UNIVERSITY OF CALIFORNIA

Los Angeles

Genomic Features Underlying
Andean High-Altitude Adaptive Hemoglobin Levels

A thesis submitted in partial satisfaction of the requirements for the degree Master of Arts in Anthropology

by

Kimberly Tanya Zhu

2021

ABSTRACT OF THE THESIS

Genomic Features Underlying

Andean High-Altitude Adaptive Hemoglobin Levels

by

Kimberly Tanya Zhu

Master of Arts in Anthropology

University of California, Los Angeles, 2018

Professor Abigail Winslow Bigham, Chair

Humans have inhabited the Andean Altiplano for over 11,000 years, where the partial pressure of oxygen is 35% lower than at sea level. Peruvian Quechua who thrive in this environment display a suite of adaptive phenotypes, such as elevated hemoglobin concentration ([Hb]). The genetic architecture contributing to this adaptive phenotype is currently unknown. To identify genomic regions associated with elevated [Hb] among Peruvian Quechua, we identified single nucleotide polymorphisms (SNPs) that display strong signatures of positive selection using four statistics: LSBL, iHS, XP-EHH, and XP-nSL. We then performed a genome-wide association study (GWAS) for elevated [Hb], restricting our analysis to SNPs showing evidence of past natural selection. As GWAS nominated SNPs often have small effect sizes and represent a small portion of all associated SNPs, we aggregated SNP effects across the genome by creating a phenotype prediction model using LASSO regression. From this investigation, we created a comprehensive list of putative regions of selection and found several genomic loci that are weakly associated with [Hb]. By investigating elevated hemoglobin concentration from a genomic perspective, this study contributes novel insights into the genetic basis of adaptive evolution among Peruvian Quechua, as well as the role of positive selection in shaping trait variation.

The thesis of Kimberly Tanya Zhu is approved.

Jessica W Lynch

Molly Mauer Fox

Abigail Winslow Bigham, Committee Chair

University of California, Los Angeles

2018

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

SUPPLEMENTARY MATERIALS

INTRODUCTION

High altitude provides a natural laboratory to understand human evolutionary change by positive natural selection. Humans have inhabited the Andean Altiplano (average altitude of 3,660 meters (m)) for over 11,000 years (Rademaker et al 2014). At this altitude, the partial pressure of oxygen is approximately 35% lower than at sea level. This reduced availability of oxygen limits aerobic metabolism, thereby challenging human growth, development, and reproduction (Beall 2006, Storz & Scott 2019). Humans evolved under the atmospheric conditions found at sea level with limited exposure to mild hypoxia (Hochachka 1998). High-altitude hypoxia provides a powerful selective pressure that necessitates human adaptation.

Human populations inhabiting the Tibetan plateau, Ethiopian highlands, and the Andean altiplano have met this challenge, each with a unique suite of adaptive phenotypes.

The Andean adaptive pattern is characterized by elevated hemoglobin concentrations ([Hb]), elevated oxygen (O_2) saturation, and an overall increase in arterial O_2 content (Beall et al. 1998). Elevated [Hb] found in Andeans works to offset the reduction in atmospheric O_2 by increasing arterial O_2 content. However, this elevated [Hb] also increases blood viscosity, resulting in erythrocytosis and potentially requires increased cardiovascular musculature. Excessive erythrocytosis is understood to play a critical role in chronic mountain sickness, which is a fatal disease that affects residents of high-altitude zones. Conversely, Tibetans show no significant variation in [Hb] at altitudes up to 4,000 m. compared to values found at sea-level (Beall et al. 1998). At altitudes above 4,000 m, mean [Hb] in Tibetans was found to be ~15.8 gm/dL (Beall 2006).

Genomic scans for natural selection performed among high-altitude native populations have revealed genomic regions associated with adaptation to hypoxia. Among Tibetans, both *EPAS1* and *EGLN1* have been found to be under strong positive selection (Beall et al. 2010, Bigham et al. 2010, Simonson et al. 2010). They are part of the hypoxia inducible factor (HIF) pathway, an evolutionary ancient system that regulates metabolic and erythropoietic responses to oxygen concentration at the organismal and cellular level. *EGLN1* encodes for PHD2, a HIF regulator. *EPAS1* is a paralog of *HIF1A* and is a regulatory gene in the HIF pathway responsible for inducing transcription in downstream genes in response to decreased oxygen levels. SNPs in these genes associate with the low hemoglobin phenotype of Tibetan adaptation perhaps through a loss-of-function mutation for *EGLN1* (Song et al., 2020), although a gain-of-function for this locus has also been proposed (Lorenzo et al, 2014). Among Andeans, multiple genes exhibiting evidence for recent positive selection in the HIF pathway including *EGLN1* and EPAS1 have been identified (Bigham et al., 2010; Foll et al., 2014). However, despite *EGLN1* and *EPAS1* SNP associations with [Hb] identified among Tibetans, no significant SNP associations with [Hb] among Andeans have been identified for these two genes (Bigham et al., 2013). Critically, there has been no research investigating the genetic architecture underlying elevated hemoglobin levels among Andeans at a genome-wide level. To date, we do not know the genes involved in regulating this phenotype, nor the extent of polygenicity that may be underlying this trait.

To investigate whether Andean elevated [Hb] is a result of positive natural selection and genomic adaptation, we performed a genome-wide selection scan and association study using genome-wide SNP genotype data from Peruvian Quechuas study. participants recruited based on a migrant study design. We identified genomic loci showing signatures of positive natural selection and expected some of these genomic loci to be significantly associated with the adaptive phenotype of elevated [Hb].

RESULTS

Study Design and Participant Characteristics

We recruited 603 individuals of high-altitude Peruvian Quechua ancestry from two locations in Peru, Cerro de Pasco (4,338 m) and Lima (154 m). Participants were recruited using a migrant study design that included three groups: high-altitude Quechua (HAQ), migrant Quechua (MQ), and low-altitude Quechua (LAQ). HAQ participants were born, raised, and resided at high-altitude (n=301). MQ participants were born at high altitude (above 3,000 m), but down migrated low altitude within their lifetime (n=150). LAQ participants were born, raised, and resided at low altitude (n=152). HAQ study participants were recruited in Cerro de Pasco, Peru whereas MQ and LAQ study participants were recruited in Lima, Peru. Study participants characteristics are provided in Table 1. Two HAQ individuals did not provide a blood sample for [Hb] measurement and were excluded from the hemoglobin analyses. All participants were between the ages of 18 and 35 (average 24.7 ± 5.13). Forty-nine percent of the total participant cohort was female. HAQ study participants weighed less and had higher average [Hb] than both LAQ and MQ. LAQ study participants were heavier and younger than MQ study participants.

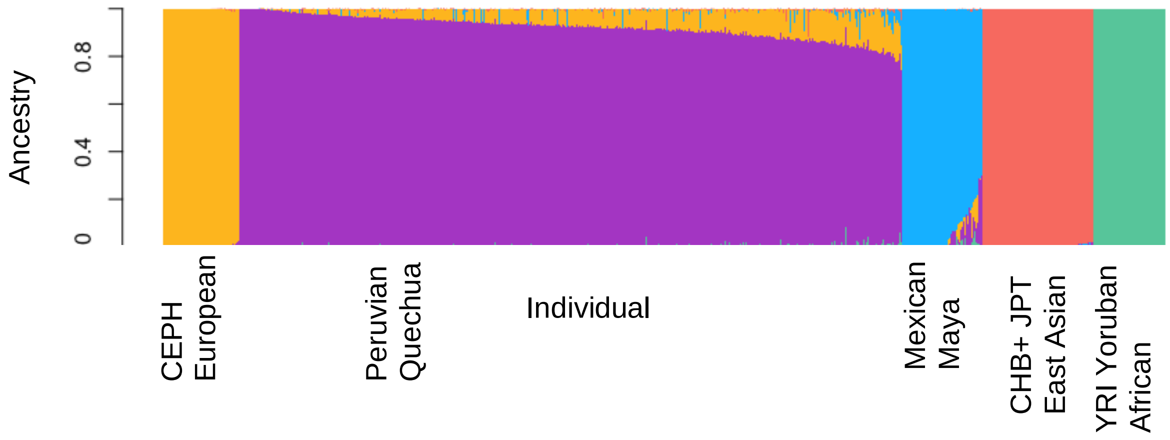
Table 1. Participant Characteristics.

