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

Tom D. Brutsaert[a,1], Melisa Kiyamu[b], Gianpietro Elias Revollendo[a], Jenna L. Isherwood[c], Frank S. Lee[d], Maria Rivera-Ch[b], Fabiola Leon-Velarde[b], Sudipta Ghosh[e], and Abigail W. Bigham[c,2]

[a]Department of Exercise Science, Syracuse University, Syracuse, NY 13244; [b]Universidad Peruana Cayetano Heredia, Lima, Peru; [c]Department of Anthropology, University of Michigan, Ann Arbor, MI 48109-1107; [d]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104; and [e]Department of Anthropology, North-Eastern Hill University, Shillong, India

Edited by Cynthia M. Beall, Case Western Reserve University, Cleveland, OH, and approved October 14, 2019 (received for review April 11, 2019)

Highland native Andeans have resided at altitude for millennia. They display high aerobic capacity (VO$_2$max) at altitude, which may be a reflection of genetic adaptation to hypoxia. Previous genomewide (GW) scans for natural selection have nominated *Egl-9 homolog 1* gene (*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 GW association and analysis of covariance performed on a total sample of 429 Peruvian Quechua and 94 US lowland referents, we identified 5 *EGLN1* SNPs associated with higher VO$_2$max (L·min^{-1} and mL·min^{-1}·kg^{-1}) in hypoxia (rs1769793, rs2064766, rs2437150, rs2491403, rs479200). For 4 of these SNPs, Quechua had the highest frequency of the advantageous (high VO$_2$max) allele compared with 25 diverse lowland comparison populations from the 1000 Genomes Project. Genotype effects were substantial, with high versus low VO$_2$max genotype categories differing by ∼11% (e.g., for rs1769793 SNP genotype TT = 34.2 mL·min^{-1}·kg^{-1} vs. CC = 30.5 mL·min^{-1}·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.

hypoxia | selection | Peruvian Quechua | evolution | aerobic capacity

Low O$_2$ availability at altitude is an environmental stressor that has a negative effect 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 (VO$_2$max, mL·min^{-1}·kg^{-1}) in Andean natives tested at high altitude (5, 7–17). Indeed, an adaptive response reflected in enhanced VO$_2$max is a long-standing idea that can be traced back to early Spanish accounts of impressive physical tolerance to altitude among the Inca in the 1500s (18). As a phenotype, VO$_2$max 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 O$_2$ delivery and/or cellular O$_2$ use.

For compelling reasons, we focused a priori attention on the *Egl-9 homolog 1* gene (*EGLN1*, also known as *PHD2*). First, previous genomewide (GW) 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 dataset of Peruvian Quechua to test for genetic association between *EGLN1* SNP variants and the VO$_2$max measured in hypoxia. Our sample comprised 429 Quechua as well as 94 non-Hispanic white lowlanders from Syracuse, New York. We performed 2 complementary analyses on the full sample, including a conventional GW association study (GWAS) on the VO$_2$max phenotype and a targeted analysis of covariance (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 4 subgroups: Quechua-high-altitude

Significance

Andean highland native populations, such as the Quechua of Peru, have enhanced exercise capacity at altitude and may be genetically adapted to altitude. We identified 5 genetic markers near the *Egl-9 homolog 1* gene (*EGLN1*) gene that were associated with higher aerobic capacity (VO$_2$max) 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$_2$max and were found at higher frequency in Quechua compared with lowland populations. These results add further context to previous studies that have provided evidence of natural selection at the *EGLN1* locus in Andeans.

Author contributions: T.D.B., F.L.-V., and A.W.B. designed research; T.D.B., M.K., G.E.R., J.L.I., M.R.-C., F.L.-V., S.G., and A.W.B. performed research; T.D.B., J.L.I., and A.W.B. contributed new reagents/analytic tools; T.D.B., J.L.I., F.S.L., and A.W.B. analyzed data; and T.D.B. and A.W.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: Genotyping data are available through the Dryad digital repository (https://doi.org/10.5068/D1XH3R).

[1]To whom correspondence may be addressed. Email: tdbrutsa@syr.edu.

[2]Present address: Department of Anthropology, University of California, Los Angeles, CA 90095.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1906171116/-/DCSupplemental.

