Identifying positive selection candidate loci for high-altitude adaptation in Andean populations

Abigail W. Bigham,[1] Xianyun Mao,[2] Rui Mei,[3] Tom Brutsaert,[4] Megan J. Wilson,[5] Colleen Glyde Julian,[5] Esteban J. Parra,[6] Joshua M. Akey,[7] Lorna G. Moore[8] and Mark D. Shriver[2]*

[1]Department of Pediatrics, The University of Washington, Seattle, WA 98195, USA
[2]Department of Anthropology, Pennsylvania State University, University Park, PA 16802, USA
[3]Affymetrix, Inc., Santa Clara, CA 95051, USA
[4]Departments of Exercise Science and Anthropology, Syracuse University, Syracuse, NY 13244-5040, USA
[5]Department of Anthropology and Altitude Research Center, University of Colorado, Denver, CO 80262, USA
[6]Department of Anthropology, University of Toronto, Mississauga, Ontario, Canada
[7]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[8]Graduate School of Arts and Sciences, Wake Forest University, Winston-Salem, NC 27109, USA
**Correspondence to:* Tel: +1 206 543 5412; Fax: +1 206 221 3795; E-mail: awb150@u.washington.edu

Date received (in revised form): 14th October 2009

Abstract

High-altitude environments (>2,500 m) provide scientists with a natural laboratory to study the physiological and genetic effects of low ambient oxygen tension on human populations. One approach to understanding how life at high altitude has affected human metabolism is to survey genome-wide datasets for signatures of natural selection. In this work, we report on a study to identify selection-nominated candidate genes involved in adaptation to hypoxia in one highland group, Andeans from the South American Altiplano. We analysed dense microarray genotype data using four test statistics that detect departures from neutrality. Using a candidate gene, single nucleotide polymorphism-based approach, we identified genes exhibiting preliminary evidence of recent genetic adaptation in this population. These included genes that are part of the hypoxia-inducible transcription factor (*HIF*) pathway, a biochemical pathway involved in oxygen homeostasis, as well as three other genomic regions previously not known to be associated with high-altitude phenotypes. In addition to identifying selection-nominated candidate genes, we also tested whether the *HIF* pathway shows evidence of natural selection. Our results indicate that the genes of this biochemical pathway as a group show no evidence of having evolved in response to hypoxia in Andeans. Results from particular *HIF*-targeted genes, however, suggest that genes in this pathway could play a role in Andean adaptation to high altitude, even if the pathway as a whole does not show higher relative rates of evolution. These data suggest a genetic role in high-altitude adaptation and provide a basis for genotype/phenotype association studies that are necessary to confirm the role of putative natural selection candidate genes and gene regions in adaptation to altitude.

Keywords: *genome scan, positive selection, Native Americans, altitude adaptation*

Introduction

Identifying gene regions showing signatures of natural selection in the human genome offers a window into our recent evolutionary past, as well as a deeper understanding of how this evolutionary force has shaped extant patterns of variation. Several recent studies have analysed dense single nucleotide polymorphism (SNP) genotype data to detect signatures of selection in three major continental groups: West Africans, East Asians and Northern Europeans.[1–6] To date, only a few studies have focused on identifying candidate genes under selection with reference to a specific selective pressure.[7,8] Here, we use high-density SNP data to search for candidate genes for altitude adaptation in Andean populations. By expanding the populations of study to the Americas and targeting a specific selective pressure, hypobaric hypoxia, we can produce a more detailed and nuanced understanding of this evolutionary process.

High-altitude environments provide scientists with a natural laboratory to study the genetic and physiological effects of hypobaric hypoxia, the decreased partial pressure of oxygen at high altitude resulting in lower circulating oxygen levels in the body, on endemic highland species.[9–11] Humans have inhabited three high-altitude (>2,500 m) zones of the world for multiple generations: the Tibetan Plateau, the Andean Altiplano and the Semien Plateau of Ethiopia (Figure 1). Each of these populations exhibits unique circulatory, respiratory and haematological adaptations to life at high altitude. For example, research has shown that Tibetan and Ethiopian populations have relatively low haemoglobin concentrations, in contrast to the 'classic' Andean physiological adaptation (also seen in high-altitude sojourners), where haemoglobin concentrations are elevated compared with low-altitude groups.[12–17] Andeans also exhibit lower levels of resting ventilation, a more 'blunted' hypoxic ventilatory response, higher levels of pulmonary arterial pressure and an increased frequency of chronic mountain sickness compared with their Tibetan counterparts.[18,19] Overall, this research has led to a substantial body of literature documenting the suite of human physiological responses to high-altitude habitation (for a review, see Hornbein and Schoene[20]).

The physiological differences between low- and high-altitude populations have been well documented, but little work has focused on understanding the genetic bases or identifying the genetic variants underlying these adaptations.[21,22] The few natural selection genetics studies conducted previously have focused on specific genes hypothesised to play a role in adaptation to altitude, but none of them have found conclusive evidence of this evolutionary force.[21–27] One recent study scanned 998 genetic markers in seven Nepalese Sherpa porters and identified genomic regions that may have been involved in adaptation to altitude.[28] However, genome scans using much larger panels of genetic markers and larger sample sizes will be necessary to expand upon these very preliminary findings. Related research has explored the heritability of specific altitude phenotypes such as arterial oxygen saturation, haemoglobin concentration and thoracic skeletal dimensions.[16,29,30] One heritability study concluded that a major autosomal dominant locus exists for high oxygen saturation, where Tibetan women carrying this high oxygen saturation allele had a higher offspring survival rate than women possessing the low oxygen saturation allele.[30] Even though research of this nature documents the potential for natural selection to act on phenotypic traits, it does not identify the gene(s) controlling the phenotype.