	All, n=603	**HAQ, n= 301**	**LAQ, n=152**	**MQ, n=150**	**LAQ+MQ, n=302**
% Female	0.5	0.48	0.5	0.53	0.52
Age (years)	24.61 ± 5.13	24.33 ± 4.99	24.45 ± 4.58 β	25.49 ± 5.47	24.97 ± 5.06
Height (cm)	158.59 ± 10.58	158.17 ± 8.18	161.02 ± 8.96	158.02 ± 7.85	159.52 ± 8.54
Weight (kg)	61.68 ± 10.22	59.51 ± 8.21 αβδ	66.34 ± 11.77 β	61.69 ± 9.44	64.02 ± 10.9
[Hb] (g/dl)	15.69 ± 2.71	17.7 ± 2.19 αβδ	13.76 ± 1.42	13.71 ± 1.45	13.73 ± 1.43

α p≤0.05 vs LAQ, β p≤0.05 vs MQ, δ p≤0.05 vs LAQ+MQ

Genome-wide SNP data were generated using the Affymetrix (Santa Clara, CA) Axiom Biobank Genotyping Array consisting of ~600,000 polymorphic genetic loci for all Peruvian Quechua study participants. In addition, Axiom Biobank Genotyping Array data were generated for 101 low-altitude Indigenous Americans of Mexican Maya descent recruited from the city of Palenque, Chiapas, Mexico (60 m). Statistical analysis was carried out with 383,930 autosomal markers passing QC filtering (supplemental Table S1). We removed 23 Peruvian Quechua individuals and 28 Mexican Maya individuals that were first, second, or third degree relatives using KING (Manichaikul et al. 2010). The resulting dataset consisted of 577 Peruvian Quechua participants and 71 Mexican Maya participants (supplemental Table S1).

In order to identify Indigenous American individuals (Peruvian Quechua and Mexican Maya) with high degrees of non-Indigenous ancesty, we estimated global ancestry using the program ADMIXTURE (Alexander & Lange 2011) and performed a principal component analysis (PCA) using Plink 2.0 (Purcell et al. 2007, Chang et al. 2015) (Figure 1). For these analyses, we included publicly available data for 60 Yorubans (YRI), 45 Han Chinese from Beijing (CHB), 45 Japanese from Tokyo (JPT), and 60 individuals of north-central European ancestry (CEU) from the Human Genome Diversity Project-Centre d'Etude du Polymorphisme Humain (HGDP-CEPH) from the International Hap Map Project (International HapMap 2003). Indigenous American ancestry ranged from 100% to 51.15% among Peruvian Quechua and 100% to 51.58% among Mexican Maya. On average, Peruvian Quechua study participants were found to have 90.51% Indigenous American ancestry, 8.28% European ancestry, and less than 1% of African and Asian ancestry (0.69% and 0.51%, respectively). Mexican Maya were found to have on average 94.59% Indigenous American ancestry, 3.98% European ancestry, 1.20% African ancestry, and 0.23% East Asian ancestry (Figure 1A). PCA revealed that Peruvian Quechua and Mexican Maya formed an Indigenous American cluster that was distinct from the CEU, YRI, and East Asian populations (Figure 1B).



B.

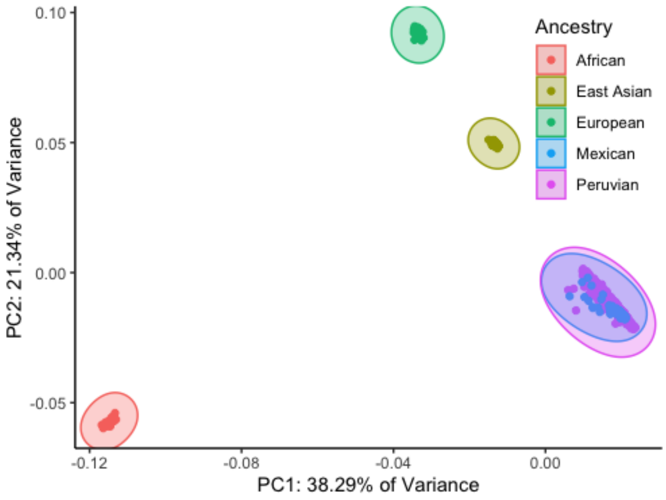


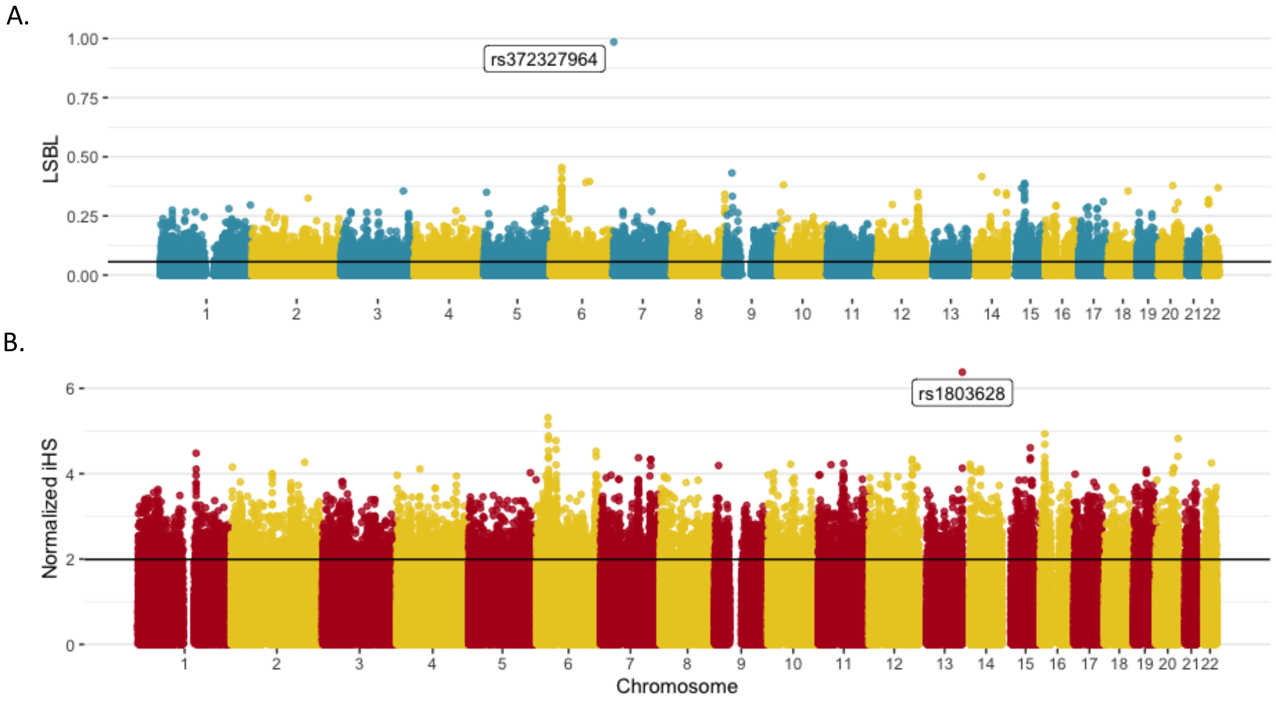
Figure 1. ADMIXTURE results after implementing a total Indigenous Ancestry cutoff of 80% for Peruvian Quechua and Mexican Maya participants (A). Principal Component Analysis (PCA) after implementing the total Indigenous Ancestry cutoff of 80% (B).

Peruvian Quechua genomes show signatures of natural selection

To test for evidence of recent positive selection, we performed a selection scan using four statistics: Locus Specific Branch Length (LSBL) (Shriver et al. 2004) and three haplotypes tests of selection 1. Cross-population number of segregating loci (XP-nSL) (Szpiech et al., 2021), 2. Cross-population extended haplotype homozygosity (XP-EHH) (Sabeti et al. 2007; Pickrell et al. 2009), and 3. Integrated haplotype score (iHS) (Voight et al. 2006). These statistics are robust at detecting recent positive selection, including soft selective sweeps and selection on variants that have not yet reached fixation. We limited the selection scan to samples with more than 90% Indigenous ancestry, resulting in 458 Peruvian Quechua individuals and 60 Mexican Maya individuals. LSBL was calculated for all polymorphic variants in the dataset (331,122 SNPs) by comparing F_{ST} between Quechua, Maya, and East Asians. XP-nSL was calculated using a log-ratio of the SL statistic for the haplotype pools of Peruvian Quechua and Mexican Maya and included 278,532 SNPs. iHS was calculated to track haplotype homozygosity decay for ancestral and derived haplotypes extending from each core SNP using 153,176 SNPs that had known ancestral alleles. XP-EHH sums iHH statistics for

Peruvian Quechua and Mexican Maya in conjunction with EHH and included 279,362 SNPs. Statistical significance thresholds were determined for all four statistics using the empirical distribution, with loci in the top 5% of the distribution for each test considered statistically significant.

We identified 16,557 statistically significant genomic markers for LSBL (Figure 2A). LSBL ranged from 0.16 to 0.98. The most extreme LSBL value was reported for the SNP, rs372327964, an intronic variant located in the gene *GET4*. We identified 7,659 statistically significant SNPs for iHS (Figure 2B). with values ranging from 1.97 to 6.01. The *GAS6* missense variant, rs1803628, displayed the most extreme iHS value. For XP-EHH, 13,968 genomic markers were identified as statistically significant (Figure 2C). Significant normalized XP-EHH values ranged from 1.66 to 8.71. The *TRIM31* missense SNP, rs3734838, with the most significant XP-nSL displayed the most extreme XP-EHH value. For XP-nSL, we identified 12,095 statistically significant SNPs with values ranging from 1.66 to 8.37 (Figure 2D). The *TRIM31* missense variant, rs3734838, exhibited the most extreme positive normalized XP-nSL value.