First published November 11, 2019.

residents (n = 195) from Cerro de Pasco, Peru, at 4,338 m above sea level; Quechua-migrants (n = 111), born at altitude but migrated permanently to sea level from Lima, Peru; Quechua-born at sea level from Lima (n = 123); and non-Hispanic whites (n = 94) from Syracuse, NY, at ∼140 m above sea level. Roughly equal numbers of men and women were recruited between the ages of 18 and 35 y. 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, 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 to 2002 from 67 male and female Quechua by the same investigators, using the same exercise testing equipment and protocol (5, 31, 32).

For Quechua-high-altitude residents, exercise testing was conducted in Cerro de Pasco, Peru. For the Quechua-migrants, Quechua-born at sea level from Lima, and Syracuse groups, exercise testing was conducted under simulated altitude conditions by lowering the fractional concentration of O_2 (F_IO_2) to ∼0.126 at sea level. VO_2max was measured using a graded testing protocol and a metabolic cart on a cycle ergometer. Only participants achieving a true VO_2max (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 ∼610,000 markers. The Biobanking Array contains 29 markers in and around (50 kb upstream and downstream) *EGLN1*. Of these 29 *EGLN1* markers, 6 met the criteria for association testing. In addition, we manually genotyped 2 *EGLN1* SNPs (rs479200, rs480902) that exhibited substantial differences in minor allele frequency (MAF) compared with Mexican control individuals from the 1000 Genomes Project (1KG) phase 3 (Table 1). This resulted in a final selection of 8 *EGLN1* SNPs for genetic analysis.

We tested *EGLN1* SNP associations with VO_2max, using 2 complementary approaches: GWAS and a priori 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 1.90 (https://www.cog-genomics.org/plink2/). GW 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 principal 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 5 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 (34). 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 subgroups in body size and composition, [Hb], and VO_2max, but these were expected, given differences in ethnicity, place of birth, altitude of residence, and acclimatization state.

Full GWAS results are reported in Dataset S1. After correction for GW significance, no SNPs were significantly associated with VO_2max. Power analysis revealed that sample size was underpowered to detect an association with VO_2max at 80% power (35). Our genomic inflation factor for the combined Quechua and Syracuse dataset was 6.266 before controlling for covariates. When adjusting for all 5 covariates, including PC1, our genomic inflation factor drops to 1.043 for the combined Quechua and Syracuse dataset. When genomic inflation was measured among Quechua or Syracuse participants independently, the genomic inflation factor 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 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 VO_2max via GWAS (uncorrected P value < 0.05) and ranked within the top 4% of SNPs tested.

From ANCOVA, 5 of 8 *EGLN*1 SNPs were associated with VO_2max and showed similar association patterns, either as a SNP main-effect or as a SNP-by-study subgroup interaction (Table 2). The remaining 3 SNPs were not significant. Interactions resulted from SNP genotype effects within study subgroups 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 VO_2max 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 4 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 VO_2max ($mL \cdot min^{-1} \cdot kg^{-1}$) by rs1769793 SNP genotype are shown in Fig. 1*A*. The rs1769793 genotype differences in VO_2max were physiologically significant, with TT = 34.16 ± 0.98 $mL \cdot min^{-1} \cdot kg^{-1}$, CT = 31.98 ± 0.40 $mL \cdot min^{-1} \cdot kg^{-1}$, and CC = 30.50 ± 0.53 $mL \cdot min^{-1} \cdot kg^{-1}$. The rs1769793 genotype frequencies were different between Quechua and Syracuse populations, with 29% of Quechua in the high VO_2max genotype category (TT) compared with only 2% of Syracuse (Fig. 1*B*). For broader global context, we compared Quechua allele frequencies for the adaptive high VO_2max allele (T) for rs1769793 with population mean values available from 1KG phase 3 (Fig. 1*C*; https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/). The Quechua sample has the highest frequency of T = 0.55 compared with all other populations included in the 1KG data (Fig. 1*C*).

