored genotype frequency changes correlated with elevation, while correcting for population structure using the Samβada software (Stucki *et al.* 2017). For this analysis, we retrieved the mean annual temperature, maximum temperature in the warmest month, minimum temperature in the coldest month and the solar radiation per month data from the WorldClim dataset (Fick and Hijmans 2017), as well as elevation data from the Shuttle Radio Topography Mission (Farr *et al.* 2007). Because these environmental variables were greatly correlated among each other, for the analysis we only used elevation. We then included the two first principal components obtained in the analyses above as independent variables to account for population structure. After splitting all markers and corresponding genotypes for all individuals into 17 independent files of 75,000 markers, we ran Samβada in its supervision module and merged the corresponding output results. The  $P$ -value of each associated genotype in the output was computed from the Wald score (pop version), and adjusted using Bonferroni correction to account for multiple testing. Next, significant SNPs ( $P$ -value  $< 0.05$ ) were annotated using VEP (McLaren *et al.* 2016) and a gene-set over-representation analysis using DAVID 2021 was performed on the obtained candidate gene list.

### Impact of Admixture on Positive Selection

Finally, since most coastal and rainforest Peruvian populations display admixture with European and African ancestries (Ruiz-Linares *et al.* 2014; Homburger *et al.* 2015; Chacón-Duque *et al.* 2018; Harris *et al.* 2018), we investigated significant local ancestry proportion deviations (SD) that could potentially indicate post-admixture selection [ $SD > 4.42$  as seen in Bhatia *et al.* (2014)]. We used RFMix v.1.5.4 (Maples *et al.* 2013) in PopPhased mode, with a window size of 0.2 cM, and with two expectation-maximization (EM) iterations. The rest of parameters were kept as defined by default. As reference populations we used the CEU ( $n = 75$ ), CHB ( $n = 75$ ), and YRI ( $n = 75$ ) individual samples from the 1 KGP (Auton *et al.* 2015), and 75 Puno individuals from the PAGE consortium (Wojcik *et al.* 2019) as the Native American proxy. Local ancestry proportions were then computed per SNP and per ecoregion from the per individual values obtained, after processing the outputs as described in Martin *et al.* (2017).



In addition, to further assess the potential effects of recent admixture in the detection of positive selection, we repeated the selection analyses in each ecoregion after masking all alleles assigned to any ancestry other than Native American, and checked the concordance between the top candidates identified in the masked and unmasked procedures (see for details [supplementary note](#), [Supplementary Material](#) online).

## Supplementary Material

[Supplementary material](#) is available at *Molecular Biology and Evolution* online.

## Acknowledgements

We extend our deepest gratitude to all participants who donated samples for this project. We also thank the Mendoza Revilla family who provided lodging and logistics support in Lima during fieldwork. We further thank A. Obregon-Tito, C. Lewis, and R. Tito for assistance with sample collection and valuable discussions about appropriate study design, as well as J. Baker, G.L. Wojcik, C. Gignoux, and C.D. Bustamante for providing access to the PAGE dataset in advance of publication. We thank the INEN-UCDavis project directors for providing access to genotype data from 79 Peruvian individuals ahead of publication. This work was supported by the Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación (AEI) (PID2019-110933GB-I00/AEI/10.13039/501100011033 to E.B.); the Unidad de Excelencia María de Maeztu funded by the Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación (DOI: 10.13039/501100011033; ref: CEX2018-000792-M to E.B. and R.C.-C.); the National Science Foundation (NSF) SBE (Postdoctoral Research Fellowship Award No. 1711982 to M.A.N.-C.), NSF-BCS (BCS-0242958 to A.C.S.) and NSF-Research Experience for Undergraduates (BCS-0242958 to A.C.S.); the Mexican National Council for Science and Technology (CONACYT) (FONCICYT/50/2016 to A.M.-E.); and the International Center for Genetic Engineering and Biotechnology (ICGEB, Italy) (CRP/MEX15-04\_EC to A.M.-E.). The PEGEN-BC study was supported by the National Cancer Institute at the National Institutes of Health (R01CA204797 to L.F.) and the Instituto Nacional de Enfermedades Neoplásicas in Lima, Peru.