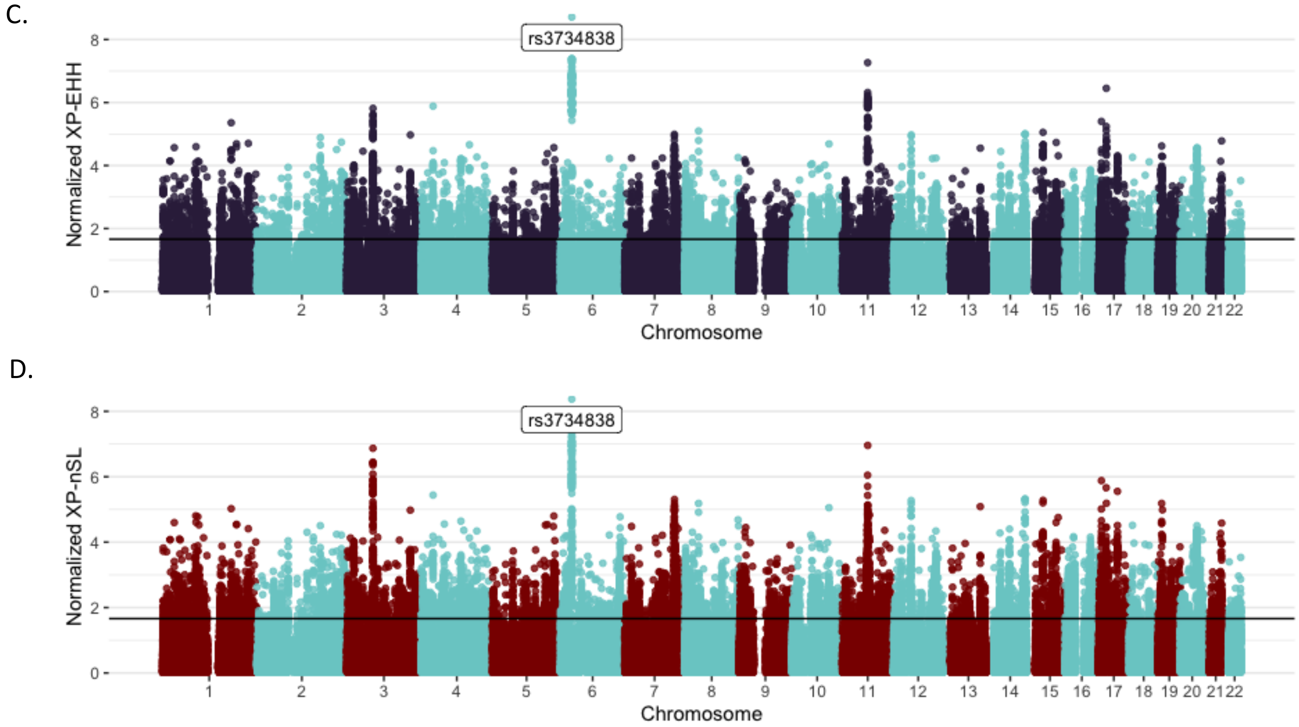


Figure 2. Selection scan results for LSBL(A), iHS (B), XP-EHH (C), and XP-nSL(D) are depicted by each plot. Top 5th percentiles are indicated by the horizonal black lines. Top SNPs for each statistic are labeled.

Several regions of the genome showed clusters of significant results. As expected, the HLA region on chromosome six, a highly variable genomic region involved in immune system functioning and protection against pathogens, showed a cluster of significant results across all four test statistics (Figure 3). In addition to the HLA region, we detected clusters of significant results for several additional regions of the genome for one or more statistics. The haplotype tests XP-EHH and XP-nSL displayed a significant cluster around Chr3:6,940,000 (Figure 3A) and Chr11: 67,200,000 (Figure 3B). For the chromosome 3 region, two SNPs located within intronic regions of *FRMD4B* were the most significant SNPS for this region. rs73107500 (normalized XP-EHH = 5.82, normalized XP-nSL = 6.43) was located at the top of this peak for XP-EHH, and rs9985338 (normalized XP-EHH = 5.35, normalized XP-nSL=6.87) was located at the top for XP-nSL. The top SNP for the chromosome 11 region was rs35363135 (normalized XP-EHH = 7.27, normalized XP-nSL = 6.05). rs35363135 was located in a non-coding region of protein coding gene *RPS6KB2*. The LSBL results showed a peak located around Chr15:450,000,000 (Figure 3C). The top SNP in this peak was rs269866 (LSBL = 0.39), located in an intronic region of *DUOX2*. The iHS and LSBL statistics showed a peak located around Chr12:112,500,000 (Figure 3D). rs7971204 (normalized iHS = 4.40, normalized LSBL = 0.48) was the most significant SNPs in this region for iHS and LSBL, respectively. rs7971204 was located in an intergenic region between *RBM19* and *RP11-100F15.*

The union of statistically significant SNPs from all four selection scans resulted in 36,399 selection nominated genomic loci, of which 439 were identified across all four statistics (Figure 4, Supplementary Table S2). This agreement rate of 1.21% was expected as these complimentary statistics use orthogonal approaches to detect signals of selection. In order to reduce false positives, we identified regions of the genome showing evidence of selection for LSBL and one of the three haplotype tests. A total of 1,168 statistically significant SNPs were shared between iHS and LSBL, and 2,501 statistically significant SNPs were shared between XPEHH and LSBL. Between XP-nSL and LSBL, 2,671 SNPs were found to statistically significant (Figure 4A-C).

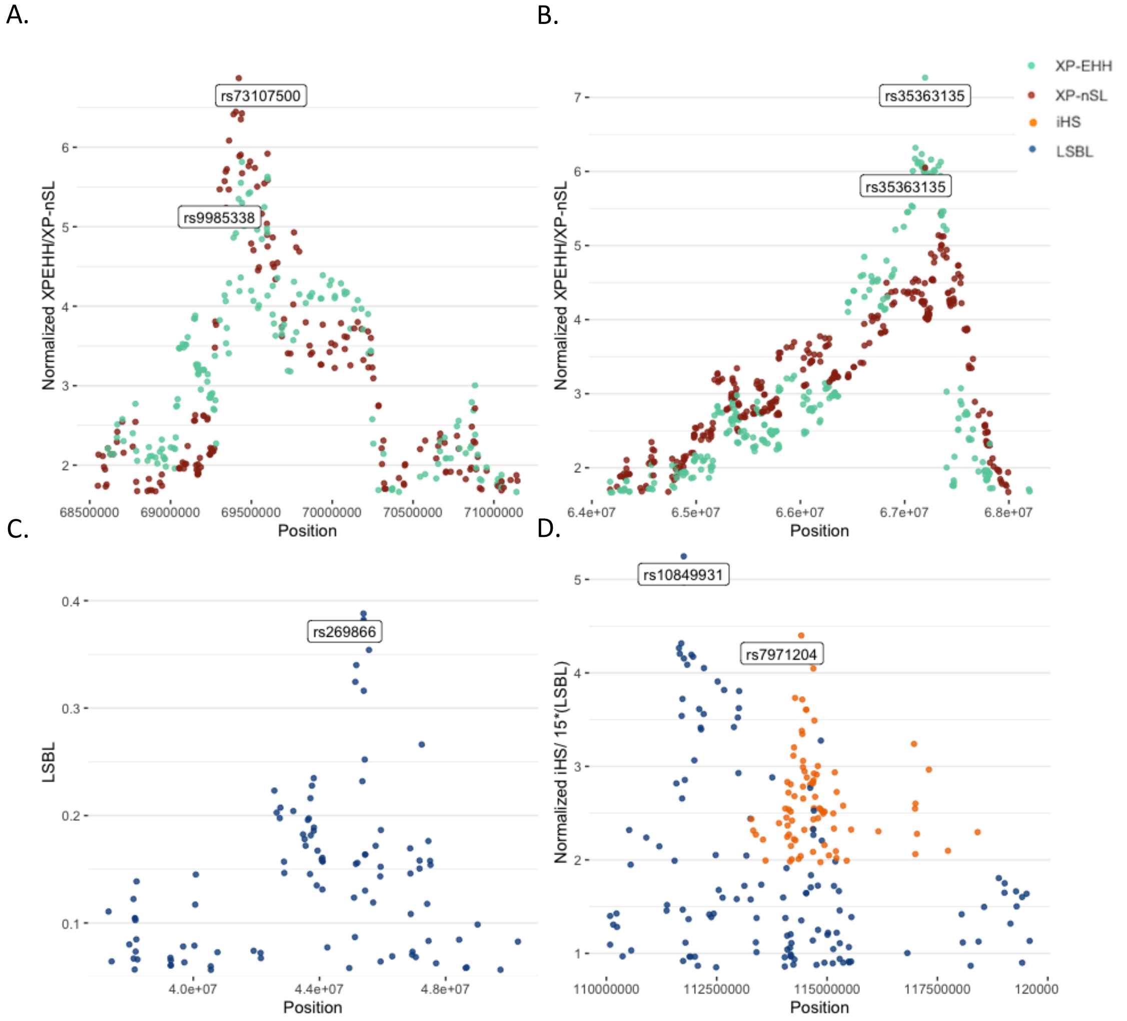


Figure 3. Selection scan statistic peaks found for Chromosomes 3 (A), 11 (B), 15 (C), and 12 (D) are depicted in each plot. For Chromosome 12, the peak is depicted with LSBL values that have been multiplied by 15.