As part of an ongoing project to understand the role of natural selection in shaping human genetic diversity in high-altitude populations, we genotyped 490,032 autosomal SNPs using the Affymetrix, Inc. (Santa Clara, CA) GeneChip® Mapping 500 K array in 195 persons of high-altitude or low-altitude descent. By comparing high-altitude populations with related populations living at low altitude, a list of selection-nominated candidate genes and gene regions was generated using four summary statistics: locus-specific branch length (LSBL), the natural log of the ratio of heterozygosities (ln*RH*), Tajima's D and the whole-genome long-range haplotype (WGLRH) test. We focused our attention on the hypoxia-inducible factor (HIF) pathway, which is a transcriptional regulator that controls cellular oxygen homeostasis and plays a key role in energy metabolism. It is upregulated in many cancers and may be involved in the accumulation of adipose tissue. This pathway, comprising at least 75 genes scattered throughout the genome, is thought to regulate many of the physiological responses to cellular hypoxia. Based on their functional roles, we have an *a priori* reason to

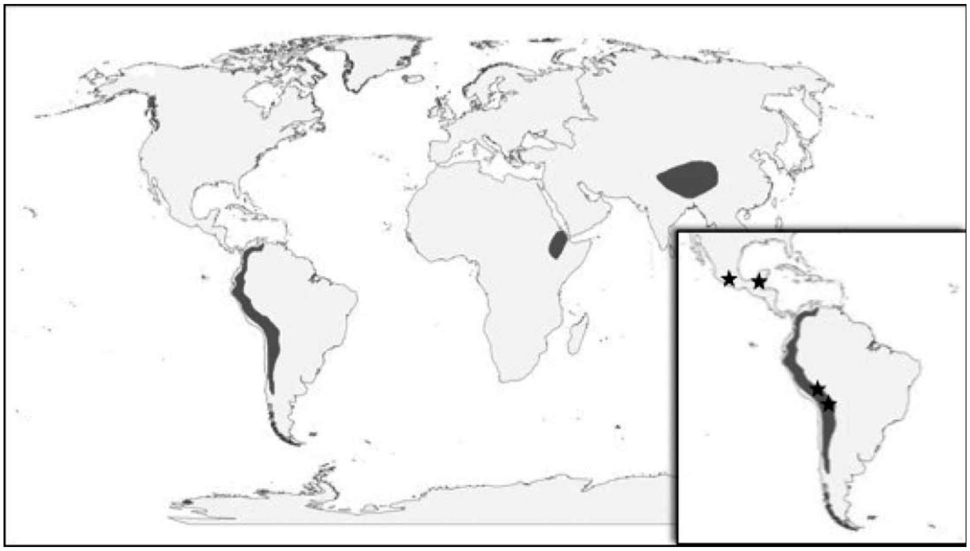

Figure 1. The geography of human adaptation to high altitude. Geographic locations where humans have adapted to life at high altitude are indicated in grey and include the Andean Altiplano of South America, the Tibetan Plateau of Central Asia and the Semien Plateau of Ethiopia. The inset indicates the sampling locations of the four Native American population samples. The populations include Peruvian Quechua, Bolivian Aymara, Nahua, Mixtec and Tlapanec speakers from Guerrero, Mexico, and Maya from the Yucatan Peninsula, Mexico.

expect that genes in this pathway might be involved in adaptation to high altitude.[31] Genomic searches for signatures of natural selection, however, are also a means of aiding the identification of gene function or to expand the current understanding of a gene's function. Several studies of natural selection have helped to identify functional roles for the loci under selection.[32–34] Therefore, we also considered non-*HIF* genes in this analysis.

Materials and methods

Populations

SNP data were generated using 105 individuals of Native American descent (previously reported on by Mao *et al.*[35]). This sample could be further divided into two groups, a high-altitude group and a low-altitude group. The high-altitude group was composed of 50 individuals of Andean descent: 25 Quechua collected in Cerro de Pasco, Peru (4,300 m), and 25 individuals of largely Aymara ancestry collected in La Paz, Bolivia (3,600 m).[36,37] The low-altitude group consisted of Native American lowlanders from Mexico, including 11 Nahua, nine Mixtec and ten Tlapanec individuals collected in Guerrero (1,600 m) and 25 Maya individuals collected in the Yucatan Peninsula (10 m). Sampling locations for each of the Native American samples is shown in Figure 1. Native American population samples, highland and lowland alike, were selected based on the proportion of Native American to European individual genetic ancestry, with persons showing high levels of Native American and low levels of European ancestry chosen for this research. Genetic ancestry was estimated using a panel of ancestry informative markers (AIMs) that distinguish between West African, Northern European and Native American populations.[38,39] Additionally, we included 90 East Asian lowlanders in this research: the 45 Haplotype Mapping Project (HapMap) Han Chinese from Beijing and the 45 HapMap Japanese from Tokyo. In this analysis, we split the high-altitude and low-altitude populations into three population groupings: 1) Andeans (Quechua and Aymara); 2) Mesoamericans (Maya, Nahua, Tlapanec and Mixtec); and 3) East Asians (Han Chinese and Japanese).

In addition to the Native American and East Asian populations, 120 West African and Northern

European individuals from the HapMap project were also genotyped using the Affymetrix 500 K array set. These included 60 Yoruba from Ibadan, Nigeria and 60 individuals from the USA who were of northern and western European ancestry, collected by the Centre d'Etude du Polymorphisme Humain (CEPH). The availability of the HapMap data made it possible to compare the results of our analysis with those of previous studies of natural selection in the same samples. By doing so, we confirm that this genotyping platform is appropriate for the analysis of Andean signatures of natural selection.

Genome scan data

The Affymetrix, Inc. Gene Chip Human mapping 500 K array set was used to generate high-density multi-locus SNP genotype scores. This mapping array has an even distribution across the genome, with an average inter-SNP distance of 5.8 kilobases (kb). It is composed of two arrays named for the restriction enzymes used in the complexity reduction step of the reaction, the *Nsp* array and the *Sty* array. Each array assays approximately 250,000 SNPs. In total, this analysis was conducted using 490,023 autosomal SNPs.