Table 3 shows the specific ANCOVA model for rs1769793. Body size, age, and sex accounted for the majority of the variance in VO_2max. After covariate control, rs1769793 SNP genotype

Table 1. ***EGLN1* SNP associations with VO_2max from GWAS**

rsID	BP	n	Beta	SE	L95	U95	STAT	P 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) ± 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 VO_2max in hypoxia from ANCOVA

										SNP genotype main effect		Interaction SNP-by-study group	
rsID	Position (HG38)	Alleles	MA*	Adaptive allele†	MAF Quechua	MAF MXL	*EGLN1* relationship	Gene	Function	*P* value	R^2	*P* value	R^2
rs2064766	231333207	A/G	A	G	0.30	0.48	50KB down	*EXOC8*	UTR-3′	0.014	0.02	ns	ns
rs2437150	231352778	C/T	T	C	0.31	0.51	50KB down	*SPRTN*	missense	0.059	0.01	0.029	0.028
rs2491403	231375439	C/T	T	C	0.31	0.50	In gene	*EGLN1*	intronic	0.037	0.01	0.045	0.026
rs479200	231408034	A/G	A	G	0.37	0.65	In gene	*EGLN1*	intronic	0.056	0.01	0.019	0.030
rs1769793	231465353	C/T	C	T	0.45	0.65	50KB up	*EGLN1*	intergenic	0.0002	0.03	0.024	0.029

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

*Defined in Peruvian Quechua.

†Adaptive allele is the allele associated with higher VO_2max in hypoxia.

(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 (36), but results were identical, whichever method of body size control was applied. The same general association pattern for rs1769793 was evident for the 4 additional significant SNPs, rs2064766, rs2437150, rs2491403, and rs479200 (*SI Appendix*, Figs. S2 and S3 and Tables S2–S6). That is, Quechua were overrepresented in the high VO_2max SNP genotype categories (*SI Appendix*, Fig. S2; odds ratio [OR] Fisher exact P values ≤ 0.001; 4 SNPs tested), and Quechua showed high frequency of the putative adaptive allele at each locus compared with 1KG phase 3 data (*SI Appendix*, Fig. S3 *A–D*). Indeed, the Quechua sample showed the highest frequency worldwide of the adaptive allele across 3 of these loci (rs2064766, rs2437150, and rs479200).

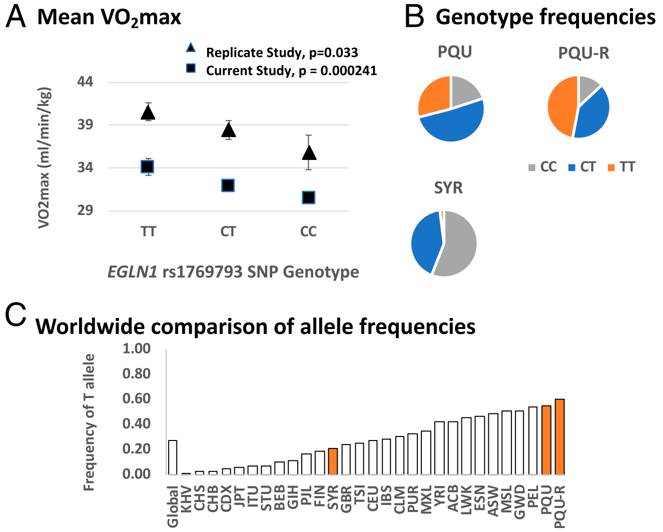

Fig. 1. Association of *EGLN1* rs1769793 with VO_2max in hypoxia. (*A*) Marginal mean values of VO_2max ($mL·min^{-1}·kg^{-1}$) for the current study and for the replication cohort (from ANCOVA models presented in Table 3 and *SI Appendix*, Table S7). Data for the replication cohort were collected in 2001 to 2002 and have been published elsewhere (5, 43, 44). Error bars are SEM. (*B*) Genotype frequencies for rs1769793 from the current study (PQU and SYR) and the replication cohort (PQU-R). The high VO_2max genotype (TT) is given in orange, the heterozygote genotype (CT) in blue, and the low VO_2max genotype (CC) in gray. (*C*) Allele frequencies for the putative adaptive allele (T) in PQU, PQU-R, and SYR samples compared with allele frequency data from the 1KG phase 3. Quechua have the highest recorded allele frequency of T worldwide. Global is the global mean value T frequency. Comparison population abbreviations given in *SI Appendix*, *Materials and Methods*.

