

Classification: Biological Sciences, Anthropology

Title: Association of *EGLN1* gene with high aerobic capacity of Peruvian Quechua at high altitude.

Authors: Tom D. Brutsaert^{1*}, Melissa Kiyamu², Gianpietro Elias Revollendo¹, Jenna L. Isherwood³, Frank S. Lee⁴, Maria Rivera-Ch.², Fabiola Leon-Velarde², Sudipta Ghosh⁵, Abigail W. Bigham³.

Affiliations:

¹Syracuse University, Department of Exercise Science, 820 Comstock Ave., Syracuse, NY, 13244.

²Universidad Peruana Cayetano Heredia, Lima, Peru.

³University of Michigan, Department of Anthropology, Ann Arbor, MI. 1085 S. University Ave., Ann Arbor, MI 48109-1107.

⁴Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, 19104.

⁵Department of Anthropology, North-Eastern Hill University, Shillong, India.

*Correspondence to: Tom D. Brutsaert, Dept. of Exercise Science, 820 Comstock Ave., Syracuse, NY, 13244. e-mail:tdbrutsa@syr.edu, phone:315-443-9696

Abstract: Highland native Andeans have resided at altitude for millennia. They display high aerobic capacity (VO_{2max}) at altitude which may be a reflection of genetic adaptation to hypoxia. Previous genome-wide (GW) scans for natural selection have nominated *EGLN1* as a candidate gene. The encoded protein, EGLN1/PHD2, is an O_2 sensor that controls levels of the Hypoxia Inducible Factor- α (HIF- α), which regulates the cellular response to hypoxia. From GWAS and ANCOVA performed on a total sample of $n=429$ Peruvian Quechua and $n=94$ US lowland referents, we identified five *EGLN1* SNPs associated with higher VO_{2max} ($l \cdot min^{-1}$ and $ml \cdot min^{-1} \cdot kg^{-1}$) in hypoxia (rs1769793, rs2064766, rs2437150, rs2491403, rs479200). For four of these SNPs, Quechua had the highest frequency of the advantageous (high VO_{2max}) allele compared to 25 diverse lowland comparison populations from the 1000 Genomes project. Genotype effects were substantial with high versus low VO_{2max} genotype categories differing by $\sim 11\%$ (e.g., for rs1769793 SNP genotype TT=34.2 $ml \cdot min^{-1} \cdot kg^{-1}$ versus CC=30.5 $ml \cdot min^{-1} \cdot kg^{-1}$). To guard against spurious association, we controlled for population stratification. Findings were replicated for *EGLN1* SNP rs1769793 in an independent Andean sample collected in 2002. These findings contextualize previous reports of natural selection at *EGLN1* in Andeans, and support the hypothesis that natural selection has increased the frequency of an *EGLN1* causal variant that enhances O_2 delivery or use during exercise at altitude in Peruvian Quechua.

Key words: Hypobaric hypoxia, human adaptation, genetic, Quechua.

Significance Statement: Andean highland native populations, like the Quechua of Peru, have enhanced exercise capacity at altitude and may be genetically adapted to altitude. In the first study of its kind, we identified five genetic markers near the *EGLN1* gene that were associated with higher aerobic capacity (VO_{2max}) in hypoxia. *EGLN1* encodes for a protein that controls the level of the Hypoxia Inducible Factor- α , which in turn regulates the cellular hypoxic response. Advantageous SNP alleles were associated with a significantly higher VO_{2max} and were found at higher frequency in Quechua compared to lowland populations. These results add further context to previous studies that have provided evidence of natural selection at the *EGLN1* locus in Andeans.

INTRODUCTION:

Low O₂ availability at altitude is an environmental stressor that has a negative impact on human reproductive capacity (1) and physical work capacity (2). Highland native Andean (Quechua) populations have resided at altitude for millennia (3) and may be genetically adapted to hypobaric hypoxia (4-6). However, direct evidence for genetic adaptation is lacking and only limited data exist linking genotype-to-phenotype. One line of indirect evidence comes from comparative physiological studies that have measured high maximal aerobic capacity (VO₂max, ml·min⁻¹·kg⁻¹) in Andean natives tested at high altitude (5, 7-17). Indeed, an adaptive response reflected in enhanced VO₂max is a long-standing idea that can be traced back to the early Spanish accounts of impressive physical tolerance to altitude among the Inca in the 1500's (18). As a phenotype, VO₂max reflects the integrated functioning of respiratory, cardiovascular, and muscular systems, all of which are significantly stressed during exercise at altitude. Thus, it is reasonable to hypothesize directional selection on favorable molecular and/or physiological phenotypes resulting in high frequency Andean population alleles that enhance systemic O₂ delivery and/or cellular O₂ use.

For compelling reasons, we focused *a priori* attention on the *Egl-9 homologue 1* gene (*EGLN1*, also known as *PHD2*). First, previous genome-wide scans have identified *EGLN1* as a likely target of natural selection in both Andeans and Tibetans (19-22). Second, unique haplotypes of *EGLN1*, as well as a second gene, *EPAS1*, which encodes the hypoxia-inducible factor (HIF-2α), have been associated with low hemoglobin concentration [Hb] in Tibetans (22-24). Third, *EGLN1* plays a central role in the cellular hypoxic response (25). That is, the encoded protein, prolyl hydroxylase (PHD2), is a key oxygen sensor that controls—via prolyl hydroxylation—the protein levels of the Hypoxia Inducible Factor-α (HIF-α), which in turn is the master transcriptional regulator of the hypoxic response (26). Fourth, genetic variation within the *EGLN1* gene is related to physiological variation. For example, heterozygous loss-of-function *EGLN1* mutations lead to erythrocytosis in humans (27, 28) as well as erythrocytosis and increased respiration under hypoxic conditions in mice (21, 29, 30).