Tests for positive selection

We used four statistics to identify candidate loci showing positive selection in the Andean population: LSBL, ln*RH*, Tajima's D and the WGLRH test.[5,40–42] LSBL, ln*RH* and the WGLRH test were implemented as previously described.[5,41,43] LSBL was calculated for each SNP in the dataset, whereas an overlapping sliding windows approach was taken to calculate ln*RH* and Tajima's D. We used a window size of 200,000 base pairs (bp), moving in 50,000 bp increments along each chromosome for ln*RH* and 100,000 bp with 10,000 bp increments for Tajima's D. Window size was determined by the genome coverage and the marker density of the Affymetrix 500 K array set. Statistical significance for each of LSBL, ln*RH* and Tajima's D was determined by using its respective genome-wide empirical distribution generated by these data. Those loci with *p*E values falling in the top (LSBL) or bottom (ln*RH* and Tajima's D) 5 per cent of the empirical distribution were considered statistically significant ($\alpha = 0.05$). For the WGLRH test, significance was assessed by comparing the relative extended haplotype homozygosity (REHH) of a specific core haplotype with the gamma distribution and applying the false discovery rate (FDR) approach to correct for multiple tests.[44]

For Tajima's D, we compared standardised Tajima's D across windows similar to the integrated haplotype score (iHS) statistic of Voight *et al.*[2] To do so, we used the following equation:

$$\textit{standardised } D = \frac{D_i - \mu(D)}{SD(D)}$$

Where D_i is the Tajima's D calculated for a sliding window in a given population panel (Andean, Mesoamerican or East Asian), μ is the mean Tajima's D for all windows and SD is the standard deviation of Tajima's D for all windows. Using *standardised* D, we identified regions of the genome that were significantly negative in the Andeans. Because we were interested in regions of the genome that have been subject to natural selection in Andeans but not in lowland Native American populations, however, we also wanted to compare the two New World populations—Andeans and Mesoamericans—to identify genomic regions that may have undergone selection in the high-altitude populations but not in the low-altitude populations. To do so, we developed a statistic to summarise the difference in Tajima's D between two populations using the following equation:

$$\textit{Tajima's}\,\mathrm{D}\ \textit{difference} = \frac{(D_{iA} - D_{iB}) - \mu(D_A - D_B)}{SD(D_A - D_B)}$$

Here, D_{iA} is Tajima's D computed for a given sliding window in population A, D_{iB} is Tajima's D computed for a given sliding window in population B, μ is the mean Tajima's D for all windows and SD is the standard deviation of Tajima's D for all windows. Again, Tajima's D was calculated for each population using an overlapping sliding window size of 100,000 bp with 10,000 bp increments. We did not include East Asians in this comparison, in order to eliminate overlooking genomic regions that may have undergone changes in East Asians after the Old World and New World populations split.

Haplotypic phase was determined for each chromosome prior to calculating two of the statistics, Tajima's D and the WGLRH test. The program FastPHASE resolved the haplotypic phase from the unphased genotype data for Tajima's D.[45] Northern Europeans, West Africans, Native Americans and East Asians were phased individually. The high-altitude and low-altitude Native Americans were phased together as one population. Missing genotypes were inferred for all populations. For the WGLRH test, haplotypic phase was computed using the expectation maximisation (EM) algorithm as implemented in the program Haploview.[46] FastPHASE was not used for this test because we strictly followed the WGLRH algorithm as it was designed, which used Haploview for phasing.[41]

Table 1. Number of SNP or SNP window comparisons for each test statistic and their empirical *p*-values at two levels of α.

Test	Number of tests	*p*E = 0.05	*p*E = 0.01
LSBL	490,032	24,502	4,900
ln*RH*	53,251	2,663	533
Tajima's D	263,882	13,194	2,639
WGLRH	43,153	55	NA

LSBL = locus-specific branch length test; ln*RH* = natural logarithm of the ratio of heterozygosities; WGLRH = whole-genome long-range haplotype.

Results

Working with 490,032 SNPs from the Affymetrix, Inc. Gene Chip Mapping 500 K array in 195 individuals from high and low altitudes, we identified gene regions ($\alpha = 0.05$ and $\alpha = 0.01$) that differed significantly between Andeans and two low-altitude populations, Mesoamericans and East Asians, using four statistics that detect departures from neutrality. The statistics included LSBL, ln*RH*, Tajima's D and the WGLRH test. The significant SNP comparisons or SNP windows for each of the four test statistics applied to these data are shown in Table 1. The empirical distribution for LSBL is shown in Figure 2.