## Author Contributions

A.M.-E., E.B., R.C.-C., and M.A.N.-C. conceived the study. A.C.S., V.R.-d.-C., K.S., B.L., T.V., and L.F. collected the samples. M.A.N.-C. processed the samples with assistance from E.R., and performed the quality control of the data, including initial global ancestry analyses. R.C.-C. performed further global and local ancestry analyses and applied the selection tests, posterior annotations, and enrichments. E.B. and R.C.-C. interpreted the results with the inputs from A.M.-E. and M.A.N.-C. E.B. and R.C.-C. wrote the first draft of the paper with substantial contributions from

M.A.N.-C. A.M.-E. and E.B. provided support for data generation and overall supervision. All authors approved the final version of the manuscript.

## Data Availability

Newly reported genotype data for Peruvian individuals recruited through collaboration with Arizona State University, Stanford University, LANGEPIO Cinvestav, Universidad Ricardo Palma and Universidad Nacional Mayor de San Marcos have been deposited in the European Genome-Phenome Archive (EGA dataset accession number: EGAD00010002261 and study accession number: EGAS00001005692). In agreement with study protocols and approvals, the genotyped data for 79 Peruvian individuals recruited through the Instituto Nacional de Enfermedades Neoplásicas (INEN) and University of California Davis are only available upon request as they are part of a larger dataset at INEN-UCDavis (for early access requests contact L Fejerman at [ifejerman@ucdavis.edu](mailto:ifejerman@ucdavis.edu)).

## References

- Aguirre C. 2005. *Breve historia De la Esclavitud en El Peru: Una Herida que No deja De sangrar*. Lima: Fondo editorial del Congreso del Peru.
- Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**(9): 1655–1664.
- Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**(Database issue):D789–D798.
- Amorim CEG, Daub JT, Salzano FM, Foll M, Excoffier L. 2015. Detection of convergent genome-wide signals of adaptation to tropical forests in humans. *PLoS ONE.* **10**(4):e0121557.
- Amorim CEG, Nunes K, Meyer D, Comas D, Bortolini MC, Salzano FM, Hünemeier T. 2017. Genetic signature of natural selection in first Americans. *Proc Natl Acad Sci.* **114**(9):2195–2199.
- Anand AR, Ganju RK. 2006. HIV-1 gp120-mediated apoptosis of T cells is regulated by the membrane tyrosine phosphatase CD45. *J Biol Chem.* **281**(18):12289–12299.
- Anderson KM, Anderson DM, McAnally JR, Shelton JM, Bassel-Duby R, Olson EN. 2016. Transcription of the non-coding RNA upper-hand controls Hand2 expression and heart development. *Nature.* **539**(7629):433–436.
- Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. 2010. Data quality control in genetic case-control association studies. *Nat Protoc.* **5**(9):1564–1573.
- Apata M, Arriaza B, Llop E, Moraga M. 2017. Human adaptation to arsenic in Andean Populations of The Atacama Desert. *Am J Phys Anthropol.* **163**(1):192–199.
- Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, et al. 2015. A global reference for human genetic variation. *Nature.* **526**(7571):68–74.
- Baaten BJ, Li C-R, Bradley LM. 2010. Multifaceted regulation of T cells by CD44. *Commun Integr Biol.* **3**(6):508–512.
- Barberena R, Menéndez L, le Roux PJ, Marsh EJ, Tessone A, Novellino P, Lucero G, Luyt J, Sealy J, Cardillo M, et al. 2020. Multi-isotopic and morphometric evidence for the migration of farmers leading up to the Inka Conquest of The southern Andes. *Sci Rep.* **10**(1): 21171.

