Identification of a Tibetan-Specific Mutation in the Hypoxic Gene *EGLN1* and Its Contribution to High-Altitude Adaptation

Kun Xiang,[†,1,2] Ouzhuluobu,[†,3] Yi Peng,[†,1] Zhaohui Yang,[1,2] Xiaoming Zhang,[1,2] Chaoying Cui,[3] Hui Zhang,[1] Ming Li,[1,2] Yanfeng Zhang,[1] Bianba,[3] Gonggalanzi,[3] Basang,[4] Ciwangsangbu,[4] Tianyi Wu,[5] Hua Chen,[6] Hong Shi,[1] Xuebin Qi,[*,1] and Bing Su[*,1]

[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

[2]University of Chinese Academy of Sciences, Beijing, China

[3]High Altitude Medical Research Center, School of Medicine, Tibetan University, Lhasa, China

[4]People's Hospital of Dangxiong County, Dangxiong, China

[5]National Key Laboratory of High Altitude Medicine, High Altitude Medical Research Institute, Xining, China

[6]Department of Epidemiology and Biostatistics, Harvard School of Public Health

†These authors contributed equally to this study.

***Corresponding authors:** E-mail: sub@mail.kiz.ac.cn; qixuebin@mail.kiz.ac.cn.

Associate editor: Joshua Akey

The 46 *EGLN1* gene sequences reported in this article have been deposited in the GenBank (accession nos. KC554019–KC554064).

Abstract

Tibetans are well adapted to high-altitude hypoxic conditions, and in recent genome-wide scans, many candidate genes have been reported involved in the physiological response to hypoxic conditions. However, the limited sequence variations analyzed in previous studies would not be sufficient to identify causal mutations. Here we conducted resequencing of the entire genomic region (59.4 kb) of the hypoxic gene *EGLN1* (one of the top candidates from the genome-wide scans) in Tibetans and identified 185 sequence variations, including 13 novel variations (12 substitutions and 1 insertion or deletion). There is a nonsynonymous mutation (rs186996510, D4E) showing surprisingly deep divergence between Tibetans and lowlander populations (F_{ST} = 0.709 between Tibetans and Han Chinese). It is highly prevalent in Tibetans (70.9% on average) but extremely rare in Han Chinese, Japanese, Europeans, and Africans (0.56–2.27%), suggesting that it might be the causal mutation of *EGLN1* contributing to high-altitude hypoxic adaptation. Neutrality test confirmed the signal of Darwinian positive selection on *EGLN1* in Tibetans. Haplotype network analysis revealed a Tibetan-specific haplotype, which is absent in other world populations. The estimated selective intensity (0.029 for the C allele of rs186996510) puts *EGLN1* among the known genes that have undergone the strongest selection in human populations, and the onset of selection was estimated to have started at the early Neolithic (~8,400 years ago). Finally, we detected a significant association between rs186996510 and hemoglobin levels in Tibetans, suggesting that *EGLN1* contributes to the adaptively low hemoglobin level of Tibetans compared with acclimatized lowlanders at high altitude.

***Key words:* EGLN1, positive selection, hypoxic adaptation, Tibetans.**

Introduction

Tibetans are among the several known populations in the world who have been living at high altitude (>3,000 m) for a long period of time and have developed physiological adaptations to high-altitude hypoxia (Wu and Kayser 2006; Beall 2007). The adaptive features of Tibetans include elevated resting ventilation, low hypoxic pulmonary vasoconstrictor response, and relatively higher level of blood oxygen saturation and relatively lower hemoglobin levels compared with acclimatized lowlanders (Wu and Kayser 2006; Beall 2007). All these features were acquired through long-lasting natural selection since the ancestors of Tibetans occupied the Himalayan region as early as 30,000 years ago (Shi et al. 2008). In addition, we expect the selectively advantageous mutations underlying the improved oxygen delivery under hypoxic conditions to be common in contemporary Tibetan populations.