We used a large data set of Peruvian Quechua to test for genetic association between *EGLN1* SNP variants and the VO₂max measured in hypoxia. Our sample comprised n=429 Quechua, and n=94 non-Hispanic white lowlanders from Syracuse, NY, USA. We performed two complementary analyses on the full sample, including a conventional GWAS on the VO₂max phenotype, and a targeted ANCOVA focused *a priori* on the *EGLN1* gene.

MATERIALS AND METHODS:

Full sample selection and methodological details are given in the SI Appendix. The genetic sample (n=523 total) included four sub-groups: (1) Quechua-High Altitude Residents (HAR, n=195) from Cerro de Pasco, Peru at 4,338 m above sea-level; (2) Quechua-Migrants (M, n=111), born at altitude but migrated permanently to sea-level from Lima, Peru; (3) Quechua-Born at sea level from Lima (BSL, n=123); and (4) non-Hispanic whites (n=94) from Syracuse, NY, USA, at ~140 m. Roughly equal numbers of males and females were recruited between the ages of 18-35. Participants provided written informed consent for study procedures approved by the *Syracuse University Office of Research Integrity and Protections*, and the *Research Ethics Committee of the Universidad Peruana Cayetano Heredia (UPCH), Lima, Peru*. The study also was approved by *The University of Michigan Institutional Review Board*. We replicated the association results for rs1769793 using data from a previously recruited Quechua cohort from Cerro de Pasco, Peru. These data were collected in 2001-2 from 67 male and female Quechua by the same investigators using the same exercise testing equipment and protocol (5, 31, 32).

For Quechua-HAR, exercise testing was conducted in Cerro de Pasco, Peru. For Quechua-M, Quechua-BSL, and Syracuse, exercise testing was conducted under simulated altitude conditions by lowering the fractional concentration of O₂ (F_iO₂) to approximately 0.126 at sea-level. VO₂max was measured using a graded testing protocol and a metabolic cart on a cycle ergometer. Only participants achieving a true VO₂max (i.e., a respiratory exchange ratio >1.1 and maximal heart rate with 10% of predicted maximum) were retained for genetic analysis.

Microarray genotype data were generated using the Affymetrix (Santa Clara, CA) Axiom Biobanking Array featuring approximately 610,000 markers. The Biobanking Array contains 29 markers in and around (50KB upstream and downstream) *EGLN1*. Of these 29 *EGLN1* markers, 6 met the criteria for association testing. In addition, we manually genotyped two *EGLN1* SNPs (rs479200, rs480902) which exhibited substantial differences in minor allele frequency (MAF) compared to Mexican controls from the 1,000 Genomes Phase 3 (Table 2). This resulted in a final selection of 8 *EGLN1* SNPs for genetic analysis.

We tested *EGLN1* SNP associations with VO₂MAX using two complementary approaches: 1) genome-wide association (GWAS) and 2) *a priori* analysis of covariance (ANCOVA). Both analyses used the entire cohort of n=523. For GWAS, we tested 215,512 autosomal variants using standard linear regression in Plink version 1.9 (<https://www.cog-genomics.org/plink2/>). Genome-wide significance was assessed by applying the false discovery rate of Benjamini and Hochberg. Sex, group, age, and height were included as covariates. Population stratification was controlled by introducing into statistical models the first principle component (PC) of a principal component analysis (PCA) performed on the array data (33). For ANCOVA, we controlled for sex, age, body weight, group, and the first five PCs of the PCA. We applied a Bonferroni correction for multiple testing with a p-value cutoff of p<0.00625 ($\alpha = 0.05$, 8 tests). If the main SNP effect was significant at p<0.00625, interactions with other factors were examined and retained using the conventional p-value cutoff of p<0.05.

Genotyping data are available through the dryad digital repository (link to data here). Associated protocols and code are available through direct communication with the corresponding author. Ethical approval is required for access to the physiological data.

RESULTS:

Sample characteristics are summarized in SI Appendix, Table S1. There were significant differences between sub-groups in body size and composition, [Hb], and VO₂max, but these were expected given differences in ethnicity, place of birth, altitude of residence, and acclimatization state.

Full GWAS results are reported in SI Appendix, Dataset S1. After correction for genome-wide significance, no SNPs were significantly associated with VO₂max. Power analysis revealed that sample size was underpowered to detect an association with VO₂max at 80% power (34). Our genomic inflation factor (GIF) for the combined Quechua and Syracuse dataset was 6.266 before controlling for covariates. When adjusting for all five covariates including PC1, our GIF drops to 1.043 for the combined Quechua and Syracuse dataset. When genomic inflation was measured among Quechua or Syracuse participants independently, the GIF was 1.012 and 1.021, respectively.

Specific *EGLN1* GWAS results are presented in detail in Table 1. Of the 8 *EGLN1* SNPs available for analysis, SNP rs1769793 was the most significant via GWAS (uncorrected p-value = 0.002) and ranked 533 out of 215,512 SNPs included i.e., within the top 0.25% of SNPs tested. Three additional *EGLN1* SNPs, rs2491403, rs2064766, and rs2437150, were also significantly associated with VO₂max via GWAS (uncorrected p-value <0.05) and ranked within the top 4% of SNPs tested.