We also tested models that introduced (pairwise) 2 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 4 SNPs (R^2 range = 0.31 to 0.41), while the other 4 SNPs showed high to moderate LD with each other ($R^2 > 0.60$; Fig. 2). Taken together, these results suggest that the 5 SNPs mark a single causal genetic locus.

We identified individuals harboring all the alleles associated with higher VO_2max for all 5 SNPs at rs1769793, rs2064766, rs2437150, rs2491403, and rs479200 (T, G, C, C, and G, respectively), and compared them with 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 6.0% (27 of 454) of Quechua had none of the high VO_2max alleles compared with 33.0% (32 of 97) of Syracuse participants. Mean values of VO_2max were 33.97 $mL·min^{-1}·kg^{-1}$ versus 30.42 $mL·min^{-1}·kg^{-1}$ 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 5 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 ± 1.04 $mL·min^{-1}·kg^{-1}$, TC = 38.46 ± 1.10 $mL·min^{-1}·kg^{-1}$, and CC 35.85 ± 2.01 $mL·min^{-1}·kg^{-1}$ (Fig. 1*A*). 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 1KG Peruvians from Lima sample (Fig. 1*C*).

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 2 coding region SNPs, D4E (rs186996510) and C127S (rs12097901), that are enriched in Tibetans and are in strong LD (37–39). In fact, the D4E/C127S haplotype is enriched more than 80-fold in Tibetans compared with in 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 5 *EGLN1* exons and ∼1 kb upstream and downstream of the first and last exons in 12 individuals from the 2001 to 2002 cohort (5, 31, 32). The D4E and C127S missense SNPs were present at frequencies of 12.5%, but increased only by

Table 3. **ANCOVA model showing association of EGLN1 SNP (rs1769793) with VO_2max 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	0.191
Age	0.31	1	0.31	3.82	0.014	0.009
Weight	10.07	1	10.07	124.00	<0.00001	0.208
Sex	36.87	1	36.87	457.77	<0.00001	0.483
Study subgroup	5.18	3	1.73	21.44	<0.00001	0.204
EGLN1 SNP rs1769793	1.37	2	0.68	8.47	0.000241	0.034
Sex*subgroup interaction	1.17	3	0.39	4.86	0.002	0.029
rs1769793*subgroup interaction	1.18	6	0.20	2.44	0.024	0.029
Error	39.30	488	0.08			
Total	2,273.55	511				

Model R^2 = 0.762.

11.7% and 0.8%, respectively, compared with Mexican control individuals from 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 429 Peruvian Quechua, using PCR and restriction enzyme digestion (*SI Appendix*, Table S9). The D4E SNP was monomorphic, whereas the MAF of C127S was 0.08% (MA = C). ANCOVA revealed no association for C127S (rs12097901) with VO_2max. Our sequencing efforts revealed no other missense SNPs or SNPs affecting splicing donor or acceptor sites. An additional 5 variable sites were identified in either the 5′ or 3′ UTR, but in no case was the difference in MAF > 0.1 compared with Mexican control individuals. Taken together, it appears unlikely that the potentially functional variant of the Andean *EGLN1* affects the coding sequence, splicing, or translation.