Recently, multiple genome-wide scans of genetic variations in Tibetans have identified more than a dozen candidate genes in the hypoxia-inducible factor (HIF) pathway that showed unusually large allelic differences between highlander Tibetans and lowlander Han Chinese (Beall et al. 2010; Bigham et al. 2010; Simonson et al. 2010; Yi et al. 2010; Peng et al. 2011; Wang et al. 2011; Xu et al. 2011), among which *EPAS1* (endothelial PAS domain protein 1; also known as HIF2α) and *EGLN1* (egl nine homolog 1; also known as HIF prolylhydroxylase 2, PHD2) are two key genes functioning at the upstream of the HIF pathway and showed consistent selective signals across multiple studies (Peng et al. 2011; Xu et al. 2011). Resequencing of *EPAS1* detected a Tibetan-specific haplotype

Article

harboring high frequencies of many linked sequence variations in Tibetans, which are much less prevalent in Han Chinese and other world populations, suggesting *EPAS1* is one of the critical genes contributing to genetic adaptation to hypoxia (Peng et al. 2011). In contrast, although the selective signal of *EGLN1* was also consistent among some of the studies, the tested sequence variations of *EGLN1* were rather limited. There were only 22 single-nucleotide polymorphisms (SNPs) analyzed by the array-based assays in approximately 60-kb *EGLN1* region (Simonson et al. 2010) and sequencing gaps (high GC content regions in particular) existed in the *EGLN1* data by exome sequencing (Yi et al. 2010). Additionally, the reported selective signal may not necessarily reflect selection on *EGLN1*. For example, DISC-1, a gene not functioning in the hypoxic pathway, is also located in the same genomic region showing signal of selection (Peng et al. 2011). Hence, the incomplete data of *EGLN1* calls for a systematic analysis of a detailed sequence variation pattern to test its contribution to high-altitude hypoxic adaptation.

EGLN1 is located on human chromosome 1 (1q42.2). It spans about 60 kb and contains five exons (Dupuy et al. 2000). In normoxic conditions, *EGLN1* post-translationally regulates the level of EPAS1 and HIF1α by hydroxylating them on specific proline residues in an oxygen-dependent way (Jaakkola et al. 2001). This enables recognition of EPAS1 and HIF1α by the VHL ubiquitin ligase complex and subsequent degradation of them by the proteasome. In hypoxic conditions, the hydroxylation is significantly decreased, and EPAS1 and HIF1α are stabilized (Jaakkola et al. 2001). Because *EPAS1* has been shown carrying adaptive sequence changes in Tibetans for hypoxic adaptation (Peng et al. 2011), as a negative regulator of *EPAS1*, whether *EGLN1* also contributes to hypoxic adaptation is yet to be tested.

To dissect the molecular mechanism of adaptation to high-altitude hypoxia in Tibetans, we conducted resequencing of the complete genomic region of *EGLN1* (59.4 kb) in Tibetans and discovered one nonsynonymous mutation highly prevalent in Tibetans (64.3–75.8%) but extremely rare (<2.5%) in lowlander reference populations including Han Chinese. Network analysis based on *EGLN1* sequence data revealed a Tibetan-specific haplotype, which is absent in other world populations, suggesting that *EGLN1* has been under Darwinian positive selection and may contribute to high-altitude hypoxic adaptation in Tibetans.

Results

Complete Sequence of *EGLN1* and Discovery of a Tibetan-Specific Mutation

To reveal the detailed pattern of sequence variations of *EGLN1* in Tibetan populations, we first conducted resequencing of the entire genomic region of *EGLN1*, spanning about 59.4 kb, covering a 55.5-kb exon–intron region as well as a 3.4-kb 5′- and a 0.5-kb 3′-regions (fig. 1). Totally, we identified 166 SNPs and 19 insertions or deletions (in-dels), among which 12 SNPs and 1 insertion are novel variations (supplementary table S1, Supplementary Material online). All in-dels occur in the noncoding regions, therefore, do not cause protein sequence changes. Among the 166 SNPs, there are two nonsynonymous SNPs located within exon-1, a G to C mutation (rs186996510) leading to an amino acid change from aspartic acid to glutamic acid (D4E) and another G to C mutation (rs12097901) causing a change from cysteine to serine (S127C) (supplementary table S1, Supplementary Material online). The linkage disequilibrium (LD) map of *EGLN1* in Tibetans is shown in figure 1.