- 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. 2019. The current genomic landscape of Western South America: Andes, Amazonia, and Pacific Coast. *Mol Biol Evol.* **36**(12):2698–2713.
- Beall CM. 2007. Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *Proc Natl Acad Sci.* **104**(Suppl 1): 8655–8660.
- Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. 2016. Pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics.* **32**(18):2817–2823.
- Bhatia G, Tandon A, Patterson N, Aldrich MC, Ambrosone CB, Amos C, Bandera EV, Berndt SI, Bernstein L, Blot WJ, et al. 2014. Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. *Am J Hum Genet.* **95**(4):437–444.
- Bigham A, Bauchet M, Pinto D, Mao X, Akey JM, Mei R, Scherer SW, Julian CG, Wilson MJ, Herráez DL, et al. 2010. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLOS Genet.* **6**(9):e1001116.
- Bigham AW, Lee FS. 2014. Human high-altitude adaptation: forward genetics meets the HIF pathway. *Genes Dev.* **28**(20):2189–2204.
- Bigham AW, Mao X, Mei R, Brutsaert T, Wilson MJ, Julian G, Parra EJ, Akey JM, Moore LG, Shriver MD. 2009. Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. *Hum Genomics.* **4**(2):79–90.
- Bigham AW, Wilson MJ, Julian CG, Kiyamu M, Vargas E, Leon-Velarde F, Rivera-Chira M, Rodriguez C, Browne VA, Parra E, et al. 2013. Andean and Tibetan patterns of adaptation to high altitude. *Am J Hum Biol.* **25**(2):190–197.
- Blotta MH, Marshall JD, DeKruyff RH, Umetsu DT. 1996. Cross-linking of the CD40 ligand on human CD4+ T lymphocytes generates a costimulatory signal that up-regulates IL-4 synthesis. *J Immunol Baltim Md 1950.* **156**(9):3133–3140.
- Blumthaler M, Ambach W, Ellinger R. 1997. Increase in solar UV radiation with altitude. *J Photochem Photobiol B.* **39**(2): 130–134.
- Borda V, Alvim I, Mendes M, Silva-Carvalho C, Soares-Souza GB, Leal TP, Furlan V, Scliar MO, Zamudio R, Zolani C, et al. 2020. The genetic structure and adaptation of Andean Highlanders and Amazonians are influenced by the interplay between geography and culture. *Proc Natl Acad Sci.* **117**(51):32557–32565.
- Brandini S, Bergamaschi P, Cerna MF, Gandini F, Bastaroli F, Bertolini E, Cereda C, Ferretti L, Gómez-Carballa A, Battaglia V, et al. 2018. The Paleo-Indian entry into South America according to mitochondrial genomes. *Mol Biol Evol.* **35**(2):299–311.
- Britto B. 2017. Update of the terrestrial ecoregions of Peru proposed in the red book of endemic plants of Peru. *Gayana Botánica.* **74**(1):15–29.
- Browning BL, Browning SR. 2013. Detecting identity by descent and estimating genotype error rates in sequence data. *Am J Hum Genet.* **93**(5):840–851.
- Caignard G, Leiva-Torres GA, Leney-Greene M, Charbonneau B, Dumaine A, Fodil-Cornu N, Pyzik M, Cingolani P, Schwartzentruber J, Dupaul-Chicoine J, et al. 2013. Genome-wide mouse mutagenesis reveals CD45-mediated T cell function as critical in protective immunity to HSV-1. *PLOS Pathog.* **9**(9):e1003637.
- Cao H, Crocker PR. 2011. Evolution of CD33-related siglecs: regulating host immune functions and escaping pathogen exploitation? *Immunology.* **132**(1):18–26.
- Cárdenas-Arroyo F, Bray TL. 1998. *Intercambio y Comercio entre Costa, Andes y Selva: Arqueología y Etnohistoria de Suramérica.* Bogotá: Departamento de Antropología, Universidad de los Andes.
- Carlin AF, Chang Y-C, Areschoug T, Lindahl G, Hurtado-Ziola N, King CC, Varki A, Nizet V. 2009. Group B streptococcus suppression of phagocyte functions by protein-mediated engagement of human siglec-5. *J Exp Med.* **206**(8):1691–1699.