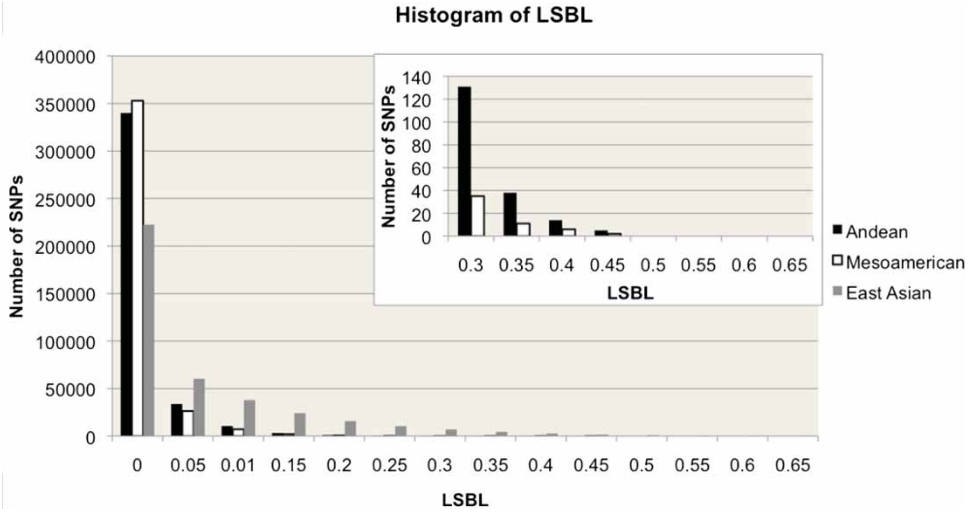

Figure 2. Empirical distribution of LSBL across three populations: Andeans, Mesoamericans and East Asians. The inset shows LSBLs above 0.3 for Andeans and Mesoamericans. These two populations have fewer LSBLs above 0.15 compared with the East Asians, which is expected, given their more recent common ancestry.
LSBL = Locus-specific branch length.

Table 2. Summary of the significant HIF pathway candidate genes for the four test statistics used to detect signatures of positive selection.

Gene name	Test for natural selection			
	LSBL	ln*RH*	Tajima's D difference	WGLRH
ADRA1B	X			
ARNT2	X			
ATP1A1		X		
ATP1A2	X			
CDH1	X	X		
COPS5	X			
CXCR4	X			
EDN1	X			
EDNRA	X	X	X	
EGLN1	X			
EGLN2	X			
ELF2	X	X		
FRAP1		X		
IL1A/ IL1B		X		
IL6	X			
IGFBP1			X	
IGFBP2	X			
MDM2	X			
MMP2	X			
NOS1	X			
NOS2A	X	X	X	
NOTCH1	X			
NRP1			X	
NRP2	X			
POLRA		X		
PIK3CA	X		X	
PIK3CG			X	
PRKAA1	X	X	X	
PRKAA2	X			
SNAI3	X			
SPRY2	X			
TF	X			
TGFA	X			
TNC	X	X		
TNF	X			
VEGF	X		X	

Continued

LSBL = locus-specific branch length; ln*RH* = natural log of the ratio of heterozygosities; WGLRH = whole-genome long-range haplotype.

We identified 14, seven and three HIF pathway genes that fell into the 5 per cent tail of the empirical distribution for LSBL, ln*RH* and Tajima's D difference, respectively. No HIF pathway candidate genes were located in a statistically significant extended haplotype region for the WGLRH test. The SNP genotyping platform used in this analysis, however, did not assay SNPs within the gene boundaries of 13 HIF pathway candidate genes. For this reason, it was important to look 50 kb upstream and downstream of the start and end coordinates of each gene for significant SNPs or SNP windows so as to not exclude a potential candidate gene from analysis. When doing so, 29, ten and eight HIF pathway genes show at least one significant SNP or window for LSBL, ln*RH* and Tajima's D difference, respectively. Table 2 enumerates the significant HIF pathway genes for each test statistic using the 50 kb upstream and downstream definition.

The number and proportion of significant SNPs or sliding windows varied for each gene. For example, all nine ln*RH* windows and six of 28 LSBLs for tenascin C (*TNC*) were statistically significant. By contrast, the gene nitric oxide synthase 2A (*NOS2A*) displayed only one significant ln*RH* window out of seven, and the gene vascular endothelial growth factor (*VEGF*) contained only one significant LSBL among 15 assayed SNPs. Moreover, the only gene with all of the windows in the gene region significant for ln*RH* was *TNC*. None of the HIF pathway candidate genes were statistically significant for all of the test statistics. However, we did identify significant HIF pathway genes using two or three out of the four statistics. *VEGF* was significant for Tajima's D difference and LSBL. *TNC* and cadherin 1 (*CDH1*) were statistically significant for LSBL and ln*RH*. Lastly, three genes, those encoding endothelin receptor type A (*ENDRA*) and protein kinase, AMP-activated, alpha 1 catalytic subunit (*PRKAA1*) and *NOS2A*, were statistically significant for LSBL, ln*RH* and the Tajima's D difference.

To evaluate if HIF pathway genes are over-represented in the 5 per cent tail of the empirical distribution for Andeans, we used Fisher's exact test. We tested the hypothesis that the proportion of significant LSBL values ($\alpha = 0.05$) is higher among HIF pathway candidate genes than among non-HIF genes using a 2×2 contingency table, where the four categories were: significant LSBLs for HIF genes, non-significant LSBLs for HIF genes, significant LSBLs for non-HIF genes and non-significant LSBLs for non-HIF genes. The results indicated that the HIF pathway candidate genes are not over-represented in the 5 per cent tail of the distribution (OR = 0.644 (95 per cent confidence interval 0.538–0.778); $p < 0.001$). A second method of testing if the HIF pathway candidate genes are over-represented in the 5 per cent tail is to compare the LSBL distribution of HIF pathway candidate genes to the LSBL distribution of all non-HIF genes using a one-sided Kolmogorov–Smirnov (K-S) test. Again, the results of this test suggested that the 5 per cent tail of the empirical LSBL distribution is not enriched with HIF genes ($D_{n,m} = 0.0205$; $p = 0.3162$). It is important to note that these results do not preclude particular HIF genes from involvement in genetic adaptation to high altitude. Rather, they denote that the HIF pathway as a whole has not evolved in response to hypoxia among Andeans.