To identify SNPs showing unusually large allelic divergences between Tibetans and lowlander reference populations, therefore potentially contributing to high-altitude adaptation, we calculated pairwise F_{ST} between Tibetans and four reference populations including Han Chinese (CHB), Japanese (JPT), Europeans (CEU), and Africans (YRI) (data from the 1000 Human Genome Project) (supplementary table S2, Supplementary Material online). As shown in figure 2, when comparing Tibetans and Han Chinese, among the 150 SNPs and 13 in-dels, the most diverged region is located in exon-1 and the upstream of intron-1. There are 114 SNPs (76%) and 10 (76.9%) in-dels showing F_{ST} values larger than the average of Chromosome-1 (where *EGLN1* is located) (F_{ST} = 0.0102, calculated using the genome-wide array data; Peng et al. 2011), an indication of elevated genetic divergence of *EGLN1* between Tibetans and Han Chinese. Surprisingly, one of the two nonsynonymous mutations, rs186996510 (D4E) has a striking divergence between Tibetans and non-Tibetans, and it is highly prevalent in Tibetans (63.27%) but extremely rare in Han Chinese (1.03%), Japanese (0.56%), Europeans (0.59%), and Africans (2.27%). The F_{ST} value between Tibetans and Han Chinese is 0.709, about 70 times larger than the average of Chromosome-1. This nonsynonymous SNP was not identified in previous studies using the panel of Affimatrix Genome-wide Human SNP Array 6.0 (Beall et al. 2010; Bigham et al. 2010; Simonson et al. 2010; Peng et al. 2011; Wang et al. 2011; Xu et al. 2011), and neither was reported by previous exome sequencing (Yi et al. 2010) because it is located in a high GC content (~70%) region that is hard to conduct sequencing. This is so far the greatest allelic divergence observed between Tibetans and non-Tibetans, a strong implication of its potential role in high-altitude adaptation.

Neutrality Test and Haplotype Analysis

Although previous data using genome-wide SNP arrays has observed a signal of positive selection on the genomic region containing *EGLN1* (Simonson et al. 2010; Peng et al. 2011; Xu et al. 2011), it was inconclusive due to the limited SNPs used and the broad genomic region analyzed. With the complete sequence data of *EGLN1* in Tibetans, we performed a neutrality test using the method by Fay and Wu (2000). The result showed that there was an excess of major alleles in Tibetans, suggesting a significant deviation from neutral expectation and a signal of Darwinian positive selection on *EGLN1* ($H = -4.02$; $P = 0.025$), consistent with previous data (Peng et al. 2011; Xu et al. 2011).

Network analysis using *EGLN1* haplotypes derived from the 150 SNPs (shared between Tibetans and non-Tibetans) also