- Castro e Silva MA, Ferraz T, Couto-Silva CM, Lemes RB, Nunes K, Comas D, Hünemeier T. 2022. Population histories and genomic diversity of South American natives. *Mol Biol Evol.* **39**(1):msab339.
- Chacón-Duque J-C, 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. 2018. Latin Americans show wide-spread converso ancestry and imprint of local native ancestry on physical appearance. *Nat Commun.* **9**:5388.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience.* **4**(1):7.
- Chen L, Qin ZS. 2016. traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals: Table 1. *Bioinformatics.* **32**(8):1214–1216.
- Cheng X, Jiang H. 2019. Long non-coding RNA HAND2-AS1 downregulation predicts poor survival of patients with end-stage dilated cardiomyopathy. *J Int Med Res.* **47**(8):3690–3698.
- Cho D-H, Hong Y-M, Lee H-J, Woo H-N, Pyo J-O, Mak TW, Jung Y-K. 2004. Induced inhibition of ischemic/Hypoxic injury by APIP, a novel Apaf-1-interacting protein. *J Biol Chem.* **279**(38): 39942–39950.
- Christophersen IE, Rienstra M, Roselli C, Yin X, Geelhoed B, Barnard J, Lin H, Arking DE, Smith AV, Albert CM, et al. 2017. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat Genet.* **49**(6):946–952.
- Clark TG, Conway SJ, Scott IC, Labosky PA, Winnier G, Bundy J, Hogan BL, Greenspan DS. 1999. The mammalian toll-like 1 gene, Tll1, is necessary for normal septation and positioning of the heart. *Dev Camb Engl.* **126**(12):2631–2642.
- Crawford JE, Amaru R, Song J, Julian CG, Racimo F, Cheng JY, Guo X, Yao J, Ambale-Venkatesh B, Lima JA, et al. 2017. Natural selection on genes related to cardiovascular health in high-altitude adapted Andeans. *Am J Hum Genet.* **101**(5):752–767.
- Dai CL, Vazifeh MM, Yeang C-H, Tachet R, Wells RS, Vilar MG, Daly MJ, Ratti C, Martin AR. 2020. Population histories of the United States revealed through fine-scale migration and haplotype analysis. *Am J Hum Genet.* **106**(3):371–388.
- D’Altroy TN. 2014. *The Incas.* 2nd ed. Hoboken (NJ): Wiley.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics.* **27**(15): 2156–2158.
- Dawes R, Hennig B, Irving W, Petrova S, Boxall S, Ward V, Wallace D, Macallan DC, Thursz M, Hill A, et al. 2006. Altered CD45 expression in C77G carriers influences immune function and outcome of Hepatitis C infection. *J Med Genet.* **43**(8):678–684.
- De Deken X, Corvilain B, Dumont JE, Miot F. 2014. Roles of DUOX-mediated hydrogen peroxide in metabolism, host defense, and signaling. *Antioxid Redox Signal.* **20**(17):2776–2793.
- Delaneau O, Marchini J, Zagury J-F. 2012. A linear complexity phasing method for thousands of genomes. *Nat Methods.* **9**(2): 179–181.
- De Queiroz JS, Silva F, Ipenza C, Hernick C, Batallanos L, Griswold D, Rogers AE. 2014. Peru tropical forest and biodiversity assessment. US Foreign Assistance Act, Section 118/119 Report. USAID.
- de Souza JG, Schaan DP, Robinson M, Barbosa AD, Aragão LEOC Jr, Marimon BH, Marimon BS, da Silva IB, Khan SS, Nakahara FR, et al. 2018. Pre-columbian earth-builders settled along the entire southern rim of the Amazon. *Nat Commun.* **9**(1):1125.
- Dillehay TD, Ramírez C, Pino M, Collins MB, Rossen J, Pino-Navarro JD. 2008. Monte verde: seaweed, food, medicine, and the peopling of South America. *Science.* **320**(5877):784–786.
- Eichstaedt CA, Antao T, Cardona A, Pagani L, Kivisild T, Mormina M. 2015. Positive selection Of AS3MT to arsenic water in Andean populations. *Mutat Res.* **780**:97–102.
- Epstein JA, Aghajanian H, Singh MK. 2015. Semaphorin signaling in cardiovascular development. *Cell Metab.* **21**(2):163–173.
- Escobar G, Beall CM. 1982. Contemporary patterns of migration in the Central Andes. *Mt Res Dev.* **2**(1):63–80.

- Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*. **131**(2):479–491.
- Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, et al. 2016. The reactome pathway knowledgebase. *Nucleic Acids Res.* **44**(D1):D481–D487.
- Fan S, Hansen MEB, Lo Y, Tishkoff SA. 2016. Going global by adapting local: A review of recent human adaptation. *Science*. **354**(6308):54–59.
- Farr TG, Rosen PA, Caro E, Crippen R, Duren R, Hensley S, Kobrick M, Paller M, Rodriguez E, Roth L, et al. 2007. The shuttle radar topography mission. *Rev Geophys*. **45**(2):1–33.
- Fehren-Schmitz L, Haak W, Mächtle B, Masch F, Llamas B, Cagigao ET, Sossna V, Schitteck K, Cuadrado JI, Eitel B, et al. 2014. Climate change underlies global demographic, genetic, and cultural transitions in pre-Columbian southern Peru. *Proc Natl Acad Sci*. **111**(26):9443–9448.
- Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol*. **37**(12):4302–4315.
- Fuselli S, Tarazona-Santos E, Dupanloup I, Soto A, Luiselli D, Pettener D. 2003. Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. *Mol Biol Evol*. **20**(10):1682–1691.
- García-Longoria L, Muriel J, Magallanes S, Villa-Galarce ZH, Ricopa L, Inga-Díaz WG, Fong E, Vecco D, Guerra-Saldaña C, Salas-Rengifo T, et al. 2022. Diversity and host assemblage of avian haemosporidians in different terrestrial ecoregions of Peru. *Curr Zool*. **68**(1):27–40.
- Goldberg A, Mychajliw AM, Hadly EA. 2016. Post-invasion demography of prehistoric humans in South America. *Nature*. **532**(7598):232–235.
- Gómez-Carballa A, Pardo-Seco J, Brandini S, Achilli A, Perego UA, Coble MD, Diegoli TM, Álvarez-Iglesias V, Martínón-Torres F, Olivieri A, et al. 2018. The peopling of South America and the trans-Andean gene flow of the first settlers. *Genome Res*. **28**(6):767–779.
- Gravel S, Zakharia F, Moreno-Estrada A, Byrnes JK, Muzzio M, Rodriguez-Flores JL, Kenny EE, Gignoux CR, Maples BK, Guiblet W, et al. 2013. Reconstructing native American Migrations from whole-genome and whole-exome data. *PLoS Genet*. **9**(12):e1004023.
- GTE Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. **45**(6):580–585.
- Guernier V, Hochberg ME, Guégan J-F. 2004. Ecology drives the worldwide distribution of human diseases. *PLOS Biol*. **2**(6):e141.
- Guyton AC, Richardson TQ. 1961. Effect of hematocrit on venous return. *Circ Res*. **9**(1):157–164.
- Han E, Carbonetto P, Curtis RE, Wang Y, Granka JM, Byrnes J, Noto K, Kermany AR, Myres NM, Barber MJ, et al. 2017. Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nat Commun*. **8**(1):14238.
- Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. **32**(Database issue):D258–D261.
- Harris DN, Song W, Shetty AC, Levano KS, Cáceres O, Padilla C, Borda V, Tarazona D, Trujillo O, Sanchez C, et al. 2018. Evolutionary genomic dynamics of Peruvians before, during, and after the Inca empire. *Proc Natl Acad Sci U S A*. **115**(28):E6526–E6535.
- Heath D, Williams DR. 1995. *High-Altitude Medicine And Pathology*. 4th ed. Oxford (NY): Oxford University Press.
- Holm H, Gudbjartsson DF, Arnar DO, Thorleifsson G, Thorgeirsson G, Stefansdottir H, Gudjonsson SA, Jonasdottir A, Mathiesen EB, Njølstad I, et al. 2010. Several common variants modulate heart rate, PR interval and QRS duration. *Nat Genet*. **42**(2):117–122.
- Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, Pons-Estel BA, Acevedo-Vasquez E, Miranda P, Langefeld CD, et al. 2015. Genomic insights into the ancestry and demographic history of South America. *PLoS Genet*. **11**(12):e1005602.
- Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. **4**(1):44–57.
- Ioannidis AG, Blanco-Portillo J, Sandoval K, Hagelberg E, Miquel-Poblete JF, Moreno-Mayar JV, Rodríguez-Rodríguez JE, Quinto-Cortés CD, Auckland K, Parks T, et al. 2020. Native American gene flow into Polynesia predating easter Island settlement. *Nature*. **583**(7817):572–577.
- Jablonski NG, Chaplin G. 2000. The evolution of human skin coloration. *J Hum Evol*. **39**(1):57–106.
- Jacovas VC, Couto-Silva CM, Nunes K, Lemes RB, de Oliveira MZ, Salzano FM, Bortolini MC, Hünemeier T. 2018. Selection scan reveals three new loci related to high altitude adaptation in native Andeans. *Sci Rep*. **8**(1):12733.
- Johnson DC, Okondo MC, Orth EL, Rao SD, Huang H-C, Ball DP, Bachovchin DA. 2020. DPP8/9 inhibitors activate the CARD8 inflammasome in resting lymphocytes. *Cell Death Dis*. **11**(8):1–10.
- Kamvar ZN, Brooks JC, Grünwald NJ. 2015. Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Front Genet*. **6**:208.
- Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. **45**(D1):D353–D361.
- Lewis CM, Lizárraga B, Tito RY, López PW, Iannaccone GC, Medina A, Martínez R, Polo SI, De La Cruz AF, Cáceres AM, et al. 2007. Mitochondrial DNA and the peopling of South America. *Hum Biol*. **79**(2):159–178.
- Lewis CM, Tito RY, Lizárraga B, Stone AC. 2005. Land, language, and loci: mtDNA in native Americans and the genetic history of Peru. *Am J Phys Anthropol*. **127**(3):351–360.
- Li Y. 2011. *Functional analysis Of Dusp27, a Novel target Gene of The JAK1/STAT1 pathway, in Myogenesis*. Hong Kong: Hong Kong University of Science and Technology.
- Lindo J, Huerta-Sánchez E, Nakagome S, Rasmussen M, Petzelt B, Mitchell J, Cybulski JS, Willerslev E, DeGiorgio M, Malhi RS. 2016. A time transect of exomes from a native American population before and after European contact. *Nat Commun*. **7**(1):13175.
- Livi-Bacci M. 2006. The depopulation of Hispanic America after the conquest. *Popul Dev Rev*. **32**(2):199–232.
- Llamas B, Fehren-Schmitz L, Valverde G, Soubrier J, Mallick S, Rohland N, Nordenfelt S, Valdiosera C, Richards SM, Rohrlach A, et al. 2016. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci Adv*. **2**(4):e1501385.
- Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, Payne AJ, Steinthorsdottir V, Scott RA, Grarup N, et al. 2018. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet*. **50**(11):1505–1513.
- Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, et al. 2007. The NCBI dbgap database of genotypes and phenotypes. *Nat Genet*. **39**(10):1181–1186.
- Martin A, Gignoux C, Walters R, Wojcik G, Neale B, Gravel S, Daly M, Bustamante C, Kenny E. 2017. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet*. **100**(4):635–649.
- Maruo Y, Nagasaki K, Matsui K, Mimura Y, Mori A, Fukami M, Takeuchi Y. 2016. Natural course of congenital hypothyroidism by dual oxidase 2 mutations from the neonatal period through puberty. *Eur J Endocrinol*. **174**(4):453–463.
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GS, Thormann A, Flicek P, Cunningham F. 2016. The ensembl variant effect predictor. *Genome Biol*. **17**:122.