In addition to studying the HIF pathway candidate genes specifically, we also scanned across each chromosome to discern genomic regions showing evidence of reduced variation in Andeans, a hallmark of directional selection. Given the large number of significant tests for LSBL, ln*RH* and Tajima's D using a 5 per cent significance cut-off we restricted our attention to regions with clusters of significant values for one or more test statistics as selection-nominated candidate gene regions. To identify such regions, we calculated the significance of one megabase non-overlapping windows moving across each chromosome for LSBL, ln*RH* and Tajima's D using the hypergeometric distribution. The *p*-value for each window was corrected for multiple tests using the Bonferroni correction.[44] In total, *p*-values for 2,718 windows were calculated for each of the LSBL, ln*RH* and Tajima's D statistics. Significant *p*-values were defined such that one false-positive would be expected for all observed windows. Using this definition, windows for which $p \leq 0.004$ were considered to be statistically significant. The results of this analysis revealed 54 regions displaying extended regions of continuously significant statistics for two or more of the three statistics. Three of these regions located on chromosomes 11, 12 and 15 were significant for all three statistics. Table 3 enumerates the chromosomal regions that were statistically significant for LSBL, ln*RH* and Tajima's D.

Table 3. One megabase windows displaying extended regions of statistical significance for LSBL, ln*RH*, and Tajima's D difference.

Chromosome	Window start	Window end	LSBL *p*-value*	ln*RH* *p*-value*	Tajima's D *p*-value*	Known genes
11	82000000	83000000	0.000000	0.000001	0.000000	19
12	109000000	110000000	0.000000	0.000000	0.000002	41
15	41000000	42000000	0.000000	0.000534	0.000000	70

**p*-values have been corrected for multiple tests using the Bonferroni correction.

Table 4. Summary of the significant core regions containing known genes in the Andean population for the WGLRH test. Core regions that do not contain a known gene are not listed.

Chromosome	SNPs in core region	Haplotype frequency	*p*-value adjusted	Genes in core region*
1	3	0.060	0.032	*KCNK2*
1	2	0.080	0.038	*KCNK2*
1	5	0.070	0.050	*KIAA1026*
1	7	0.070	0.032	*PEX14*
2	2	0.520	0.016	*ERBB4*
2	2	0.540	0.006	*INPP5D*
2	2	0.080	0.038	*TMEM169*
3	2	0.330	0.011	*CPNE4*
4	2	0.051	0.007	*LDB2*
4	3	0.090	0.000	*RBM47*
5	7	0.070	0.034	*FCHO2*
6	3	0.120	0.018	*F13A1*
6	4	0.080	0.001	*F13A1*
6	3	0.120	0.036	*KCNK5*
8	2	0.130	0.036	*ADRA1A*
9	2	0.176	0.047	*ADAMTSL1*
9	2	0.610	0.044	*PTPRD*
9	3	0.070	0.011	*VAV2*
10	2	0.170	0.015	*AK056561*
10	2	0.500	0.021	*FAM107B*
10	2	0.101	0.040	*GFRA1*
10	3	0.540	0.023	*OLAH*
11	2	0.090	0.047	*LDLRAD3*
11	3	0.060	0.002	*PSMD13*
12	2	0.330	0.032	*CHST11*
12	5	0.080	0.035	*SLC4A8*
14	7	0.080	0.002	*KCNK10*
14	3	0.440	0.043	*PPP2R5C*
17	2	0.390	0.036	*C17orf54*
17	6	0.080	0.003	*MPRIP*

Continued

Table 4. Continued

Chromosome	SNPs in core region	Haplotype frequency	*p*-value adjusted	Genes in core region*
18	2	0.450	0.038	*PTPRM*
19	2	0.295	0.002	*GNA15*
20	2	0.100	0.000	*RP5-1022P6.2*
21	2	0.490	0.001	*ERG*

*Genes are listed for the core region only and not the 500 kb extended haplotype regions identified by each core.

The WGLRH test identified 43,153 extended haplotype/core regions throughout the genome in the Andean panel. Only 57 of these regions were statistically significant after identifying ‘flipped’ SNPs and applying an FDR correction for multiple testing. Two of these extended haplotypes were also identified as significant in the Mesoamericans. After removing these two regions from the Andean analysis, 55 significant extended haplotype regions remained. Those significant extended haplotypes containing known genes in their core regions are listed in Table 4. No common core haplotypes were shared between the East Asians and the Andeans. Of the 55 significant 500 kb extended haplotype regions, seven contained SNPs that were statistically significant for LSBL. None of the core regions identified using the WGLRH test overlapped with the statistically significant gene windows for ln*RH* or Tajima’s D.

To validate that this dataset was appropriate for identifying signatures of positive selection in Andean populations, we performed an identical analysis using all four statistical tests for positive selection on the HapMap project populations.[47] The samples included in this analysis have been used in previous genome scans conducted on larger datasets.[2,6] The populations included 60 Yoruba from Ibadan, Nigeria; 60 individuals of northern and western European ancestry from the USA collected by the CEPH; and 90 East Asians from China and Japan. The East Asians used in this analysis corresponded to the East Asians used for the Andean analysis. This analysis identified significant gene regions consistent with previous studies. For example, SNPs found in the gene solute carrier family 24, member 5 (*SLC24A5*)—a gene associated with skin pigmentation and shown to be under positive selection in European populations but not in East Asian and West African populations—possessed statistically significant LSBL and ln*RH* values in the European population;[32,48] however, this was not observed for Tajima’s D difference or the WGLRH test. Another gene, ectodysplasin A receptor (*EDAR*), known to be involved in hair and tooth development, consistently shows evidence of positive directional selection among East Asian populations.[49] For the East Asians in this analysis, SNPs falling within *EDAR* showed statistically significant LSBL values and Tajima’s D window, but non-significant ln*RH* windows. Additionally, it was not identified as a significant core haplotype for the WGLRH test. The absence of significant ln*RH* windows is not surprising in this population of Chinese and Japanese individuals, however, given that the haplotype under selection has not swept to fixation in the Japanese population. By extending our analysis to this group of well-studied old-world populations, we verified that the signatures of selection found in these three populations overlapped with those signatures identified using other SNP datasets, supporting our contention that the SNP dataset and analytical methods used here are appropriate for identifying signatures of positive selection in high-altitude Andeans.