From ANCOVA, five of 8 *EGLN1* SNPs were associated with VO₂max and showed similar association patterns, either as a SNP main-effect or as a SNP-by-study sub-group interaction (Table 2). The remaining three SNPs were not significant. Interactions resulted from SNP genotype effects within study sub-groups that were of degree rather than direction. That is, SNP associations were evident as a trend in the Syracuse referent population as well. SNP associations with VO₂max were not spurious as they persisted even after control for stratification. The most compelling association was for *EGLN1* SNP rs1769793, which was significant after Bonferroni correction for multiple testing ($p=0.00625$, $\alpha=0.05$, 8 tests, Table 2). The four other SNPs (rs2064766, rs2437150, rs2491403, and rs479200) were significant by the conventional $p<0.05$ criteria either as a main or interaction effect (Table 2).

Marginal mean values of VO₂max (ml·min⁻¹·kg⁻¹) by rs1769793 SNP genotype are shown in Fig. 1a. The rs1769793 genotype differences in VO₂max were physiologically significant, with TT=34.16±0.98 ml·min⁻¹·kg⁻¹, CT=31.98±0.40 ml·min⁻¹·kg⁻¹, and CC= 30.50±0.53 ml·min⁻¹·kg⁻¹. The rs1769793 genotype frequencies were different between Quechua and Syracuse populations, with 29% of Quechua in the high VO₂max genotype category (TT) compared to only 2% of Syracuse Fig. 1b. For broader global context, we compared Quechua allele frequencies for the adaptive high VO₂max allele (T) for rs1769793 to population mean values available from 1000 Genomes Phase 3 (Fig. 1c) (<https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/>). The Quechua sample has the highest frequency of T=0.55 compared to all other populations included in the 1000 Genomes data (Fig. 1c).

Table 3 shows the specific ANCOVA model for rs1769793. Body-size, age, and sex accounted for the majority of the variance in VO₂max. After covariate control, rs1769793 SNP genotype

(CC, CT, or TT) explained 3.4% of the variance in VO_2max as a main effect ($p=0.000241$) and 2.9% of the variance as an interaction effect ($p=0.024$). Statistical models controlled for weight as a covariate according to the recommended practice (35), but results were identical whichever method of body-size control was applied. The same general association pattern for rs1769793 was evident for the four additional significant SNPs, rs2064766, rs2437150, rs2491403, and rs479200 (SI Appendix, Figs. S2-S3, Tables S2-6). That is: 1) Quechua were over-represented in the high VO_2max SNP genotype categories (SI Appendix, Fig. S2, odds ratio [OR] fisher exact p -values ≤ 0.001 ; four SNPs tested), and 2) Quechua showed high frequency of the putative adaptive allele at each locus compared to 1000 Genomes Phase 3 data (SI Appendix, Fig. S3 a-d). Indeed, the Quechua sample showed the highest frequency worldwide of the adaptive allele across three of these loci (rs2064766, rs2437150, and rs479200).

We also tested models that introduced (pairwise) two SNPs into the same statistical model. The introduction of the most significant SNP (rs1769793) eliminated the effect of the other SNPs with lower significance. In addition, SNP rs1769793 showed low to moderate linkage disequilibrium (LD) with these four SNPs (r^2 range = 0.31-0.41) while the other four SNPs showed high to moderate LD with each other ($r^2 > 0.60$) (Fig. 2). Taken together, these results suggest that the five SNPs mark a single causal genetic locus.

We identified individuals harboring all of the alleles associated with higher VO_2max for all five SNPs at rs1769793, rs2064766, rs2437150, rs2491403, and rs479200 (T, G, C, C, and G respectively), and compared them to individuals harboring none of the adaptive SNP alleles. Among the Quechua, 18.7% (85 of 454) of participants harbored all high VO_2max alleles versus only 2.1% of Syracuse participants (2 of 97). Similarly, only 5.9% (27 of 454) of Quechua had none of the high VO_2max alleles compared to 33% (32 of 97) of Syracuse participants. Mean values of VO_2max were $33.97 \text{ ml}\cdot\text{min}^{-1}\cdot\text{kg}^{-1}$ versus $30.42 \text{ ml}\cdot\text{min}^{-1}\cdot\text{kg}^{-1}$ in participants with all or none of the high VO_2max alleles, respectively ($p=0.006$). This difference (~13%) in mean VO_2max by haplotype was similar to the 11% reported for individual SNPs. Thus, SNP effects were not additive between loci suggesting again that these five SNPs mark a single causal genetic locus.

In the replication sample, rs1769793 was significant ($p=0.033$) and explained 10.6% of the variance in VO_2max (SI Appendix, Table S7). Genotype differences in VO_2max were similar to the differences in the larger current sample, with TT= $40.59\pm 1.04 \text{ ml}\cdot\text{min}^{-1}\cdot\text{kg}^{-1}$, TC= $38.46\pm 1.10 \text{ ml}\cdot\text{min}^{-1}\cdot\text{kg}^{-1}$, and CC $35.85\pm 2.01 \text{ ml}\cdot\text{min}^{-1}\cdot\text{kg}^{-1}$ (Figure 1a). The frequency of the advantageous T-allele was 68%, which is higher than the 55% in the current Quechua sample and the 54% documented in the 1000-genomes “Peruvians from Lima (PEL)” sample (Figure 1c).