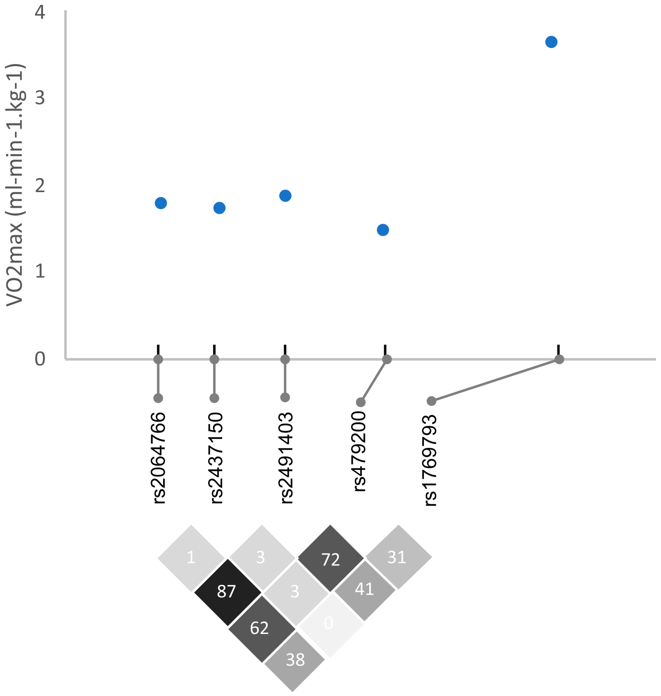

Fig. 2. Difference in marginal mean values of VO_2max (mL·min^{-1}·kg^{-1}) for the 5 significant SNP markers associated with VO_2max. The difference in adjusted mean values of VO_2max between the highest VO_2max and lowest VO_2max genotype categories for each SNP is plotted. The genomic coordinate of 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.

Discussion

This study reveals a genetic association between an unknown *EGLN1* variant/haplotype and VO_2max measured in hypoxia. Results contextualize previously published evidence of natural selection at *EGLN1* in Andeans (19–21, 40). Indeed, the genotype–phenotype association and high frequency of several SNP alleles associated with higher VO_2max 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 4 *EGLN1* SNPs, including SNP rs1769793, which ranked in the top 0.25% of all SNPs tested at $P =$ 0.002 (uncorrected). The nonsignificant *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, 5 SNPs showed strikingly similar association patterns with VO_2max in hypoxia. Four of these SNPs were the same as those identified by GWAS. These 5 SNPs were associated with VO_2max 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% to 11% of the variance in VO_2max, and genotype effects were large, with differences between high and low VO_2max genotypes of ∼11% to 13% (Fig. 1*A*). For perspective, moderate aerobic training produces mean gains of ∼14% (41). More important, Quechua were overrepresented in all high VO_2max SNP-genotype categories compared with Syracuse lowland natives, consistent with the hypothesis of directional selection on favorable variants. For example, for rs1769793, 29% of Quechua were TT (high VO_2max) compared with only 2% of Syracuse participants (Fig. 1*C*). Considered another way, more than half of the Syracuse sample (56%) were CC (low VO_2max) compared with only 20% of Quechua. In terms of allele frequencies, at 4 of the 5 SNPs, Quechua had the highest frequency of the high VO_2max allele compared with samples from 1KG (Fig. 1*C*).

The argument for genetic adaptation depends on several criteria: identification of an adaptive phenotype, association of the phenotype with a gene or genes, allele frequencies that are consistent with the mode of selection hypothesized, statistical genetic evidence of past natural selection on the genomic region harboring the gene or genes, and 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 to 3, whereas previous work by our group and others supports criteria 4 for

EGLN1 (19–21, 40). 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 1KG, our top associated SNP with VO_2max, rs1769793, is in high LD (R^2 = 0.977) with the highest-ranking SNP, rs1769792, in the locus-specific branch length analysis from ref. 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_2max was likely a target of past selection in the Andes. However, it is also possible that high VO_2max 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, criterion 5 has not been examined in the literature.