Discussion

Using a dense genome-wide panel of SNPs (Affymetrix 500 K chip), we compared patterns of genetic diversity between high-altitude Andeans, low-altitude Mesoamericans and East Asians to identify selection-nominated candidate genes or gene regions in Andeans. Four tests based on different characteristics of the data were used in our analysis: LSBL, ln*RH*, Tajima's D and the WGLRH test. We selected these complementary methods because each statistic possesses a varying degree of efficacy for identifying signatures of natural selection depending on the allelic background of the populations used in the analysis, the strength of selection and the length of time elapsed since the start of the selective event. Given the aspects of genetic variation summarised by these statistics, it is not expected that the results of tests will overlap. Rather, these methods should be considered as complementary tests that can be useful for the identification of regions under positive selection.

Based on the results of this study, the HIF pathway genes exhibiting the most compelling evidence of positive directional selection across the test statistics are *ENDRA*, *PRKAA1* and *NOS2A*. *ENDRA*, expressed in vascular smooth muscle, encodes a vasoconstrictor whose actions are mediated through endothelin-1.[50] *PRKAA1* encodes a heterotrimeric enzyme belonging to the ancient 5′-AMP-activated protein kinase gene family involved in regulation of cellular ATP (reviewed by Kemp *et al.*[51]). *PRKAA1* functions as a cellular energy sensor under ATP-deprived conditions, such as those that are experienced in hypoxia. Thus, it provides metabolic adaptations to the oxygen-starved cellular environment. *NOS2A*, in combination with additional nitric oxide synthase enzymes, synthesises nitric oxide (NO) from arginine and oxygen. NO increases blood flow in the arteries and helps to regulate blood pressure. Erzurum *et al.*[52] have recently shown that NO production is increased in Tibetans resident at 4,200 m compared with sea-level controls. Recent work has also demonstrated higher uterine artery blood flows during pregnancy in Andean than European high-altitude residents, possibly due to greater uterine artery vasodilation.[37] These studies suggest that vascular factors, not just haematological or pulmonary systems, contribute to altitude adaptation in Tibetan and Andean populations. Here, we showed preliminary evidence of positive selection in *NOS2A* in Andean populations. It would be worthwhile to extend the work conducted in this study to Tibetan populations who show physiological adaptations with respect to NO production, to determine if a similar genetic signal is present in this Himalayan population.

The three chromosomal regions showing extended regions of statistically significant test results are excellent candidates for further study. They include regions on chromosomes 11, 12 and 15. In addition to the two chromosomal regions, the 55 candidate regions identified by the WGLRH test are also strong candidates for further study; however, the WGLRH test only considers derived alleles whose frequencies have risen to >0.85 in the populations under consideration. One problem with only considering those haplotypes with high frequencies of the derived allele is that natural selection could also act to select the ancestral allele, and these signatures cannot be detected with the WGLRH test. Given the low altitudes inhabited by human ancestors, however, it is more likely that selection acted on a novel mutation in one or more of the genes involved in adaptation to altitude, as opposed to an ancestral variant already present in the population. This is especially relevant with regard to the HIF pathway, as this is an evolutionarily ancient system important in embryogenesis, development and homeostasis.

One potential problem with this and other genome scans is that it uses pre-ascertained SNPs to look at the underlying pattern of nucleotide diversity. For example, Tajima's D was first designed for sequence-based tests of selection wherein the nucleotide diversity is known for an entire gene or gene region. Using this statistic on genome scans for natural selection, one must be aware of the ascertainment bias inherent in the analysis. Given the selection criteria of the Affymetrix 500 K panel, uncommon, low-frequency alleles will be under-represented and common alleles will be over-represented. This bias is more likely to miss candidate natural selection genes rather than increase our false-positive rate. Moreover, we used Tajima's D in conjunction with other tests for positive selection so that genes that might have been overlooked using this statistic could have been identified with one or more of the other three statistics. To illustrate this point, consider the genes *TNC* and *CDH1*, which, in our analysis, showed signatures of selection by LSBL and ln*RH*, but not either of the Tajima's D statistics. This pattern is the same as that observed for the gene *SLC24A5*, which is known to be the target of natural or sexual selection in European populations.[32,48] Therefore, it is possible that with unbiased complete sequence data of *TNC* or *CDH1* in Andean and Mesoamerican populations, Tajima's D will reveal a pattern of nucleotide diversity consistent with positive directional selection.

In future work, it would be interesting to compare the overall genetic signals of natural selection found in Andean populations with those found in Tibetan populations, as these two populations are distinct in geographical locale as well as in duration of time living at altitude. Archaeological data indicate that Himalayan populations first inhabited the Tibetan Plateau as early as 25,000 years ago, whereas populations first moved onto the Andean Altiplano 11,000 years ago.[53,54] By understanding how similar environmental pressures with varying evolutionary time frames can result in either the same or different genetic adaptations, we will be better situated to understand the molecular basis for convergent human adaptations. After identification, all putative natural selection regions identified in Andeans and Tibetans must be confirmed by further research, such as genotype/phenotype association studies and functional assays.