- Merbs CF. 1992. A new world of infectious disease. *Am J Phys Anthropol.* **35**(S15):3–42.
- Mikolajczak SA, Ma BY, Yoshida T, Yoshida R, Kelvin DJ, Ochi A. 2004. The modulation of CD40 ligand signaling by transmembrane CD28 splice variant in human T cells. *J Exp Med.* **199**(7):1025–1031.
- Missaggia BO, Reales G, Cybis GB, Hünemeier T, Bortolini MC. 2020. Adaptation and co-adaptation of skin pigmentation and vitamin D genes in native AMERICANS. *Am J Med Genet C Semin Med Genet.* **184**(4):1060–1077.
- Moore LG. 2001. Human genetic adaptation to high altitude. *High Alt Med Biol.* **2**(2):257–279.
- Moore LG. 2010. Uterine blood flow as a determinant of fetoplacental development. In: Moffett A, Barker DJP, Burton GJ and Thornburg K, editors. *The Placenta And Human Developmental Programming.* Cambridge: Cambridge University Press. p. 126–146.
- Moore LG. 2017. Human genetic adaptation to high altitudes: current status and future prospects. *Quat Int.* **461**:4–13.
- 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. 2018. Early human dispersals within the Americas. *Science.* **362**(6419):eaav2621.
- Murra JV. 2002. *El Mundo Andino: Población, Medio ambiente Y economía:* Fondo Editorial PUCP. Plaza Francia 1164, Lima 1, Peru.
- Nakatsuka N, Lazaridis I, Barbieri C, Skoglund P, Rohland N, Mallick S, Posth C, Harkins-Kinkaid K, Ferry M, Harney É, et al. 2020. A paleogenomic reconstruction of the deep population history of the Andes. *Cell.* **181**(5):1131–1145.e21.
- Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. 2017. Tracing the peopling of the world through genomics. *Nature.* **541**(7637):302–310.
- Novembre J, Barton NH. 2018. Tread lightly interpreting polygenic tests of selection. *Genetics.* **208**(4):1351–1355.
- O’Fallon BD, Fehren-Schmitz L. 2011. Native Americans experienced a strong population bottleneck coincident with European contact. *Proc Natl Acad Sci.* **108**(51):20444–20448.
- Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GVN, Underwood EC, D’amico JA, Itoua I, Strand HE, Morrison JC, et al. 2001. Terrestrial ecoregions of the world: a new map of life on earth: a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience.* **51**(11):933–938.
- Patterson KB, Runge T. 2002. Smallpox and the native American. *Am J Med Sci.* **323**(4):216–222.
- Penaloza D, Arias-Stella J. 2007. The heart and pulmonary circulation at high altitudes. *Circulation.* **115**(9):1132–1146.
- Ponce de León Bardalez RG. 1994. *El Peru Y sus Recursos: Atlas geográfico Y económico.* Lima: Auge S.A. Editores. p. 1–135.
- Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, Rohland N, Nägele K, Adamski N, Bertolini E, et al. 2018. Reconstructing the deep population history of central and South America. *Cell.* **175**(5):1185–1197.e22.
- Prins BP, Mead TJ, Brody JA, Sveinbjornsson G, Ntalla I, Bihlmeyer NA, van den Berg M, Bork-Jensen J, Cappellani S, Van Duijvenboden S, et al. 2018. Exome-chip meta-analysis identifies novel loci associated with cardiac conduction, including ADAMTS6. *Genome Biol.* **19**(1):87.
- Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol.* **20**(4):R208–R215.
- Quilter J. 2013. *The Ancient Central Andes.* 1st ed. London: Routledge.
- Raghavan M, Steinrucken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Avila-Arcos MC, Malaspina A-S, et al. 2015. Genomic evidence for the pleistocene and recent population history of native Americans. *Science.* **349**(6250):aab3884.
- Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, et al. 2012. Reconstructing native American population history. *Nature.* **488**(7411):370–374.