As yet unidentified SNPs residing in the coding region of *EGLN1* could be driving the rs1769793 association. This is important given that Tibetan *EGLN1* sequencing studies have revealed two coding region SNPs, D4E (rs186996510) and C127S (rs12097901), that are enriched in Tibetans and are in strong linkage disequilibrium (LD) (36-38). In fact, the D4E/C127S haplotype is enriched over 80-fold in Tibetans compared to Han Chinese. Furthermore, this double amino acid substitution is associated with the low [Hb] phenotype characteristic of Tibetan adaptation to altitude, although differing models for the functional effects of these changes have been proposed (21). To identify coding sequence variation in Quechua, we sequenced all five *EGLN1* exons and approximately 1 kb upstream and downstream of the first and last exons in 12 individuals from the 2001-2 cohort (5, 31, 32). The D4E and C127S missense SNPs were present at frequencies of 12.5%, but increased only by 11.7% and 0.8%, respectively compared to Mexican controls from the 1000 Genomes Project (1KG). In contrast

to Tibetans, they were not in LD ($r^2 = 0.012$). We genotyped the D4E (rs186996510) and C127S (rs12097901) SNPs in the current cohort of n=429 Peruvian Quechua using PCR and restriction enzyme digestion (SI Appendix, Table S9). The D4E SNP was monomorphic, whereas the minor allele frequency (MAF) of C127S was 0.08% (minor allele = C). ANCOVA revealed no association for C127S (rs12097901) with $VO_2\text{max}$. Our sequencing efforts revealed no other missense SNPs or SNPs affecting splicing donor or acceptor sites. An additional five variable sites were identified in either the 5' or 3' UTR, but in no case was the difference in MAF > 0.1 compared to Mexican controls. Taken together, it appears unlikely that the potentially functional variant of the Andean *EGLN1* affects the coding sequence, splicing, or translation.

DISCUSSION:

This study reveals a genetic association between an unknown *EGLN1* variant/haplotype and $VO_2\text{max}$ measured in hypoxia. Results contextualize previously published evidence of natural selection at *EGLN1* in Andeans (19-21, 39). Indeed, the genotype-phenotype association and high frequency of several SNP alleles associated with higher $VO_2\text{max}$ in Quechua provide support for the hypothesis that Andeans are genetically adapted to altitude. GWAS was not definitive on the issue of genetic association, as 11,922 SNPs emerged as significant by conventional p-value criteria, but were not significant when corrected for multiple comparisons. These included four *EGLN1* SNPs, including SNP rs1769793 which ranked in the top 0.25% of all SNPs tested at $p=0.002$ (uncorrected). The non-significant p-values for GWAS were not unexpected given our relatively small sample. Thus, we used an additional *a priori* ANCOVA approach to test for *EGLN1* genetic association, allowing deeper interrogation of interaction effects and genotype differences. This was justified on the promising GWAS results and previous research providing compelling reasons to focus *a priori* on *EGLN1* (see paper introduction). From ANCOVA, five SNPs showed strikingly similar association patterns with $VO_2\text{max}$ in hypoxia. Four of these SNPs were the same as those identified by GWAS. These five SNPs were associated with $VO_2\text{max}$ after control for population stratification, and we replicated the strongest SNP signal for rs1769793 in an independent cohort of Quechua.

Significant SNPs explained from ~2-11% of the variance in $VO_2\text{max}$ and genotype effects were large, with differences between high vs. low $VO_2\text{max}$ genotypes of ~11-13% (Fig. 1a). For perspective, moderate aerobic training produces mean gains of ~14% (40). More important, Quechua were over-represented in all high $VO_2\text{max}$ SNP-genotype categories compared to Syracuse lowland natives, consistent with the hypothesis of directional selection on favorable variants. For example, for rs1769793, 29% of Quechua were TT (=high $VO_2\text{max}$) compared to only 2% of Syracuse participants (Fig. 1c). Considered another way, more than half of the Syracuse sample (56%) were CC (=low $VO_2\text{max}$) compared to only 20% of Quechua. In terms of allele frequencies, at four of the five SNPs, Quechua had the highest frequency of the high $VO_2\text{max}$ allele compared to samples from 1000 Genomes (Fig. 1c).

The argument for genetic adaptation depends on several criteria: 1) identification of an adaptive phenotype; 2) association of the phenotype with a gene or genes; 3) allele frequencies that are consistent with the mode of selection hypothesized; 4) statistical genetic evidence of past natural selection on the genomic region harboring the gene or genes; and 5) association of genotypes and phenotypes with direct measures of fitness i.e., fertility and/or mortality. The present study provides evidence to support criteria 1-3, while previous work by our group and others supports criteria 4 for *EGLN1* (19-21, 39). Direct comparison of the data here with published data showing evidence of natural selection at *EGLN1* is not possible given that the variants included on the Affymetrix Biobanking Array are not the same as the variants included on the Affy 6.0 array used previously. Nonetheless, among Peruvians from the 1,000 Genomes Project, our top associated SNP with $VO_2\text{max}$, rs1769793, is in high LD ($r^2 = 0.977$) with the highest-ranking SNP, rs1769792, in the locus specific branch length analysis (LSBL) from (19). This is indirect evidence that the previously identified region under natural selection is the same region identified in this study. Thus, results are consistent that high $VO_2\text{max}$ was likely a target of past selection in the Andes. However, it is also possible that high $VO_2\text{max}$ is secondary to selection on a related and as yet unidentified phenotype. A full understanding will require elucidation of the specific causal genetic pathway involving *EGLN1* and other genes in the HIF-

system that determine VO_2max at altitude. Additionally, criteria 5 has not been examined in the literature.