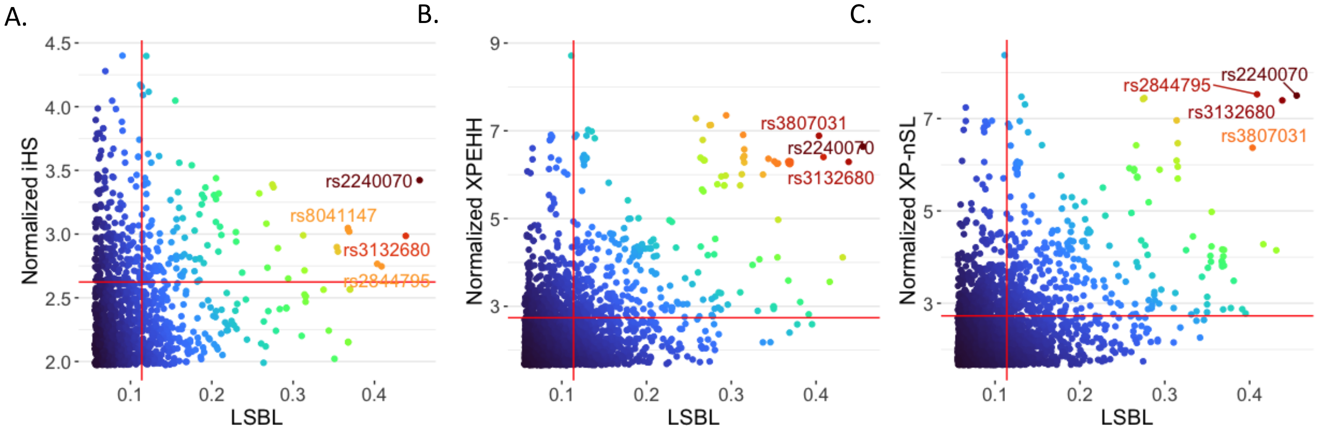


Figure 4. Intersection of statistically significant SNPs shared between LSBL and iHS (A), XPEHH (B), and XP-nSL (C). 99th percentile lines are indicated in red, and several highly significant SNPs in each intersection are highlighted.

Signatures of Selection in the HIF Genes

From our putative list of statistically significant genomic loci, we investigated loci falling within genes that are involved in the HIF pathway. The HIF pathway is the central oxygen regulating system in the human body. Previous genome-wide scans for natural selection identified putative evidence of natural selection for genes that are part of the HIF pathway (Beall et al. 2010, Bigham et al. 2010, Simonson et al. 2010). Therefore, we looked specifically within genes in this pathway for evidence of natural selection as these signatures may be indicative of the genomic response to high-altitude hypoxia specifically. Of the 90 genes involved in HIF-1 Signaling Pathway, the Axiom Biobank array assayed genomic variants in 66 of these genes. Of the 66 genes that our dataset covers, 22 containing a total of 52 SNPs show evidence of natural selection (Supplementary Table S3). Of the 52 significant SNPs located within HIF pathway genes, 41 are intronic, five are located in 3’ or 5’ untranslated regions (3’UTR or 5’UTR), and six are non-synonymous coding SNPs. Two of the non-synonymous SNPs, rs149348765 and rs11549465, are located in HIF1A. rs149348765 was found to show evidence of natural selection via XPEHH and XP-nSL statistics (XPEHH= 3.74, XP-nSL= 4.04). rs11549465 was found to show evidence of natural selection via the XP-nSL statistic (XP-nSL= 1.88). The other four non-synonymous SNPs were located in EP300 (rs20551, iHS=3.04), PSMD2 (rs11545172, XPEHH=1.76 and rs11545169, XPEHH=1.76), and PSMD9 (rs2230681, LSBL=0.06). In addition to the two non-synonymous SNPs located in HIF1A, two intronic SNPs were also found to show evidence of natural selection, rs1951795 and rs2301113. rs1951795 was detected by XP-nSL (XP-nSL=1.9), while signatures of selection for rs2301113 was detected by both XPEHH and XP-nSL (XPEHH=1.87, XP-nSL=2.07). For *EPAS1* and *EGLN1*, two genes previously identified as showing evidence of positive selection among Andean highlanders, evidence for natural selection was found for *EPAS1* intronic SNP rs1992846 and *EGLN1* intronic SNP rs2491403. rs1992846 was found to show evidence of natural selection by two statistics, iHS and XP-nSL (iHS=2.29, XP-nSL=1.67). rs2491403 was found to be significant for LSBL (LSBL= 0.12). In addition, a single intronic SNP, rs3733829, was found to be significant for *EGLN2*. This SNP showed evidence of selection from all four statistics (LSBL=0.12, iHS=2.17, XPEHH= 1.4, and XP-nSL=2.6) (Supplementary Table S3).

Signatures of Selection in Hb Genes

From our putative list of statistically signficicant genomic loci, we investigated loci falling within genes that are related to hemoglobin. Eleven SNPs located in *HBE1* were found to have statistically significant signals of selection. Of these 11 SNPs, all were located in intronic regions of *HBE1*. Six of these SNPs were located within 433bp of each other and were all detected by LSBL. Within this cluster of six SNPs, rs5006886 had the highest LSBL value (LSBL=0.067), while the other five SNPs shared an LSBL value of 0.064. Twelve SNPs located in *HBG2* showed significant signals of selection, all located in intronic regions of the gene. As *HBG2* is partially overlapped with *HBE1,* 11 of the SNPs found to be significant in HBG2 comprise the totality of the SNPs found to be significant in *HBE1.* The one SNP located in HB2 that does not overlap with HBE1 was rs3802978 (LSBL=0.10) (Supplementary Table S4)

Pathway Analysis

We performed a pathway overrepresentation analysis in SNPNexus and Reactome (Dayem Ullah et al. 2012, Fabregat et al. 2017, Griss et al. 2020, Oscanoa et al. 2020) to identify gene groups and associations that were significantly overrepresented in our selection scan results. We limited this analysis to include all statistically significant genomic loci that are overlapped with protein coding genes, and p-values were corrected via Bonferroni Correction. Twenty-nine pathways were found to be significantly overrepresented, of which almost all fall into Immune system, Metabolism of proteins, Extracellular matrix organization, Developmental biology, and Metabolism of RNA parent pathways (Table 2).

Table 2: Overrepresented Pathways.

Pathway ID	**Description**	**Parent(s)**	**Unadjusted p-value**
R-HSA-6805567	Keratinization	Developmental Biology	0.000000
R-HSA-5663205	Infectious disease	Disease	0.000000
R-HSA-1474244	Extracellular matrix organization	Extracellular matrix organization	0.000000
R-HSA-1474290	Collagen formation	Extracellular matrix organization	0.000000
R-HSA-166786	Creation of C4 and C2 activators	Immune System	0.000000
R-HSA-5690714	CD22 mediated BCR regulation	Immune System	0.000000
R-HSA-1799339	SRP-dependent cotranslational protein targeting to membrane	Metabolism of proteins	0.000000
R-HSA-381753	Olfactory Signaling Pathway	Signal Transduction	0.000000
R-HSA-418555	G alpha (s) signalling events	Signal Transduction	0.000000
R-HSA-2168880	Scavenging of heme from plasma	Vesicle-mediated transport	0.000000
R-HSA-1461973	Defensins	Immune System	0.000001
R-HSA-166663	Initial triggering of complement	Immune System	0.000001
R-HSA-72766	Translation	Metabolism of proteins	0.000001
R-HSA-112316	Neuronal System	Neuronal System	0.000001

R-HSA-2299718	Condensation of Prophase Chromosomes	Cell Cycle	0.000002
R-HSA-156842	Eukaryotic Translation Elongation	Metabolism of proteins	0.000002
R-HSA-2871837	FCERI mediated NF-kB activation	Immune System	0.000003
R-HSA-1650814	Collagen biosynthesis and modifying enzymes	Extracellular matrix organization	0.000004
R-HSA-2029481	FCGR activation	Immune System	0.000004
R-HSA-9010553	Regulation of expression of SLITs and ROBOs	Developmental Biology	0.000006
R-HSA-73857	RNA Polymerase II Transcription	Gene expression (Transcription)	0.000006
R-HSA-156902	Peptide chain elongation	Metabolism of proteins	0.000006
R-HSA-6809371	Formation of the cornified envelope	Developmental Biology	0.000009
R-HSA-5689880	Ub-specific processing proteases	Metabolism of proteins	0.000009
R-HSA-72163	mRNA Splicing - Major Pathway	Metabolism of RNA	0.000010
R-HSA-72172	mRNA Splicing	Metabolism of RNA	0.000011
R-HSA-9633012	Response of EIF2AK4 (GCN2) to amino acid deficiency	Cellular responses to external stimuli	0.000013
R-HSA-6803157	Antimicrobial peptides	Immune System	0.000013
R-HSA-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	Metabolism of RNA	0.000013