- Riley JC. 2010. Smallpox and American Indians revisited. *J Hist Med Allied Sci.* **65**(4):445–477.
- Rocha LJ, Godinho R, Brito JC, Nielsen R. 2021. Life in deserts: the genetic basis of mammalian desert adaptation. *Trends Ecol Evol.* **36**(7):637–650.
- Rodriguez-Delfin LA, Rubin-de-Celis VE, Zago MA. 2001. Genetic diversity in an Andean population from Peru and regional migration patterns of Amerindians in South America: data from Y chromosome and mitochondrial DNA. *Hum Hered.* **51**(1–2):97–106.
- Rothhammer F, Dillehay TD. 2009. The late pleistocene colonization of South America: an interdisciplinary perspective. *Ann Hum Genet.* **73**(5):540–549.
- Ruiz-Linares A, Adhikari K, Acuña-Alonzo V, Quinto-Sanchez M, Jaramillo C, Arias W, Fuentes M, Pizarro M, Everardo P, de Avila F, et al. 2014. Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet.* **10**(9):e1004572.
- Rupert J, Hochachka PW. 2001. Genetic approaches to understanding human adaptation to altitude in the Andes. *J Exp Biol.* **204**(18):3151–3160.
- Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in Human populations. *Nature.* **449**(7164):913–918.
- Sandoval JR, Salazar-Granara A, Acosta O, Castillo-Herrera W, Fujita R, Pena SD, Santos FR. 2013. Tracing the genomic ancestry of Peruvians reveals a major legacy of Pre-Columbian ancestors. *J Hum Genet.* **58**(9):627–634.
- Sanna S, Pitzalis M, Zoledziewska M, Zara I, Sidore C, Murru R, Whalen MB, Busonero F, Maschio A, Costa G, et al. 2010. Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. *Nat Genet.* **42**(6):495–497.
- Siebenmann C, Cathomen A, Hug M, Keiser S, Lundby AK, Hilty MP, Goetze JP, Rasmussen P, Lundby C. 2015. Hemoglobin mass and intravascular volume kinetics during and after exposure to 3,454-m altitude. *J Appl Physiol.* **119**(10):1194–1201.
- Sieron L, Lesiak M, Schisler I, Drzazga Z, Fertala A, Sieron AL. 2019. Functional and structural studies of toll-like 1 mutants associated with atrial-septal defect 6. *Biosci Rep.* **39**(1):BSR20180270.
- Skoglund P, Reich D. 2016. A genomic view of the peopling of the Americas. *Curr Opin Genet Dev.* **41**:27–35.
- Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, et al. 2016. The genecards suite: from gene data mining to disease genome sequence analyses. *Curr Protoc Bioinforma.* **54**:1.30.1–1.30.33.
- Stephan W. 2016. Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Mol Ecol.* **25**(1):79–88.
- Stucki S, Orozco-terWengel P, Forester BR, Duruz S, Colli L, Masembe C, Negri R, Landguth E, Jones MR, NEXTGEN Consortium, et al. 2017. High performance computation of landscape genomic models including local indicators of spatial association. *Mol Ecol Resour.* **17**(5):1072–1089.
- Szpiech ZA, Hernandez RD. 2014. Selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol.* **31**(10):2824–2827.
- Tarazona-Santos E, Carvalho-Silva DR, Pettener D, Luiselli D, De Stefano GF, Labarga CM, Rickards O, Tyler-Smith C, Pena SDJ, Santos FR. 2001. Genetic differentiation in South American Indians is related to environmental and cultural diversity: evidence from the Y chromosome. *Am J Hum Genet.* **68**(6):1485–1496.
- The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**(Database issue):D506–D515.
- van der Vliet A, Danyal K, Heppner DE. 2018. Dual oxidase: a novel therapeutic target in allergic disease. *Br J Pharmacol.* **175**(9):1401–1418.
- Vicuña L, Fernandez MI, Vial C, Valdebenito P, Chaparro E, Espinoza K, Ziegler A, Bustamante A, Eyheramendy S. 2019. Adaptation to