Acknowledgments

We would like to thank the people of high altitude who participated in this research and two anonymous reviewers for helpful comments on the manuscript. A.W.B was supported by a National Science Foundation Graduate Research Fellowship. This work was supported by the National Science Foundation (grant number 0622337 to A.W.B); the Wenner-Gren Foundation (grant number 7538 to A.W.B); and the National Institutes of Health (grant numbers 079647, 001188 to L.G.M.). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Altshuler, D. and The International HapHap Consortium (TIHC) (2005), 'A haplotype map of the human genome', *Nature* Vol. 437, pp. 1299–1320.
2. Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006), 'A map of recent positive selection in the human genome', *PLoS Biol.* Vol. 4, p. e72.
3. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E. *et al.* (2005), 'Whole-genome patterns of common DNA variation in three human populations', *Science* Vol. 307, pp. 1072–1079.
4. Akey, J.M., Zhang, G., Zhang, K., Jin, L. *et al.* (2002), 'Interrogating a high-density SNP map for signatures of natural selection', *Genome Res.* Vol. 12, pp. 1805–1814.
5. Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A. *et al.* (2004), 'The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs', *Hum. Genomics* Vol. 1, pp. 274–286.
6. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J. *et al.* (2007), 'Genome-wide detection and characterization of positive selection in human populations', *Nature* Vol. 449, pp. 913–918.
7. Hancock, A.M., Witonsky, D.B., Gordon, A.S., Eshel, G. *et al.* (2008), 'Adaptations to climate in candidate genes for common metabolic disorders', *PLoS Genet.* Vol. 4, p. e32.
8. McEvoy, B., Beleza, S. and Shriver, M.D. (2006), 'The genetic architecture of normal variation in human pigmentation: An evolutionary perspective and model', *Hum. Mol. Genet.* Vol. 15 Spec No 2, pp. R176–181.
9. Baker, P.L. (1976), *Man in the Andes: A Multidisciplinary Study of High-Altitude Quechua*, Dowden, Hutchinson, and Ross, Inc., Stroudsbourg, PA, USA.
10. Schull, W.J. and Rothhammer, F. (1990), *The Aymara: Strategies in Human Adaptation to a Rigorous Environment*, Kluwer Academic Publishers, Boston, MA, USA.
11. Moore, L.G., Shriver, M., Bemis, L., Hickler, B. *et al.* (2004), 'Maternal adaptation to high-altitude pregnancy, an experiment of nature: A review', *Placenta* Vol. 25(Suppl. A), pp. S60–S71.
12. Beall, C., Brittenham, G., Macuaga, F. and Barragan, M. (1990), 'Variation in hemoglobin concentration among samples of high altitude natives in the Andes and the Himalayas', *Am. J. Hum. Biol.* Vol. 2, pp. 639–651.
13. Beall, C.M., Decker, M.J., Brittenham, G.M., Kushner, I. *et al.* (2002), 'An Ethiopian pattern of human adaptation to high-altitude hypoxia', *Proc. Natl. Acad. Sci. USA* Vol. 99, pp. 17215–17218.
14. Beall, C.M. and Goldstein, M.C. (1987), 'Hemoglobin concentration of pastoral nomads permanently resident at 4,850–5,450 meters in Tibet', *Am. J. Phys. Anthropol.* Vol. 73, pp. 433–438.
15. Adams, W.H. and Strang, L.J. (1975), 'Hemoglobin levels in persons of Tibetan ancestry living at high altitude', *Proc. Soc. Exp. Biol. Med.* Vol. 149, pp. 1036–1039.
16. Beall, C.M., Brittenham, G.M., Strohl, K.P., Blangero, J. *et al.* (1998), 'Hemoglobin concentration of high-altitude Tibetans and Bolivian Aymara', *Am. J. Phys. Anthropol.* Vol. 106, pp. 385–400.
17. Beall, C.M. and Reichsman, A.B. (1984), 'Hemoglobin levels in a Himalayan high altitude population', *Am. J. Phys. Anthropol.* Vol. 63, pp. 301–306.
18. Zhuang, J., Droma, T., Sun, S., Janes, C. *et al.* (1993), 'Hypoxic ventilatory responsiveness in Tibetan compared with Han residents of 3,658 m', *J. Appl. Physiol.* Vol. 74, pp. 303–311.
19. Groves, B.M., Droma, T., Sutton, J.R., McCullough, R.G. *et al.* (1993), 'Minimal hypoxic pulmonary hypertension in normal Tibetans at 3,658 m', *J. Appl. Physiol.* Vol. 74, pp. 312–318.
20. Hornbein, T.F. and Schoene, R.B. (eds) (2001), *High Altitude: An Exploration of Human Adaptation*, Marcel Dekker, Inc., New York, NY, USA.
21. Rupert, J.L., Kidd, K.K., Norman, L.E., Monsalve, M.V. *et al.* (2003), 'Genetic polymorphisms in the renin-angiotensin system in high-altitude and low-altitude Native American populations', *Ann. Hum. Genet.* Vol. 67, pp. 17–25.