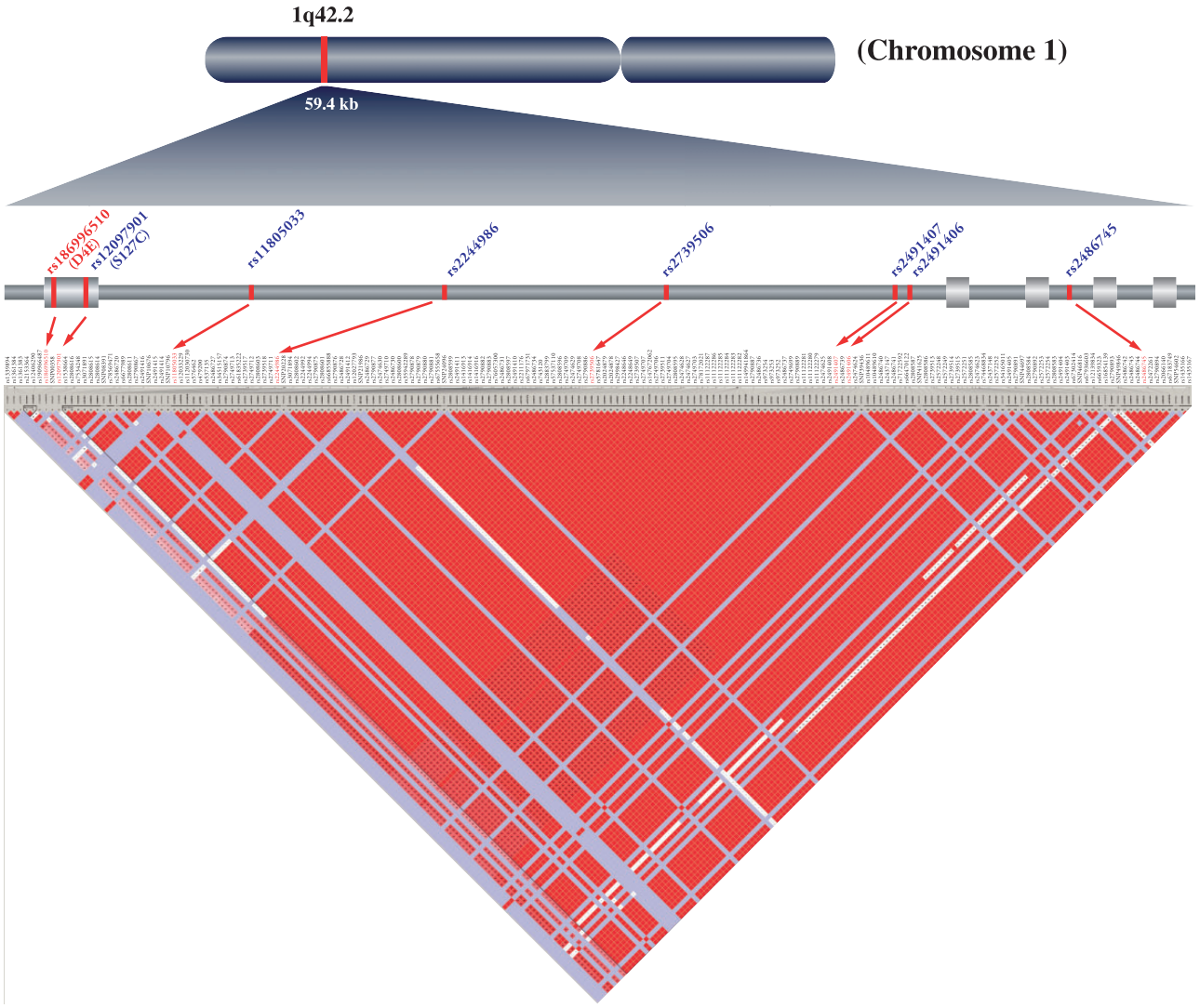

Fig. 1. The schematic map of *EGLN1* gene structure and LD. The gray boxes are the exons. The seven genotyped SNPs and rs12097901 are labeled. The LD map of *EGLN1* in Tibetans was constructed using all the 166 SNPs, and the 19 in-dels were excluded. TIB, Tibetans; CHB, Han Chinese; JPT, Japanese; CEU, Europeans; YRI, Africans.