The current study is one of only two association studies on Andeans (4). Thus, it is of interest to compare findings with the association studies of Tibetans. Like our study, several Tibetan studies identify SNPs in HIF system genes (*EGLN1*, *EPAS1*) associated with a phenotype related to oxygen transport i.e., a low concentration of [Hb] (22-24). Unlike high VO_2max , the adaptive benefit of lower [Hb] in Tibetans is not clear given opposing effects of [Hb] on O_2 delivery i.e., increased blood flow and tissue perfusion versus decreased blood oxygen content. Interestingly, one study shows higher VO_2max in Tibetans with lower [Hb] (41) and underscores a strength of the current study. That is, inferring genetic adaptation depends on identifying a phenotype with adaptive benefit (criteria 1). VO_2max , as marker of physical work capacity, meets this evolutionary standard (5, 7-17). A disadvantage of VO_2max is that the measure is influenced by physical activity patterns, but we have no *a priori* reason to suspect correlation of SNP genotypes activity patterns. Also, group differences in VO_2max were controlled via the statistical approach used. It is also of interest to note that the highest VO_2max values were recorded on Quechua participants, not on lowlanders. For example, one Quechua male had a value of $60 \text{ ml}\cdot\text{min}^{-1}\cdot\text{kg}^{-1}$ at 4,338 m! This is much higher than the highest Syracuse male value of $51.7 \text{ ml}\cdot\text{min}^{-1}\cdot\text{kg}^{-1}$, recorded in a competitive runner.

An interesting quantity is the proportion of the Andean VO_2max advantage that could be explained from allele frequency differences and the effect size. Population data going back 50 years show an increasing Andean advantage with altitude, reaching $\sim 5 \text{ ml}\cdot\text{min}^{-1}\cdot\text{kg}^{-1}$ at 4,400 meters (42). The mean effect size here (i.e., the difference between high vs. low VO_2max genotypes) across all SNPs was $\sim 2 \text{ ml}\cdot\text{min}^{-1}\cdot\text{kg}^{-1}$, with a larger effect size for SNP rs1769793 ($4 \text{ ml}\cdot\text{min}^{-1}\cdot\text{kg}^{-1}$, Fig. 3). From this, we calculate that SNP rs1769793 explains 54% of the expected $5 \text{ ml}\cdot\text{min}^{-1}\cdot\text{kg}^{-1}$ Quechua advantage i.e., $2.7 \text{ ml}\cdot\text{min}^{-1}\cdot\text{kg}^{-1}$. Physiologically, this is a small difference, but the expectation is that VO_2max is polygenic with many genes under selection pressure. Even so, relatively rare variants at *EGLN1* could underlie the observation that some altitude sojourners of non-native ancestry do well at altitude, while others struggle with altitude pathologies (43). For example, only two of 97 Syracuse participants harbored all high VO_2max alleles at all 5 SNP loci. These two individuals performed well in hypoxia, falling at the 90th and 96th percentiles of the Syracuse VO_2max distribution.

A strength of this study is that the top association signal (rs1769793) was replicated in an independent sample. In the replication sample, rs1769793 was significantly associated with VO_2max ($p=0.033$) and explained 10.6% of the variance (Fig. 1a and SI Appendix, Table S7). The frequency of the high VO_2max allele (T) at rs1769793 in the replication sample was high at 68% compared to the current sample at 55% (Figure 1b and c). The higher frequency of T in the replication sample could be the result of stochastic error, or more likely from higher rates of admixture in the current sample. That is, the replication sample was more rural, and consistent with the argument of genetic adaptation, higher frequencies of adaptive alleles are expected in rural areas where Spanish admixture is lower. Replication of the top association signal strongly suggests that findings are not spurious.

The G allele of rs479200 is one of two *EGLN1* SNP alleles linked to high-altitude adaptation in Asian Indians (44). In this population, the G allele frequency is 0.71 (similar to the Quechua

sample (0.80), SI Appendix, Fig. S3) compared to 0.36 in low-altitude Indian populations. This difference in allele frequency is comparable to that between Quechua and the Syracuse referents (~0.40, SI Appendix, Fig. S3) as well as 1KG Mexicans (0.35). Moreover, the non-adaptive allele of rs479200 is associated with high-altitude pulmonary edema in Indians. This raises the possibility of convergent evolution in altitude-adapted Andeans and Asian Indians.