22. Moore, L.G., Zamudio, S., Zhuang, J., Droma, T. *et al.* (2002), 'Analysis of the myoglobin gene in Tibetans living at high altitude', *High Alt. Med. Biol.* Vol. 3, pp. 39–47.
23. Hochachka, P.W. and Rupert, J.L. (2003), 'Fine tuning the HIF-1 "global" O2 sensor for hypobaric hypoxia in Andean high-altitude natives'. *Bioessays* Vol. 25, pp. 515–519.
24. Suzuki, K., Kizaki, T., Hitomi, Y., Nukita, M. *et al.* (2003), 'Genetic variation in hypoxia-inducible factor 1alpha and its possible association with high altitude adaptation in Sherpas'. *Med. Hypotheses* Vol. 61, pp. 385–389.
25. Rupert, J.L., Monsalve, M.V., Devine, D.V. and Hochachka, P.W. (2000), 'Beta2-adrenergic receptor allele frequencies in the Quechua, a high altitude native population'. *Ann. Hum. Genet.* Vol. 64, pp. 135–143.
26. Rupert, J.L., Devine, D.V., Monsalve, M.V. and Hochachka, P.W. (1999), 'Angiotensin-converting enzyme (ACE) alleles in the Quechua, a high altitude South American native population'. *Ann. Hum. Biol.* Vol. 26, pp. 375–380.
27. Bigham, A.W., Kiyamu, M., Leon-Velarde, F., Parra, E.J. *et al.* (2008), 'Angiotensin-converting enzyme genotype and arterial oxygen saturation at high altitude in Peruvian Quechua', *High Alt. Med. Biol.* Vol. 9, pp. 167–178.
28. Malacrida, S., Katsuyama, Y., Droma, Y., Basnyat, B. *et al.* (2007), 'Association between human polymorphic DNA markers and hypoxia adaptation in Sherpa detected by a preliminary genome scan', *Ann. Hum. Genet.* Vol. 71, pp. 630–638.
29. Kramer, A.A. (1992), 'Heritability estimates of thoracic skeletal dimensions for high-altitude Peruvian populations', in: Melton, T.W. and Eckhardt, R.B. (eds), *Populations Studies on Human Adaptation and Evolution in the Peruvian Andes*, The Pennsylvania State University Press, University Park, PA, USA, pp. 25–49.
30. Beall, C.M., Blangero, J., Williams-Blangero, S. and Goldstein, M.C. (1994), 'Major gene for percent of oxygen saturation of arterial hemoglobin in Tibetan highlanders', *Am. J. Phys. Anthropol.* Vol. 95, pp. 271–276.
31. Moore, L.G., Shriver, M., Bemis, L. and Vargas, E. (2006), 'An evolutionary model for identifying genetic adaptation to high altitude', *Adv. Exp. Med. Biol.* Vol. 588, pp. 101–118.
32. Lamason, R.L., Mohideen, M.A., Mest, J.R., Wong, A.C. *et al.* (2005), 'SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans', *Science* Vol. 310, pp. 1782–1786.
33. Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F. *et al.* (2007), 'Convergent adaptation of human lactase persistence in Africa and Europe', *Nat. Genet.* Vol. 39, 31–40.
34. Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T. *et al.* (2004), 'Genetic signatures of strong recent positive selection at the lactase gene', *Am. J. Hum. Genet.* Vol. 74, pp. 1111–1120.
35. Mao, X., Bigham, A.W., Mei, R., Gutierrez, G. *et al.* (2007), 'A genomewide admixture mapping panel for Hispanic/Latino populations', *Am. J. Hum. Genet.* Vol. 80, pp. 1171–1178.
36. Brutsaert, T.D., Parra, E.J., Shriver, M.D., Gamboa, A. *et al.* (2003), 'Spanish genetic admixture is associated with larger V(O2), max decrement from sea level to 4338 m in Peruvian Quechua', *J. Appl. Physiol.* Vol. 95, pp. 519–528.
37. Wilson, M.J., Lopez, M., Vargas, M., Julian, C. *et al.* (2007), 'Greater uterine artery blood flow during pregnancy in multigenerational (Andean), than shorter-term (European), high-altitude residents', *Am. J. Physiol. Regul. Integr. Comp. Physiol.* Vol. 293, pp. R1313–1324.
38. Bonilla, C., Shriver, M.D., Parra, E.J., Jones, A. *et al.* (2004), 'Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York city', *Hum. Genet.* Vol. 115, pp. 57–68.
39. Shriver, M.D., Parra, E.J., Dios, S., Bonilla, C. *et al.* (2003), 'Skin pigmentation, biogeographical ancestry and admixture mapping', *Hum. Genet.* Vol. 112, pp. 387–399.
40. Storz, J.F., Payseur, B.A. and Nachman, M.W. (2004), 'Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa', *Mol. Biol. Evol.* Vol. 21, pp. 1800–1811.
41. Zhang, C., Bailey, D.K., Awad, T., Liu, G. *et al.* (2006), 'A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations', *Bioinformatics* Vol. 22, pp. 2122–2128.
42. Tajima, F. (1989), 'Statistical method for testing the neutral mutation hypothesis by DNA polymorphism', *Genetics* Vol. 123, pp. 585–595.
43. Schlotterer, C. (2002), 'A microsatellite-based multilocus screen for the identification of local selective sweeps', *Genetics* Vol. 160, pp. 753–763.
44. Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *J. R. Stat. Soc.* Vol. 57, pp. 289–300.
45. Scheet, P. and Stephens, M. (2006), 'A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase', *Am. J. Hum. Genet.* Vol. 78, pp. 629–644.
46. Qin, Z.S., Niu, T. and Liu, J.S. (2002), 'Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms', *Am. J. Hum. Genet.* Vol. 71, pp. 1242–1247.
47. International HapMap Consortium (2003), 'The International HapMap Project', *Nature* Vol. 426, pp. 789–796.
48. Norton, H.L., Kittles, R.A., Parra, E., McKeigue, P. *et al.* (2007), 'Genetic evidence for the convergent evolution of light skin in Europeans and East Asians', *Mol. Biol. Evol.* Vol. 24, pp. 710–722.
49. Carlson, C.S., Thomas, D.J., Eberle, M.A., Swanson, J.E. *et al.* (2005), 'Genomic regions exhibiting positive selection identified from dense genotype data', *Genome Res.* Vol. 15, pp. 1553–1565.
50. Arai, H., Hori, S., Aramori, I., Ohkubo, H. *et al.* (1990), 'Cloning and expression of a cDNA encoding an endothelin receptor', *Nature* Vol. 348, pp. 730–732.
51. Kemp, B.E., Stapleton, D., Campbell, D.J., Chen, Z.P. *et al.* (2003), 'AMP-activated protein kinase, super metabolic regulator', *Biochem. Soc. Trans.* Vol. 31, pp. 162–168.
52. Erzurum, S.C., Ghosh, S., Janocha, A.J., Xu, W. *et al.* (2007), 'Higher blood flow and circulating NO products offset high-altitude hypoxia among Tibetans', *Proc. Natl. Acad. Sci. USA* Vol. 104, pp. 17593–17598.
53. Moseley, M. (2001), *The Incas and their Ancestors*, Thames and Hudson, London, UK.
54. Aldenderfer, M. (2003), 'Moving up in the world', *Am. Sci.* Vol. 91, pp. 542–549.