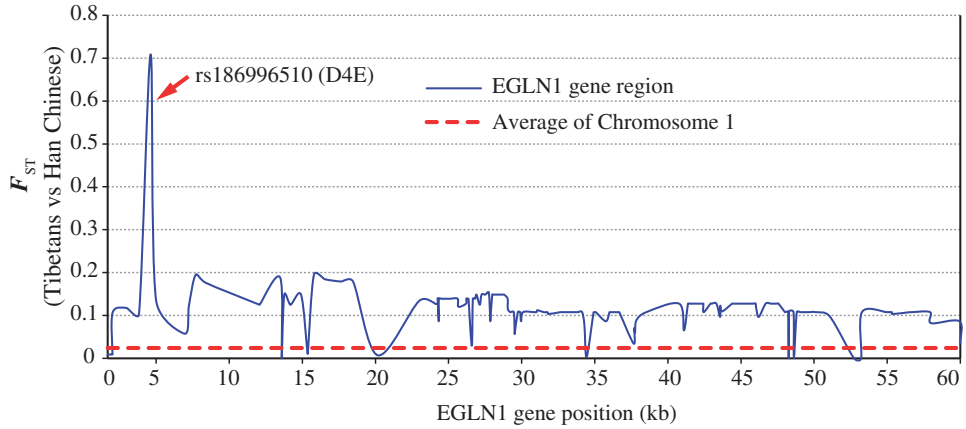

Fig. 2. Distribution of allelic divergence (measured by F_{ST}) between Tibetans and Han Chinese across the entire genomic region of *EGLN1*. The deeply diverged SNP rs186996510 is indicated.