Significant SNPs resided in non-coding regions of *EGLN1* or outside the *EGLN1* gene boundaries (Table 1, Fig. 2). We also note the following: 1) The most significant SNP, rs1769793, is a regulatory region variant. The high VO₂Max allele, T, modifies a transcription factor binding site. This SNP is linked to rs1769792, the *EGLN1* SNP with the highest LSBL value from (19). rs1769792 is also a regulatory variant affecting transcription factor binding. Furthermore, rs1769793 is associated with reticulocyte count and percentage within the UK Biobank (44). This lends additional support that this variant contributes to phenotypes involved in oxygen sensing and delivery; 2) The two coding region SNPs associated with [Hb] in Tibetans, D4E and C127S, were present, but not associated with VO₂max. No other missense SNPs or SNPs affecting splicing donor or acceptor sites were identified; 3) Two SNPs reported here, rs2064766 and rs2437150, reside upstream of *EGLN1* in the 3' UTR of *EXOC8* or in the coding region (P296L) of *SPRTN*, respectively. *EXOC8* is a component of the exocyst complex involved in targeting of secretory vesicles (45). *SPRTN* is a nuclear metalloprotease implicated in DNA repair, with human mutations associated with genomic instability (46); 4) *EGLN1* SNP rs479200 resides in a region characterized by H3K27Ac marks, DNase I hypersensitivity, and ChIP-seq transcription factor binding across multiple cell types. Of note, there are no miRNAs or snoRNAs within the *EGLN1* gene, and the closest lincRNAs are over 40 kb away from intron 1. Thus, it is unlikely that this SNP affects these classes of RNAs. On balance, it seems more likely that this SNP affects regulation of the *EGLN1* gene. Taken together, the functional variant(s) of the Andean *EGLN1* allele appears to be unlikely to affect the coding sequence or translation of the protein. Rather, we hypothesize that the functional variant is regulatory or intronic, the nature of which will require further investigation.

Summary and conclusions.

This study reveals an association between *EGLN1* SNP variants and VO₂max in hypoxia. For most SNPs, the adaptive alleles were found at higher frequency in Quechua consistent with directional selection on an unknown, linked causal variant. These results, along with previous statistical genetic evidence of natural selection on *EGLN1* support the hypothesis of genetic adaptation in Quechua via the selection of genetic variants conferring an advantage with respect to work/exercise performance at altitude. The strongest SNP association (rs1769793) was replicated and was strongly evident in genome-wide analysis. The non-coding location of all SNPs supports the hypothesis that the putative Andean *EGLN1* adaptation is regulatory.

ACKNOWLEDGEMENTS:

This study was supported by NSF BCS1132310 to TDB, AWB, and FLV. The authors would like to thank the many participants in Peru and the USA.

REFERENCES:

1. L. G. Moore, Fetal growth restriction and maternal oxygen transport during high altitude pregnancy. *High Alt Med Biol* **4**, 141-156 (2003).
2. E. R. Buskirk, J. Kollias, R. F. Akers, E. K. Prokop, E. P. Reategui, Maximal performance at altitude and on return from altitude in conditioned runners. *J Appl Physiol* **23**, 259-266 (1967).
3. K. Rademaker *et al.*, Paleoindian settlement of the high-altitude Peruvian Andes. *Science* **346**, 466-469 (2014).
4. A. W. Bigham *et al.*, Maternal PRKAA1 and EDNRA genotypes are associated with birth weight, and PRKAA1 with uterine artery diameter and metabolic homeostasis at high altitude. *Physiol Genomics* **46**, 687-697 (2014).
5. T. Brutsaert *et al.*, Spanish genetic admixture is associated with larger VO₂max decrement from sea level to 4,338 m in Peruvian Quechua. *J. Appl. Physiol.* **95**, 519-528 (2003).
6. M. D. Shriver *et al.*, Finding the genes underlying adaptation to hypoxia using genomic scans for genetic adaptation and admixture mapping. *Adv Exp Med Biol* **588**, 89-100 (2006).
7. P. T. Baker, "Work performance of highland natives" in *Man in the Andes: A multidisciplinary study of high-altitude Quechua natives*, P. T. Baker, M. A. Little, Eds. (Wovden, Hutchinson, and Ross, Inc., Stroudsburg, PA, 1976).
8. R. W. Elsner, A. Blostad, C. Forno, "Maximum oxygen consumption of Peruvian Indians native to high altitude" in *The physiological effects of high altitude*, W. H. Weihe, Ed. (Pergamon Press, New York, 1964), pp. 217-223.
9. A. R. Frisancho, C. Martinez, T. Velasquez, J. Sanchez, H. Montoye, Influence of developmental adaptation on aerobic capacity at high altitude. *J Appl Physiol* **34**, 176-180 (1973).
10. J. Kollias *et al.*, Work capacity of long-time residents and newcomers to altitude. *J Appl Physiol* **24**, 792-799 (1968).
11. R. B. Mazess, Exercise performance of Indian and white high altitude residents. *Hum Biol* **41**, 494-518 (1969).
12. R. B. Mazess, Exercise performance at high altitude in Peru. *Fed Proc* **28**, 1301-1306 (1969).
13. P. T. Baker, Human adaptation to high altitude. *Science* **163**, 1149-1156 (1969).
14. P. W. Hochachka *et al.*, Metabolic and work efficiencies during exercise in Andean natives. *J Appl Physiol* **70**, 1720-1730 (1991).
15. T. Velasquez, B. Reynafarje, Metabolic and physiological aspects of exercise at high altitude. II. Response of natives to different levels of workload breathing air and various oxygen mixtures. *Fed Proc* **25**, 1400-1404 (1966).
16. J. A. Vogel, L. H. Hartley, J. C. Cruz, Cardiac output during exercise in altitude natives at sea level and high altitude. *J Appl Physiol* **36**, 173-176. (1974).
17. A. B. Way, Exercise capacity of high altitude peruvian Quechua Indians migrant to low altitude. *Hum Biol* **48**, 175-191 (1976).
18. C. Monge, *Acclimatization in the Andes* (The Johns Hopkins Press, Baltimore, 1948).
19. A. Bigham *et al.*, Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet* **6** (2010).
20. M. Foll, O. E. Gaggiotti, J. T. Daub, A. Vatsiou, L. Excoffier, Widespread signals of convergent adaptation to high altitude in Asia and america. *Am J Hum Genet* **95**, 394-407 (2014).
21. A. W. Bigham, F. S. Lee, Human high-altitude adaptation: forward genetics meets the HIF pathway. *Genes Dev* **28**, 2189-2204 (2014).