Signatures of Selection in Pathways of Interest

The Cellular response to hypoxia pathway included 48 SNPs in six genes that showed evidence of selection. Two of three participants in the Cellular response to hypoxia pathway, Regulation of gene expression by Hypoxia-inducible Factor (R-HSA-1234158), and Oxygen-dependent proline hydroxylation of Hypoxia-inducible Factor Alpha (R-HSA-1234176) were found to contain 14 SNPs in six genes and 41 SNPs in 17 genes, respectively. The Hemostasis Parent Pathway (R-HSA-109582) was found to contain 786 SNPs and 237 genes from our putative list of selection signals (Table 2). Each of the seven pathways that interact in this parent pathway included SNPs with evidence for selection. The top three pathways in the Hemostasis Parent Pathway by p-value were Platelet homeostasis, Platelet Adhesion to exposed collagen, and Factors involved in megakaryocyte development and platelet production (Table 3). Of the 10 Muscle contraction participating pathways that were associated with loci from our list of SNPs under selection, eight were related to cardiac conduction. Overall, the Cardiac conduction pathway (R-HSA-5576891) was found to include 291 SNPs in 61 genes found to be under selection, and all of this pathway's seven participating pathways contained SNPs with evidence of selection (Table 2). These participating pathways include Phase 0-rapid depolarization (R-HSA-5576892), Phase1- inactivation of fast Na+ channels (R-HSA-5576894), Phase 2- plateau phase (R-HSA-5576893), Phase 3- rapid repolarization (R-HSA-5576890), Phase 4- resting membrane potential (R-HSA-5576886), Ion homeostasis (R-HSA-5578775), and Physiological factors (R-HSA-5578768) (Table 3).

Genome Wide Association Study

To identify associations between loci that show signals of positive selection and our phenotype of interest, we performed two GWA studies. We first performed a GWAS that included all 519 Peruvian Quechua individuals with 80% or more total indigenous ancestry, for which we had [Hb] phenotype data. Since our study is focused on finding a genetic basis for an adaptive phenotype that we hypothesize to have recently undergone positive natural selection, we limited our analysis to include SNPs that were statistically significant at the 5% level for one or more tests for positive selection. This limitation was imposed to focus our analysis on genomic loci under recent positive selection while maximizing our statistical power given our sample size. We used an additive model of inheritance and included both sex and recruitment altitude as covariates. None of the PCs were identified to be significantly associated with [Hb], so no control for population stratification was added. We did not identify any significant associations with [Hb]. The top three variants were rs884510 (MA = C, MAF = 0.25, Beta = 0.39, 95% confidence interval (CI): 0.22 to 0.57), rs16829653 (MA = G, MAF = 0.18, Beta = -0.43, 95% CI: -0.64 to -0.22, FDR_BH = 0.76), and rs12696086 (MA = C, MAF = 0.13, Beta = -0.45, 95% CI: -0.67 to -0.23, FDR_BH = 0.757921) (Figure 5A, Table 4). rs884510 was located at Chr12:54,972,299 in the non-coding region of *PDE1B* and in the intronic region of *PPP1R1A*. rs16829653 was located at Chr3:158,569,149, and rs12696086 was located at Chr3:158,663,753. They are both located in between an upstream pseudogene, *GPR79*, and a downstream protein coding gene *IQCJ-SCHIP1*.

Table 3: Pathways of Interest.

Parent Pathway	**Pathway identifier**	**Pathway name**	**Unadjusted p-value**
Cellular Responses to External Stimuli	R-HSA-8953897	Cellular responses to external stimuli	0.000
	R-HSA-2262752	Cellular responses to stress	0.000
	R-HSA-1234174	Cellular response to hypoxia	0.384
	R-HSA-1234158	Regulation of gene expression by Hypoxia-inducible Factor	0.192
	R-HSA-1234176	Oxygen-dependent proline hydroxylation of Hypoxia-inducible Factor Alpha	0.352

Hemostasis Parent	R-HSA-109582	Hemostasis	0.035
	R-HSA-418346	Platelet homeostasis	0.003
	R-HSA-392851	Prostacyclin signalling through prostacyclin receptor	0.461
	R-HSA-418360	Platelet calcium homeostasis	0.038
	R-HSA-392154	Nitric oxide stimulates guanylate cyclase	0.000
	R-HSA-432142	Platelet sensitization by LDL	0.604
	R-HSA-983231	Factors involved in megakaryocyte development and platelet production	0.174
	R-HSA-75892	Platelet Adhesion to exposed collagen	0.097
Cardiac Conduction	R-HSA-397014	Muscle contraction	0.000
	R-HSA-5576891	Cardiac conduction	0.002
	R-HSA-5576892	Phase 0 - rapid depolarisation	0.007
	R-HSA-5576894	Phase 1 - inactivation of fast Na+ channels	0.222
	R-HSA-5576893	Phase 2 - plateau phase	0.051
	R-HSA-5576890	Phase 3 - rapid repolarisation	0.718
	R-HSA-5576886	Phase 4 - resting membrane potential	0.806
	R-HSA-5578775	Ion homeostasis	0.054
	R-HSA-5578768	Physiological factors	1.000

We performed a second GWAS, restricting our analysis to HAQ (n = 277). This GWAS was also limited to include SNPs that were statistically significant at for one or more tests for positive selection in order to focus the analysis on genomic loci showing signatures of recent positive selection and to maximize statistical power. We used an additive model of inheritance and included sex as a covariate. Again, we did not identify any significant associations with [Hb]. The top three variants from this GWAS were rs2971753 (MA = C, MAF = 0.061, Beta = -3.08, 95% CI: -3.45 to -2.72, FDR_BH = 0.32), rs56069023 (MA = G, MAF = 0.20, Beta = -0.71, 95% CI: 0.38 to 1.04, FDR_BH = 0.32), and rs587706 (MA = T, MAF = 0.40, Beta = -0.55, 95% CI: -

0.81 to -0.29, FDR_BH = 0.32) (Figure 5B, Table 4). rs2971753 was located at Chr7:2,971,753 and overlapped the AC093106.5 pseudogene. rs56069023 was located at Chr13:24,932,135 and was located in between lincRNA LINC00566 and *CYCSP33*, a pseudogene. rs587706 was located at Chr6:124,807,814 and was found in an intronic region of *NKAIN2.*