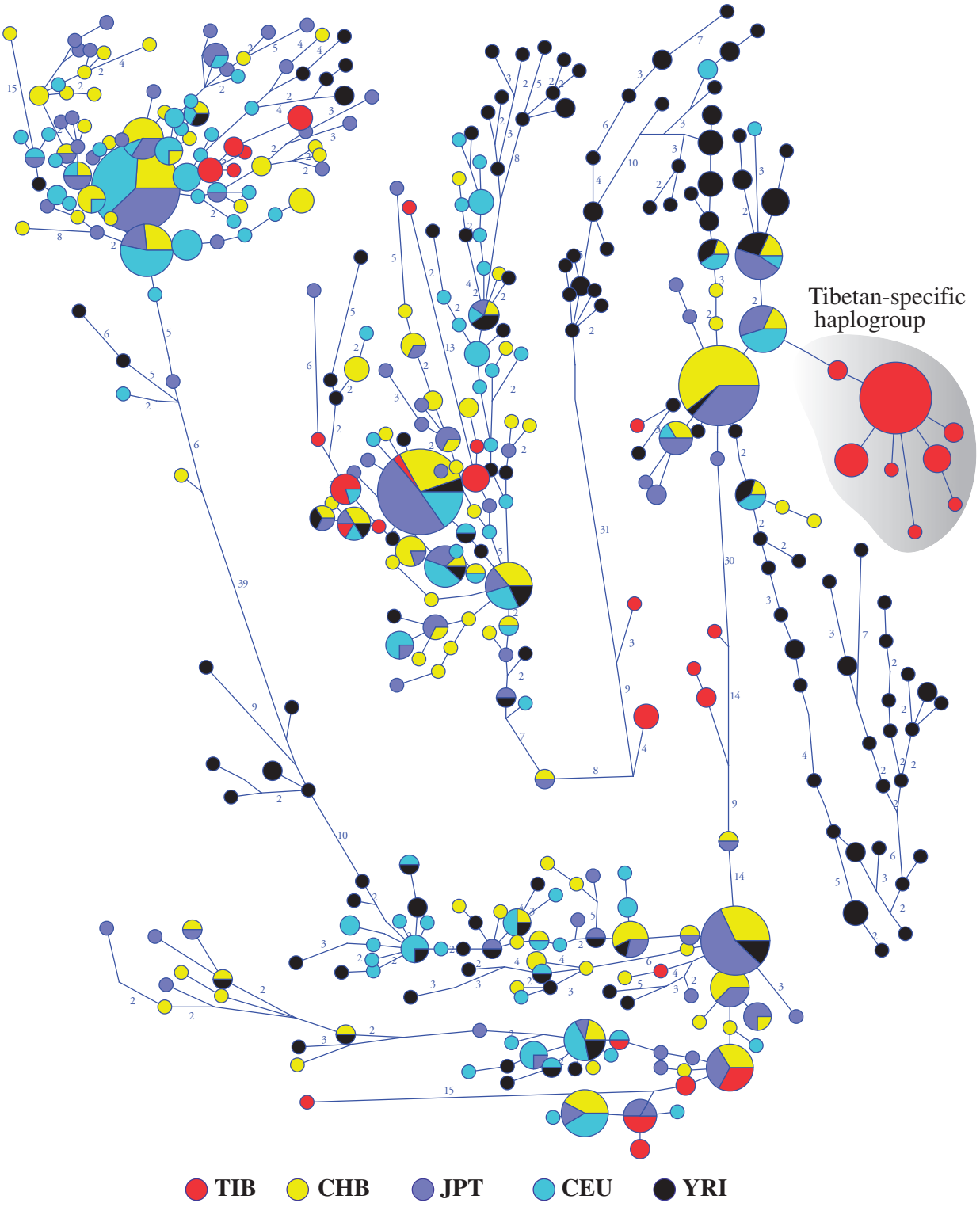

Fig. 3. Median-joining network of *EGLN1* showing a Tibetan-specific haplogroup. Each node represents a haplotype, and the size is proportional to its frequency. The length of a branch is roughly proportional to the mutation steps. When there is more than one mutation, the numbers are labeled on the branches. The Tibetan-specific haplogroup is highlighted, which is defined by three SNPs, rs186996510, rs973254, and rs1538664.

supports the action of positive selection. We observed a Tibetan-specific haplotype, which is prevalent in Tibetans (48.91%), but absent in non-Tibetan populations (fig. 3). This Tibetan-specific haplotype is defined by three SNPs including the nonsynonymous SNP rs186996510 (D4E) and two other SNPs (rs973254 and rs1538664) but not the other nonsynonymous SNP rs12097901 (S127C), suggesting that these two nonsynonymous SNPs are not tightly linked, as reflected by the low level of linkage between them ($r^2 = 0.29$) (fig. 1). Collectively, the pattern of the *EGLN1*

Table 1. Allele Frequencies of the Seven Tag SNPs in Tibetan and Non-Tibetan Populations (*$P < 0.05$).

	Tibetan								Correlation with Altitude (Spearman's)	Non-Tibetan			
Population	Nyingchi	Lhasa	Chamdo-1	Chamdo-2	Xigaze	Shannan	Nagri	Nagqu		CHB	JPT	CEU	YRI
Altitude (m)	2,750	3,400	3,500	3,850	3,890	3,935	4,700	4,800		—	—	—	—
Sample Size	118	119	47	70	93	118	120	118		97	89	85	88
rs186996510(C)	0.6992	0.6429	0.7234	0.6549	0.7151	0.7119	0.7583	0.7542	0.714*	0.0103	0.0056	0.0059	0.0227
rs11805033(C)	0.8000	0.7344	0.8111	0.7500	0.7697	0.7682	0.8419	0.7650	0.095	0.4021	0.4270	0.2118	0.3239
rs2244986(T)	0.1727	0.1823	0.1667	0.2000	0.1798	0.1955	0.1154	0.1709	−0.238	0.5412	0.5562	0.6529	0.2727
rs2739506(T)	0.7955	0.7500	0.8222	0.7500	0.7978	0.8091	0.8376	0.7906	0.311	0.4485	0.4382	0.3471	0.7273
rs2491407(G)	0.8045	0.7396	0.8222	0.7500	0.8090	0.7773	0.8419	0.7863	0.238	0.4485	0.4382	0.3471	0.7273
rs2491406(A)	0.1955	0.2604	0.1778	0.2500	0.1910	0.2182	0.1752	0.2137	−0.238	0.5464	0.5562	0.6471	0.2727
rs2486745(T)	0.7955	0.7396	0.8222	0.7500	0.8034	0.7818	0.8376	0.7906	0.238	0.4485	0.4382	0.3471	0.7273

haplotype network is consistent with the proposed act of positive selection on *EGLN1*, leading to the prevalence of population-specific haplotypes in Tibetans.