- extreme environments in an admixed human population from the Atacama desert. *Genome Biol Evol.* **11**(9):2468–2479.
- Vidal JP. 2014. Las ocho Regiones naturales Del Peru. *Terra Bras Nova Sér Rev Rede Bras História Geogr E Geogr Histórica.* **3**. doi:10.4000/terrabrasilis.1027.
- Voight BF, Kudravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* **4**(3):e72.
- Vujkovic M, Keaton JM, Lynch JA, Miller DR, Zhou J, Tcheandjieu C, Huffman JE, Assimes TL, Lorenz K, Zhu X, et al. 2020. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat Genet.* **52**(7):680–691.
- Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**(16):e164.
- Wang B, Zhang Y-B, Zhang F, Lin H, Wang X, Wan N, Ye Z, Weng H, Zhang L, Li X, et al. 2011. On the origin of Tibetans and their genetic basis in adapting high-altitude environments. *PLoS ONE.* **6**(2):e17002.
- Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution.* **38**(6):1358–1370.
- Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. 2014. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**(D1):D1001–D1006.
- Windheim M, Southcombe JH, Kremmer E, Chaplin L, Urlaub D, Falk CS, Claus M, Mihm J, Braithwaite M, Dennehy K, et al. 2013. A unique secreted adenovirus E3 protein binds to the leukocyte common antigen CD45 and modulates leukocyte functions. *Proc Natl Acad Sci.* **110**(50):E4884–E4893.
- Witt KE, Huerta-Sánchez E. 2019. Convergent evolution in human and domesticated adaptation to high-altitude environments. *Philos Trans R Soc B Biol Sci.* **374**(1777):20180235.
- Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland HM, Patel YM, Sorokin EP, Avery CL, et al. 2019. Genetic analyses of diverse populations improves discovery for complex traits. *Nature.* **570**(7762):514–518.
- Woolcott OO, Ader M, Bergman RN. 2015. Glucose homeostasis during short-term and prolonged exposure to high altitudes. *Endocr Rev.* **36**(2):149–173.
- Xue A, Wu Y, Zhu Z, Zhang F, Kemper KE, Zheng Z, Yengo L, Lloyd-Jones LR, Sidorenko J, Wu Y, et al. 2018. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat Commun.* **9**(1):2941.
- Yi X, Liang Y, Huerta-Sánchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliusson TS, et al. 2010. Sequencing of fifty human exomes reveals adaptations to high altitude. *Science.* **329**(5987):75–78.
- Zhou D, Udpa N, Ronen R, Stobdan T, Liang J, Appenzeller O, Zhao HW, Yin Y, Du Y, Guo L, et al. 2013. Whole-genome sequencing uncovers the genetic basis of chronic mountain sickness in Andean highlanders. *Am J Hum Genet.* **93**(3):452–462.