22. T. S. Simonson *et al.*, Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**, 72-75 (2010).
23. C. M. Beall *et al.*, Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. *Proc Natl Acad Sci U S A* **107**, 11459-11464 (2010).
24. X. Yi *et al.*, Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75-78 (2010).
25. A. J. Majmundar, W. J. Wong, M. C. Simon, Hypoxia-inducible factors and the response to hypoxic stress. *Mol Cell* **40**, 294-309 (2010).
26. G. L. Semenza, Regulation of oxygen homeostasis by hypoxia-inducible factor 1. *Physiology (Bethesda)* **24**, 97-106 (2009).
27. F. S. Lee, M. J. Percy, The HIF pathway and erythrocytosis. *Annu Rev Pathol* **6**, 165-192 (2011).
28. M. J. Percy *et al.*, A family with erythrocytosis establishes a role for prolyl hydroxylase domain protein 2 in oxygen homeostasis. *Proc Natl Acad Sci U S A* **103**, 654-659 (2006).
29. P. R. Arsenault *et al.*, A knock-in mouse model of human PHD2 gene-associated erythrocytosis establishes a haploinsufficiency mechanism. *J Biol Chem* **288**, 33571-33584 (2013).
30. T. Bishop *et al.*, Carotid body hyperplasia and enhanced ventilatory responses to hypoxia in mice with heterozygous deficiency of PHD2. *J Physiol* **591**, 3565-3577 (2013).
31. A. W. Bigham *et al.*, Angiotensin-converting enzyme genotype and arterial oxygen saturation at high altitude in Peruvian Quechua. *High Alt Med Biol* **9**, 167-178 (2008).
32. T. D. Brutsaert *et al.*, Ancestry explains the blunted ventilatory response to sustained hypoxia and lower exercise ventilation of Quechua altitude natives. *Am J Physiol Regul Integr Comp Physiol* (2005).
33. A. L. Price *et al.*, Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).
34. S. Purcell, S. S. Cherny, P. C. Sham, Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149-150 (2003).
35. A. M. Nevill, R. Ramsbottom, C. Williams, Scaling physiological measurements for individuals of different body size. *Eur J Appl Physiol* **65**, 110-117 (1992).
36. F. R. Lorenzo *et al.*, A genetic mechanism for Tibetan high-altitude adaptation. *Nat Genet* **46**, 951-956 (2014).
37. N. Petousi *et al.*, Tibetans living at sea level have a hyporesponsive hypoxia-inducible factor system and blunted physiological responses to hypoxia. *J Appl Physiol* (1985) **116**, 893-904 (2014).
38. K. Xiang *et al.*, Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. *Mol Biol Evol* **30**, 1889-1898 (2013).
39. A. W. Bigham *et al.*, Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. *Hum Genomics* **4**, 79-90 (2009).
40. J. A. Timmons *et al.*, Using molecular classification to predict gains in maximal aerobic capacity following endurance exercise training in humans. *J Appl Physiol* (1985) **108**, 1487-1496 (2010).
41. P. D. Wagner *et al.*, Sea-level haemoglobin concentration is associated with greater exercise capacity in Tibetan males at 4200 m. *Exp Physiol* **100**, 1256-1262 (2015).
42. T. D. Brutsaert, Do high-altitude natives have enhanced exercise performance at altitude? *Appl Physiol Nutr Metab* **33**, 582-592 (2008).

43. H. E. Montgomery *et al.*, Human gene for physical performance. *Nature* **393**, 221-222 (1998).
44. G. McInnes *et al.*, Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. *Bioinformatics* 10.1093/bioinformatics/bty999 (2018).
45. B. Wu, W. Guo, The Exocyst at a Glance. *J Cell Sci* **128**, 2957-2964 (2015).
46. D. Lessel *et al.*, Mutations in SPRTN cause early onset hepatocellular carcinoma, genomic instability and progeroid features. *Nat Genet* **46**, 1239-1244 (2014).

TABLES:

Table 1. *EGLN1* SNP associations with VO₂max from GWAS

rsID	BP	<i>n</i>	Beta	SE	L95	U95	STAT	<i>P</i> value (uncorrected)	Rank out of 215,512
rs1769793	231601099	511	0.06	0.02	0.02	0.10	3.16	0.00	533
rs2491403	231511185	516	-0.04	0.02	-0.08	-0.01	-2.21	0.03	6,704
rs2064766	231468953	522	-0.04	0.02	-0.08	0.00	-2.20	0.03	6,823
rs2437150	231488524	519	-0.04	0.02	-0.08	0.00	-2.12	0.03	8,409
rs2749713	231537921	503	-0.04	0.02	-0.08	0.00	-1.95	0.05	12,399
rs479200	231543780	523	-0.03	0.02	-0.07	0.00	-1.78	0.08	17,740
rs480902	231531627	523	-0.02	0.02	-0.06	0.02	-1.07	0.29	63,526
rs12030600	231605379	523	-0.02	0.03	-0.07	0.04	-0.55	0.58	127,033

Regression coefficients (beta) $\pm$ the 95% confidence intervals are shown. Base pairs (BP) are provided for Human Genome Build 19 (HG19). STAT is the coefficient of the *t* statistic.