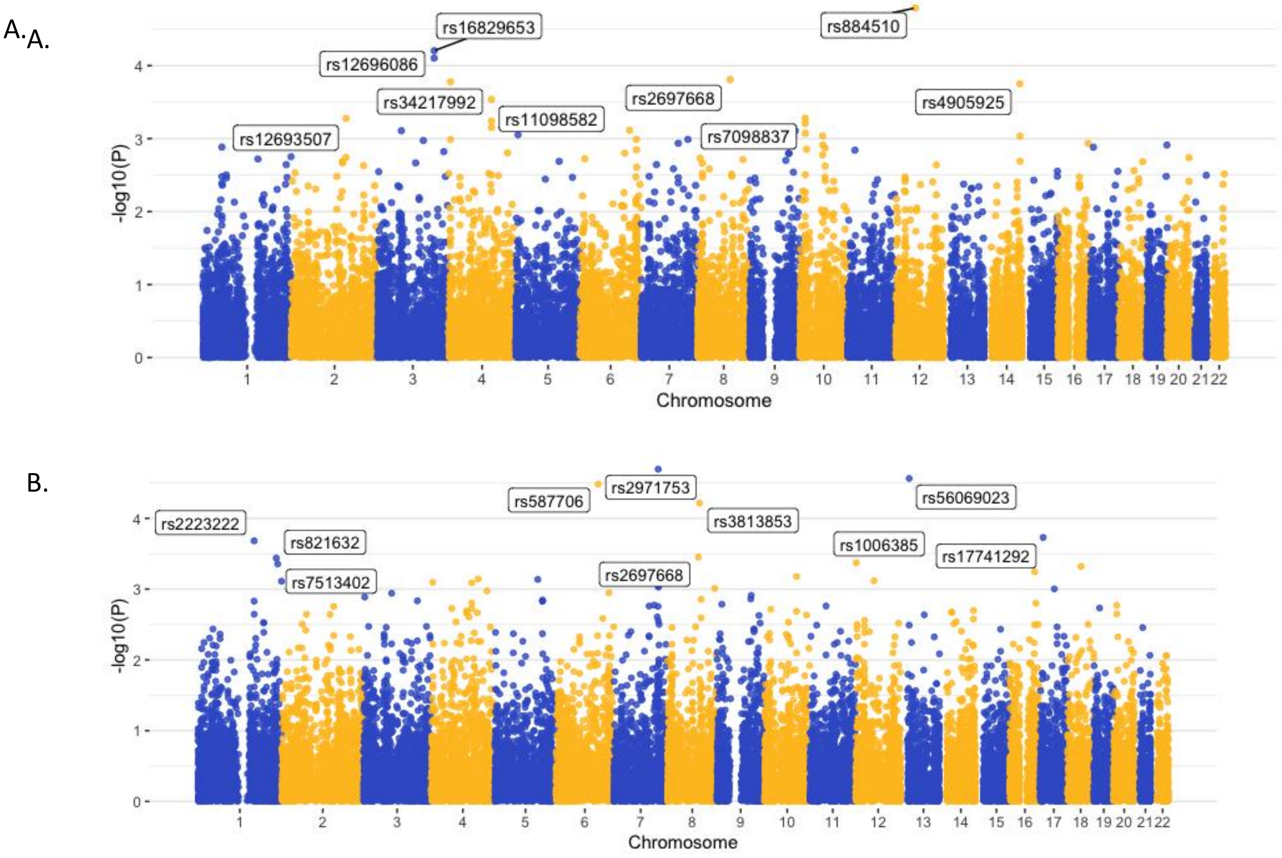
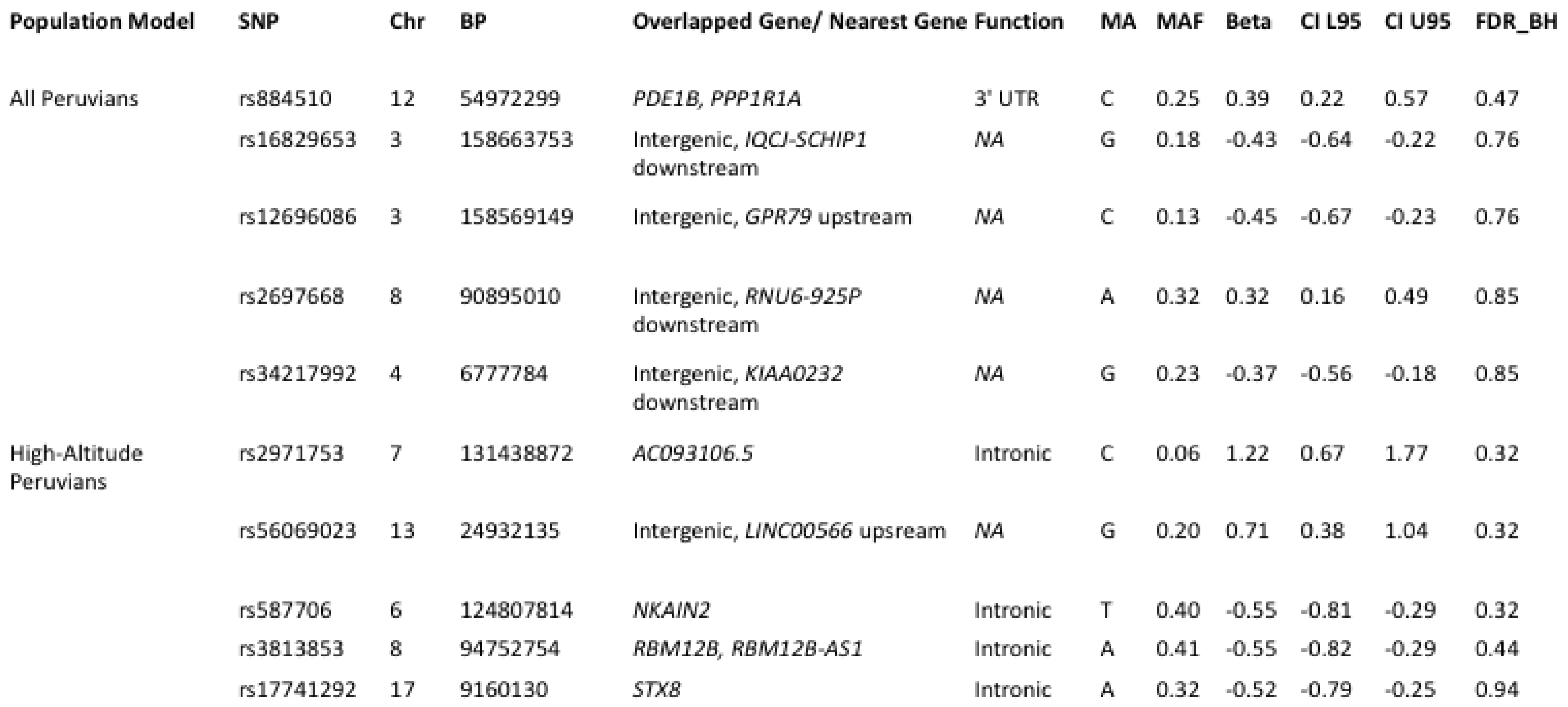


Figure 5. Manhattan plots for the genome wide association study for [Hb] phenotype**.** The top Manhattan plot shows results for the GWAS performed with all Peruvian Quechua participants passing QC and admixture filtering (A). The bottom Manhattan plot shows results for the GWAS restricted to participants that passed QC and admixture filtering (B).

Table 4: Genome Wide Association Study



Population Model	SNP	Chr	BP	Overlapped Gene/ Nearest Gene	Function	MA	MAF	Beta	CI L95	CI U95	FDR_BH
All Peruvians	rs884510	12	54972299	*PDE1B, PPP1R1A*	3' UTR	C	0.25	0.39	0.22	0.57	0.47
	rs16829653	3	158663753	Intergenic, *IQCJ-SCHIP1* downstream	*NA*	G	0.18	-0.43	-0.64	-0.22	0.76
	rs12696086	3	158569149	Intergenic, *GPR79* upstream	*NA*	C	0.13	-0.45	-0.67	-0.23	0.76
	rs2697668	8	90895010	Intergenic, *RNU6-925P* downstream	*NA*	A	0.32	0.32	0.16	0.49	0.85
	rs34217992	4	6777784	Intergenic, *KIAA0232* downstream	*NA*	G	0.23	-0.37	-0.56	-0.18	0.85
High-Altitude Peruvians	rs2971753	7	131438872	*AC093106.5*	Intronic	C	0.06	1.22	0.67	1.77	0.32
	rs56069023	13	24932135	Intergenic, *LINC00566* upsream	*NA*	G	0.20	0.71	0.38	1.04	0.32
	rs587706	6	124807814	*NKAIN2*	Intronic	T	0.40	-0.55	-0.81	-0.29	0.32
	rs3813853	8	94752754	*RBM12B, RBM12B-AS1*	Intronic	A	0.41	-0.55	-0.82	-0.29	0.44
	rs17741292	17	9160130	*STX8*	Intronic	A	0.32	-0.52	-0.79	-0.25	0.94

LASSO Regression

GWAS nominated SNPs typically have small effect sizes and represent a small portion of all truly associated SNPs. Penalized regression approaches such as the least absolute shrinkage and selection operator (LASSO) can result in lower mean squared errors (MSEs) and improved prediction accuracy (Hastie et al. 2009) and have the capability to reduce the number of predictor variables (genomic loci) that typically greatly outweigh the number of response variables (samples) in genome wide association tests (Waldmann et al. 2013). Therefore, we performed a LASSO regression as an independent validation to test the robusticity of the GWAS results. We divided the data into a training dataset (66%) and a test dataset (33%) via random selection. A regression model was fit to the training data in glmnet in R (Friedman et al 2010, R Core Team). We tested elastic net models with alphas ranging from 0 to 1. We found that the mean squared error was minimized when alpha = 1, supporting use of a LASSO model as opposed to elastic net or RIDGE regression models. The model fit with this alpha parameter included no genomic loci, verifying the GWAS results.

DISCUSSION

The genetic architecture underlying elevated [Hb] as an adaptive phenotype under high-altitude hypoxia is not well understood. We performed a genome wide selection scan and association study to test for genotype associations with [Hb]. We identified 1) several genomic loci showing signatures of selection, 2) several genomic loci under positive selection that show weak associations with [Hb], and 3) no genomic loci under positive selection that are significantly associated with [Hb]. The selection scan we performed resulted in an extensive list of putative regions of selection. This list included numerous genes that were associated with responses to hypoxia, cardiac muscle contraction, hemoglobin, and nitric oxide. The lack of evidence for a genetic basis of high-altitude adaptive elevated [Hb] provided by this study does not preclude the possibility that this trait has a genetic basis, nor does it prove that this trait is entirely developmentally or epigenetically based.

By performing a genome wide selection scan, we have produced a comprehensive list of putative genomic regions under selection. These genomic regions include several genes in the HIF pathway, two hemoglobin related genes, and several genes related to cardiac conduction. The HIF pathway is activated when cellular oxygen demand surpasses oxygen supply (cellular hypoxia). HIF pathway activation results in up-regulation of genomic regions involved in glycolysis, erythropoiesis, and angiogenesis. The up-regulation of these genes promotes the cellular response to hypoxia (Cavadas et al. 2013). In particular, we identified evidence of positive selection for four genes important in the HIF pathway, *EGLN1*, *EGLN2*, *EPAS1*, and *HIF1A*. *EGLN1* codes for the PHD2 enzyme, which interacts with and degrades the HIF-2a protein. Under conditions of hypoxia, PHD2 enzymes have reduced activity, resulting in an increased stabilization of HIF-2a proteins, which increases blood cell and blood vessel formation capacity. Similarly, *EGLN2* encodes for the PHD1 enzyme, and both *EGLN1* and *EGLN2* are involved in the degradation of HIF-1a (Zhang et al. 2019). *HIF1A* encodes for a protein subunit of HIF-1, which acts as a master regulator of cellular and systemic responses to hypoxia. HIF-1 activates transcription of an array of genomic regions to increase oxygen delivery and metabolic adaptation to hypoxia. Two non-synonymous coding variants located in *HIF1A*, rs149348765 and rs11549465, showed signatures of selection. Further investigation of these two variants and their implication in vitro could yield important insights on the genetic alterations that alter the activity of genes in the HIF pathway, and ultimately Andean adaptations to high-altitude hypoxia.