The current study is 1 of only 2 association studies on Andeans (4). Thus, it is of interest to compare our findings with those of association studies of Tibetans. Similar to 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 vs. decreased blood oxygen content). Interestingly, 1 study shows higher VO_2max in Tibetans with lower [Hb] (42) 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 a 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 with 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, 1 Quechua man had a value of 60 $mL{\cdot}min^{-1}{\cdot}kg^{-1}$ at 4,338 m. This is much higher than the highest Syracuse male value of 51.7 $mL{\cdot}min^{-1}{\cdot}kg^{-1}$, 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 y show an increasing Andean advantage with altitude, reaching ~5 $mL{\cdot}min^{-1}{\cdot}kg^{-1}$ at 4,400 m (43). The mean effect size here (i.e., the difference between high and low VO_2max genotypes) across all SNPs was ~2 $mL{\cdot}min^{-1}{\cdot}kg^{-1}$, with a larger effect size for SNP rs1769793 (4 $mL{\cdot}min^{-1}{\cdot}kg^{-1}$; Fig. 2). From this, we calculate that SNP rs1769793 explains 54% of the expected 5 $mL{\cdot}min^{-1}{\cdot}kg^{-1}$ Quechua advantage (i.e., 2.7 $mL{\cdot}min^{-1}{\cdot}kg^{-1}$). 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 nonnative ancestry do well at altitude, while others struggle with altitude pathologies (44). For example, only 2 of 97 Syracuse participants harbored all high VO_2max alleles at all 5 SNP loci. These 2 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. 1*A* 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 with the current sample, at 55% (Fig. 1 *B* 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 1 of 2 *EGLN1* SNP alleles linked to high-altitude adaptation in Asian Indians (45). In this population, the G allele frequency is 0.71 (similar to the Quechua sample (0.80), *SI Appendix*, Fig. S3) compared with 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 nonadaptive 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 noncoding regions of *EGLN1* or outside the *EGLN1* gene boundaries (Table 1 and Fig. 2). We also note the following: first, the most significant SNP, rs1769793, is a regulatory region variant. The high VO_2max allele, T, modifies a transcription factor binding site. This SNP is linked to rs1769792, the *EGLN1* SNP with the highest locus-specific branch length analysis value from ref. 19. rs1769792 is also a regulatory variant affecting transcription factor binding. Furthermore, rs1769793 is associated with reticulocyte count and percentage within the UK Biobank (46). This lends additional support that this variant contributes to phenotypes involved in oxygen sensing and delivery. Second, the 2 coding region SNPs associated with [Hb] in Tibetans, D4E and C127S, were present, but not associated with VO_2max. No other missense SNPs or SNPs affecting splicing donor or acceptor sites were identified. Third, 2 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 (47). *SPRTN* is a nuclear metalloprotease implicated in DNA repair, with human mutations associated with genomic instability (48). Fourth, *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 more than 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 variants of the Andean *EGLN1* allele appear 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 VO_2max 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 GWAS. The noncoding location of all SNPs supports the hypothesis that the putative Andean *EGLN1* adaptation is regulatory.

ACKNOWLEDGMENTS. This study was supported by NSF BCS1132310 to T.D.B., A.W.B., and F.L.-V. The authors would like to thank the many participants in Peru and the United States.

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. D. Brutsaert *et al*., Spanish genetic admixture is associated with larger V(O2) max decrement from sea level to 4338 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. (Wowden, Hutchinson, and Ross, IInc., 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. Velásquez, 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**, e1001116 (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.* **289**, R225–R234 (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. T. D. Brutsaert *et al*., Association of EGLN1 gene with high aerobic capacity of Peruvian Quechua at high altitude. Dryad Digital Repository. https://doi.org/10.5068/D1XH3R. Deposited 24 October 2019.
35. 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).
36. A. M. Nevill, R. Ramsbottom, C. Williams, Scaling physiological measurements for individuals of different body size. *Eur. J. Appl. Physiol. Occup. Physiol.* **65**, 110–117 (1992).
37. F. R. Lorenzo et al., A genetic mechanism for Tibetan high-altitude adaptation. *Nat. Genet.* **46**, 951–956 (2014).
38. 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.* **116**, 893–904 (2014).
39. 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).
40. A. W. Bigham *et al*., Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. *Hum. Genomics* **4**, 79–90 (2009).
41. J. A. Timmons *et al*., Using molecular classification to predict gains in maximal aerobic capacity following endurance exercise training in humans. *J. Appl. Physiol.* **108**, 1487–1496 (2010).
42. 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).
43. T. D. Brutsaert, Do high-altitude natives have enhanced exercise performance at altitude? *Appl. Physiol. Nutr. Metab.* **33**, 582–592 (2008).
44. H. E. Montgomery *et al*., Human gene for physical performance. *Nature* **393**, 221–222 (1998).
45. S. Aggarwal et al., *EGLN1* involvement in high-altitude adaptation revealed through genetic analysis of extreme constitution types defined in Ayurveda. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 18961–18966 (2010).
46. G. McInnes *et al*., Global Biobank Engine: Enabling genotype-phenotype browsing for biobank summary statistics. *Bioinformatics* **35**, 2495–2497 (2019).
47. B. Wu, W. Guo, The exocyst at a glance. *J. Cell Sci.* **128**, 2957–2964 (2015).
48. D. Lessel *et al*., Mutations in SPRTN cause early onset hepatocellular carcinoma, genomic instability and progeroid features. *Nat. Genet.* **46**, 1239–1244 (2014).