Table 2. *EGLN1* SNPs associated with VO₂max in hypoxia from ANCOVA

rsID	Position (HG38)	Alleles	MA*	Adaptive allele†	MAF Quechua	MAF MXL	<i>EGLN1</i> relationship	Gene	Function	SNP genotype main effect		Interaction SNP-by- study group	
										<i>P</i> value	R ²	<i>P</i> value	R ²
rs2064766	231333207	A/G	A	G	0.30	0.48	50KB down	<i>EXOC8</i>	UTR-3'	0.014	0	ns	ns
rs2437150	231352778	C/T	T	C	0.31	0.51	50KB down	<i>SPRTN</i>	missense	0.059	0	0.029	0.028
rs2491403	231375439	C/T	T	C	0.31	0.50	In gene	<i>EGLN1</i>	intronic	0.037	0	0.045	0.026
rs479200	231408034	A/G	A	G	0.37	0.65	In gene	<i>EGLN1</i>	intronic	0.056	0	0.019	0.030
rs1769793	231465353	C/T	C	T	0.45	0.65	50KB up	<i>EGLN1</i>	intergenic	0.0002	0	0.024	0.029

MA = minor allele; MAF = minor allele frequency; MXL = 1KG Mexican Americans from Los Angeles.

Table 3. ANCOVA model showing association of EGLN1 SNP (rs1769793) with VO₂max in hypoxia

Source	Type III sum of squares	df	Mean square	F	Sig.	R ²
Intercept	0.49	1	0.49	6.06	<0.00001	0.19
Age	0.31	1	0.31	3.82	0.014	0.01
Weight	10.07	1	10.07	124	<0.00001	0.21
Sex	36.87	1	36.87	457.8	<0.00001	0.48
Study subgroup	5.18	3	1.73	21.44	<0.00001	0.20
EGLN1 SNP rs1769793	1.37	2	0.68	8.47	0.000241	0.03
Sex*subgroup interaction	1.17	3	0.39	4.86	0.002	0.03
rs1769793*subgroup interaction	1.18	6	0.2	2.44	0.024	0.03
Error	39.3	488	0.08			
Total	2,273.55	511				

Model R² = 0.762.

FIGURE LEGENDS:

Fig. 1. Association of *EGLN1* rs1769793 with VO₂max in hypoxia. **(a)** Marginal mean values of VO₂max (ml·min⁻¹·kg⁻¹) for the current study and for the replication cohort (from ANCOVA models presented in Table 3 and SI Appendix Table S7). Data for the replication cohort were collected in 2001-2002 and have been published elsewhere (5, 42, 43). Error bars are standard error of the mean (SEM). **(b)** Genotype frequencies for rs1769793 from the current study (PQU and SYR) and the replication cohort (PQU-R). The high VO₂max genotype (TT) is given in orange, the heterozygote genotype (CT) in blue, and the low VO₂max genotype (CC) in grey. **(c)** Allele frequencies for the putative adaptive allele (T) in PQU, PQU-R, and SYR samples compared to allele frequency data from the 1000-genomes Phase 3. Quechua have the highest recorded allele frequency of T world-wide. Global is the global mean value T frequency. Comparison population abbreviations given in SI Appendix, Materials and Methods.

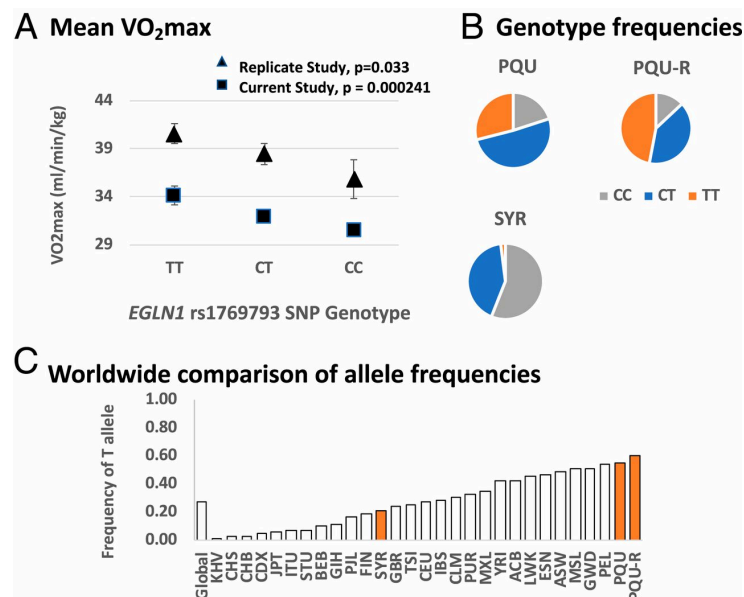


Fig. 2. Difference in marginal mean values of VO₂max (ml·min⁻¹·kg⁻¹) for the five significant SNP markers associated with VO₂max. The difference in adjusted mean values of VO₂max between the highest VO₂max and lowest VO₂max genotype categories for each SNP is plotted. The

genomic coordinate of each SNP along chromosome 1 is shown along the x-axis. Linkage disequilibrium is shown via R^2 values for each SNP by SNP comparison.