We identified evidence of selection for two genes related to hemoglobin, *HBE1* and *HBG2*. *HBE1* is a protein coding gene that encodes hemoglobin subunit epsilon, a component of embryonic hemoglobins Hb Gower I and Hb Gower II. *HBG2* is a protein coding gene that encodes of a one of two gamma chains that along with two alpha chains constitutes fetal hemoglobin (HbF). All of the 11 SNPs located in *HBE1* and the 12 SNPs located in *HBG2* that were found to be under selection, were intronic. The signals of selection found in these genes may be the result of genetic adaptation involved in the increased blood flow and oxygen to uteroplacental circulation. This elevated blood flow and oxygen delivery has been found to be an important factor in protecting high-altitude Andeans and Tibetans from fetal growth restrictions related to hypoxia (Julian et al. 2009, Wilson et al. 2007). Alternatively, both *HIF2A* and *EPAS1* are involved in increased erythropoiesis and erythroid expansion. Characteristics of increased erythropoiesis includes an augmentation of red blood cells containing HbF. Low-altitude residing adults that are exposed to high altitude and high-altitude hypoxia have been found to have an elevated accumulation of HbF that recedes to normal levels after return to low altitude (Risso et al. 2011). Future studies including flow cytometry analysis of HbF expression using blood samples obtained from the Peruvian Quechua participants recruited at high and low altitudes could provide valuable insights as to whether this increased HbF accumulation occurs in Peruvian Quechua participants, and if so, is it the result of genetic adaptation in *HBG2*.

Our GWAS did not reveal any associations with [Hb] that passed genome-wide significance. Several genomic loci were found to be weakly associated with [Hb]. The top three included rs884510, rs16829653, and rs12696086. *PDE1B* or *PPP1R1A* SNP rs884510 is involved in several hemostasis pathways including: cGMP effects (R-HSA-418457), Nitric oxide stimulates guanylate cyclase (R-HSA-392154), Platelet homeostasis (R-HSA-418346), and Hemostasis (R-HSA-109582). This gene is also involved in several signal transduction pathways. *PPP1R1A* is located in the Cellular responses to stress parent pathway, and its sub-pathway, the Response of *EIF2AK1* (HRI) to heme deficiency (R-HAS-9648895). rs16829653 and rs12696086 are located in between an upstream pseudogene, *GPR79*, and a downstream read-through transcription gene *IQCJ-SCHIP1.* The top three associated variants for the HAQ restricted GWAS were rs2971753, rs56069023, and rs587706. rs587706 is located in an intronic region of *NKAIN2,* a protein coding gene associated with Cardiovascular and

Hematological disease classes, specifically blood viscosity and cell adhesion molecules (GAD) that have been implicated in cardiovascular morphology. Cell adhesion molecules also include members of the immunoglobin family. Together, our findings suggest that the genomic regions surrounding these genes can be promising candidate regions for future investigations of the genetic architecture underlying adaptive [Hb] in Andeans.

The lack of a significant genomic association found in this study is subject to select limitations. First, data was collected using a SNP array, rather than whole genome sequencing (WGS) or exome sequencing (ES). This limited representation genomic data may not include genomic loci located within statistically significant regions of selection that are significantly associated with [Hb]. Second, haplotype and genotype imputation with programs such as RFmix (Maples et al. 2013) was not performed on our data set. This imputation would replace chromosomal sections and impute indigenous chromosomal sections in their place. Instead, admixed individuals were removed for a more conservative approach. Finally, For LASSO regression and prediction, our sample size is quite small, limiting the predictive accuracy of our model (Fryett et al. 2020).

Future work using our list of putative regions showing signatures of natural selection in addition to data that could potentially yield novel genomic loci that are significantly associated with [Hb]. Our list of putative regions and our data set can be used to test significant associations with other high-altitude adaptive phenotypes such as exhaled nitric oxide (FENO). On a broader scale, a future study that investigates the genomic architecture underlying [Hb] found in Tibetans, in addition to the predictive modeling of these trait in Han Chinese would be invaluable to understanding how differences in ancestral genomic architecture have resulted in the unique suites of high-altitude adaptations found in Tibetans and Andeans.

MATERIALS AND METHODS

Participant Recruitment and Data Collection

We recruited 603 Peruvian Quechua study participants from two locations in Peru. 301 participants of Peruvian Quechua descent from the city of Cerro de Pasco, Peru (4, 338m) who had been born and raised at high altitude. We also recruited 300 participants of Peruvian Quechua descent from Lima, Peru (154m) who were born and raised at low altitude or had migrated down to low altitude during their lifetime. At the time of recruitment, all study participants provided written informed consent in Spanish. This study was approved by the Institutional Review Boards of the University of Michigan and the Universidad Peruana Cayetano Heredia. Additionally, we recruited 101 Indigenous American low altitude study participants of Mexican Maya descent who spoke the Tzeltal, Tzotil, or Ch'ol from the city of Palenque, Chiapas Mexico (60m). Again, study participants provided written informed consent in Spanish at the time of enrollment. The study was approved by the institutional review boards at the University of Michigan and Centro de Investigación y Docencia Económicas (CIDE Mexico City, Mexico). Peruvian and Mexican study participants provided three mililiters of whole blood for DNA extraction and measurement of [Hb]. [Hb] was measured using a HemoCue Hb201+ analyzer (AngelHolm Sweden). Blood was field stabilized in cell lysis buffer. Stabilized blood samples were then hand-carried to the University of Michigan where DNA extraction was performed in the Bigham Lab for Anthropological Genomics using the Puregene Protocol (Qiagen, Valencia, CA). In addition, study participants provided basic biometric data including weight, height, sex, and age.

Genotype data was generated using the Affymetrix Biobank Genotyping Array, which returned genetic data for 592,123 genomic loci. Publicly available data were obtained for 60 Yorubans (YRI), 45 Han Chinese from Beijing, 45 Japanese from Tokyo, and 60 individuals of north-central European ancestry from the Centre d'Etude du Polymorphisme Humain (CEPH) from the International Hap Map Project (International HapMap 2003). These data were used as reference data during admixture analysis and in the selection scan.

Quality Control filtering

Genotype data were quality control filtered using Plink 2.0 (Purcell et al. 2007, Chang 2015) to only include autosomal markers with genotyping missingness rates less than 5%. Three thousand four hundred and ninety variants were removed from the HapMap data set. No variants were removed from the Peruvian or Mexican datasets. No individuals were removed based off of the missingness cutoff of 1%. Thirty-one loci were manually identified as being called for the incorrect allele and removed. These loci were removed from downstream analyses. Sample relatedness was estimated using KING (Manichaikul et al. 2010) implemented in Plink 2.0. Twenty-three Peruvian Quechua individuals 28 Mexican Maya individuals showing 3rd degree relationships or higher were identified and removed. For pairs of related individuals, the individual with the higher genotyping rate was retained in the dataset.

Phase estimation

The quality control filtered data set was first converted from a Plink .bed file format to .vcf file format, then indexed using BCFtools (Li et al. 2009). Haplotype phasing and genotype imputation were performed using the Michigan Imputation Server (Das et al. 2016). Phasing was performed using Eagle (Loh et al. 2016) and the 1000 Genomes reference panel v5 (Genomes Project et al, 2015).

Admixture Analysis

We performed an admixture analysis using the unphased data set. This reduces the over-estimation of admixture produced by switching errors that can occur during haplotype phasing.

We removed 40,960 insertion/deletion markers, filtered alleles with a minor allele frequency less than 2.5%, and pruned out markers that were in tight linkage disequilibrium (LD) (r^2>0.8). Admixture estimates for each individual were obtained using 256,150 markers in the program ADMIXTURE (Alexander & Lange 2011). The number of populations (K) was estimated to be K=5 corresponding to individuals from Peruvian Quechua, Mexican Maya, CEU, YRI, and EAS populations.

Selection Scan

We extracted ancestral allele information from publicly available 1000 Genomes data using BCFtools (Li et al. 2009), then input these ancestral alleles as reference allele in Plink 2.0 (Purcell et al. 2007). LSBL (Shriver et al. 2004) was calculated using 458 Peruvian Quechua and 60 Mexican Maya individuals with over 90% Indigenous ancestry, as well as 88 East Asian individuals from data provided by the International Hap Map Project. Monomorphic SNPs were removed from the dataset, leaving 331,122 variants. Pairwise F_{ST} was calculated between each population and was then used to calculate LSBL. Genetic markers falling in the top 5% of LSBL empirical distribution were determined to be statistically significant. Three haplotype-based tests of selection were applied using the programs Selscan v1.3.0 and Norm (Szpiech & Hernandez 2014). XP-nSL (Szpiech et al. 2021) was calculated for Peruvian Quechua were compared to the Mexican Maya as reference. XP-nSL results were normalized using Norm. Second, iHS (Voight et al. 2006) was calculated for Peruvian Quechua with Mexican Maya was reference, then normalized using Norm. Third, XP-EHH (Sabeti et al. 2007) was calculated for Peruvian Quechua, again using the Mexican Maya as reference, then normalized using Norm.

Pathway Analysis

All SNPs found to be significantly under selection were input to SNPNexus (Dayem Ullah et al. 2012, Oscanoa et al. 2020) to determine overlapped genes or nearest upstream and downstream genes. Pathway analysis was performed in Reactome (Fabregat et al. 2017, Griss et al. 2020).

Genome Wide Association Study

We combined these statistically significant markers to create a genomic data set to be used in the GWAS. Covariates were selected from a pool of participant phenotype data and the PCA performed with all remaining participants after the admixture analysis. We tested a linear regression model with Sex, Altitude, Age, Height, Weight, and principal components 1 to 5. Of these, only Sex and Altitude were found to be significant covariates, which was then verified using a student's t-test and a Barlett test. With the 519 Peruvian Quechua samples retained in the 0.8 Admixture cutoff, we performed a GWAS that included the top 5% of SNPs under selection. After a minor allele frequency cutoff of 0.05 we performed a linear association using an additive inheritance model to test using [Hb] as the key phenotype, and sex and altitude as covariates. An additional GWAS was performed with the same dataset but limited to the 277 Peruvian Quechua participants who were recruited at high altitude.

LASSO Regression

To build our LASSO regression model, we used the covariate, phenotype, and genotype data that was previously used in the GWAS. We converted this data into 0 1 2 genotypes, and centered and scaled our predictor variable ([Hb]) and parameters (covariates and SNPs). We then used glmnet in R (Friedman et al 2010, R Core Team) to fit a regression model to the data. Using a training dataset that contained 66% of the total data, we tested Ridge, LASSO, and

Elastic net models (alpha ranging from 0 to 1 in increments of 0.02) and found that the mean squared error of our model was minimized when alpha=1, which results in a LASSO regression model. We then created a set (consisting of the 33% of the total data that was excluded from the training set) to predict [Hb], then compared the predicted values against the observed.

REFERENCES

1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68.

Alexander D.H., & Lange K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*(1).

Beall, C.M., Brittenham, G.M., Strohl, K.P., Blangero, J., Williams-Blangero, S., Goldstein, M.C., Decker, M.J., Vargas, E., Villena, M., Soria, R., Alarcon, A.M. and Gonzales, C. (1998). Hemoglobin concentration of high-altitude Tibetans and Bolivian Aymara. *American Journal of Physical Anthropology, 106*(3), 385-400.

Beall, C.M. (2006). Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. *Integrative and Comparative Biology*, *46*(1), 18-24.

Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., ... & Zheng, Y. T. (2010). Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences*, *107*(25), 11459-11464.

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J. M., Mei, R., ... & Shriver, M. D. (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS genetics*, *6*(9).

Bigham, A. W., Wilson, M. J., Julian, C. G., Kiyamu, M., Vargas, E., Leon-Velarde, F., ... & Shriver, M. D. (2013). Andean and Tibetan patterns of adaptation to high altitude. *American Journal of Human Biology*, *25*(2), 190-197.

Cavadas, M. A., Nguyen, L. K., & Cheong, A. (2013). Hypoxia-inducible factor (HIF) network: insights from mathematical models. *Cell communication and signaling*, *11*(1), 1-16.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1).

Das S., Forer L., Schönherr S., Sidore C., Locke A.E., Kwong A., … & Fuchsberger C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics* *48*(10), 1284–1287.

Dayem Ullah., A. Z., Lemoine, N. R., & Chelala, C. (2012). SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic acids research*, *40*(W1), W65-W70.

Fabregat, A., Sidiropoulos, K., Viteri, G., Forner, O., Marin-Garcia, P., Arnau, V., ... & Hermjakob, H. (2017). Reactome pathway analysis: a high-performance in-memory approach. *BMC bioinformatics*, *18*(1), 1-9.

Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A., & Excoffier, L. (2014). Widespread signals of convergent adaptation to high altitude in Asia and America. *The American Journal of Human Genetics*, *95*(4), 394-407.

Friedman J., Hastie T., Tibshirani R. (2010). “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software*, *33*(1), 1.

Fryett, J. J., Morris, A. P., & Cordell, H. J. (2020). Investigation of prediction accuracy and the impact of sample size, ancestry, and tissue in transcriptome-wide association studies. *Genetic epidemiology*, *44*(5), 425-441.

Griss, J., Viteri, G., Sidiropoulos, K., Nguyen, V., Fabregat, A., & Hermjakob, H. (2020). Reactomegsa-efficient multi-omics comparative pathway analysis. *Molecular & Cellular Proteomics*, *19*(12), 2115-2125.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd Edn.* New York, NY: Springer-Verlag

Hochachka, P. W. (1998). Mechanism and evolution of hypoxia-tolerance in humans. *Journal of Experimental Biology,* 201(8), 1243-1254.

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 426(6968), 789-796.

Julian, C. G., Wilson, M. J., Lopez, M., Yamashiro, H., Tellez, W., Rodriguez, A., ... & Moore, L. G. (2009). Augmented uterine artery blood flow and oxygen delivery protect Andeans from altitude-associated reductions in fetal growth. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *296*(5), R1564-R1575.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078–2079.

Lorenzo, F. R., Huff, C., Myllymäki, M., Olenchock, B., Swierczek, S., Tashi, T., ... & Prchal, J. T. (2014). A genetic mechanism for Tibetan high-altitude adaptation. *Nature genetics*, *46*(9), 951-956.

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y.A., Finucane, H.K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics*, *48*(11), 1443–1448.

Manichaikul A., Mychaleckyj J.C., Rich S.S., Daly K., Sale M., Chen W.M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics 26*(22), 2867-2873.

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, *93*(2), 278-288.

Nunes, K., Maia, M. H. T., Dos Santos, E. J. M., Dos Santos, S. E. B., Guerreiro, J. F., Petzl-Erler, M. L., ... & Meyer, D. (2021). How natural selection shapes genetic differentiation in the MHC region: A case study with Native Americans. *Human Immunology*, *82*(7), 523-531.

Oscanoa, J., Sivapalan, L., Gadaleta, E., Dayem Ullah, A. Z., Lemoine, N. R., & Chelala, C. (2020). SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic acids research*, *48*(W1), W185-W192.

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., ... & Pritchard, J. K. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome research*, *19*(5), 826-837.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., Maller J., Sklar P., de Bakker P.I.W., Daly M.J. & Sham P.C. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics, 81*(3), 559-575.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rademaker K., Hodgins G., Moore K., Zarrillo S., Miller C., Bromley G.R., Leach P., Reid D.A., Álvarez W.Y., Sandweiss D.H. (2014). Paleoindian settlement of the high-altitude Peruvian Andes. *Science, 346*(6208):466-9.

Risso, A., Fabbro, D., Damante, G., & Antonutto, G. (2012). Expression of fetal hemoglobin in adult humans exposed to high altitude hypoxia. *Blood Cells, Molecules, and Diseases*, *48*(3), 147-153.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., ... & Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913-918.

Shiina, T., Hosomichi, K., Inoko, H., & Kulski, J. K. (2009). The HLA genomic loci map: expression, interaction, diversity and disease. *Journal of human genetics*, *54*(1), 15-39.

Shriver M.D., Kennedy G.C., Parra E.J., Lawson H.A., Sonpar V., Huang J., Akey J.M., Jones K.W. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human Genomics 1*(4), 1-13.

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., ... & Ge, R. (2010). Genetic evidence for high-altitude adaptation in Tibet. *Science*, *329*(5987), 72-75.

Song, D., Navalsky, B. E., Guan, W., Ingersoll, C., Wang, T., Loro, E., ... & Lee, F. S. (2020). Tibetan PHD2, an allele with loss-of-function properties. *Proceedings of the National Academy of Sciences*, *117*(22), 12230-12238.

Storz, Jay F., and Graham R. Scott. (2019). Life ascending: mechanism and process in physiological adaptation to high-altitude hypoxia. *Annual review of ecology, evolution, and systematics, 50*, 503-526.

Szpiech, Z. A., & Hernandez, R. D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular biology and evolution*, *31*(10), 2824-2827.

Szpiech, Z. A., Novak, T. E., Bailey, N. P., & Stevison, L. S. (2021). Application of a novel haplotype-based scan for local adaptation to study high-altitude adaptation in rhesus macaques. *Evolution Letters, 5,* 408-421.

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS biology*, *4*(3), 0446-0458.

Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., & Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in genetics*, *4*, 270.

Wilson, M. J., Lopez, M., Vargas, M., Julian, C., Tellez, W., Rodriguez, A., ... & Moore, L. G. (2007). Greater uterine artery blood flow during pregnancy in multigenerational (Andean) than shorter-term (European) high-altitude residents. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *293*(3), 1313-1324.