Genotyping of Tag SNPs in Extensive Tibetan Populations

To test whether the observed sequence variation pattern from the resequencing data of 46 Tibetan individuals also holds for the entire Tibetan region, we selected eight Tibetan populations (a total of 803 individuals) covering all seven prefectures of Tibet (supplementary fig. S1, Supplementary Material online), and we genotyped seven *EGLN1* SNPs (rs12097901 was not genotyped, fig. 1). As listed in table 1, all seven SNPs showed similar allele frequency spectrum with those observed in the resequencing data (supplementary table S1, Supplementary Material online), and no large differences were seen among different Tibetan populations, implying that populations living in Tibet are relatively homogeneous. In these Tibetan populations, the frequency of rs186996510 (D4E) ranges from 64.3% to 75.8%, indicating that natural selection on *EGLN1* has been consistent across Tibet.

Assuming that *EGLN1* contributes to high-altitude hypoxic adaptation, it is expected that Tibetan populations living at different altitudes would have been under different selective strengths. To test this, we analyzed the correlations between the seven tag SNPs and altitudes. Interestingly, only rs186996510 (D4E) showed a significant correlation (Spearman's $\rho = 0.714$, $P = 0.047$, two tailed) (table 1 and fig. 4), again an implication of its contribution to high-altitude hypoxic adaptation in Tibetans.

Estimation of Selection Intensity and Age of the Selected Allele

Following our previous method (Peng et al. 2011), we assessed the intensity and age of selection on *EGLN1* in Tibetans. The C allele of rs186996510 that is highly prevalent in Tibetans was considered as the selected allele. The estimated age of selection is 8,391 years (95% confidence interval: 8,264–8,519), a time falling into the early Neolithic. Assuming the selected allele started in Tibetans with a very low frequency ($f = 0.0001$), the estimated selection intensity is 0.029

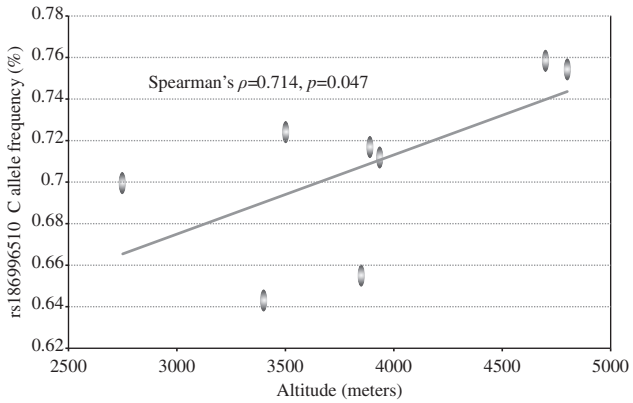

Fig. 4. Correlation between the C allele frequency of rs186996510 (D4E) and altitude.

(95% confidence interval: 0.028–0.031). We also considered a soft sweep model, in which the selected allele was at a relatively high frequency in the ancestral Tibetan population before the selective sweep. Using the allele frequency of the selected mutant in CHB ($f = 0.0103$) as an approximation of the allele frequency before selection in ancestral Tibetans, we estimated the selection intensity as 0.016. Whether starting with extremely low or low frequency before selection, *EGLN1* is among the top list of genes showing adaptive evolution in human populations (Tishkoff et al. 2001, 2007; Bersaglieri et al. 2004; Peng et al. 2011).

Genetic Association Analysis of Adaptive Phenotypes in Tibetans

To detect whether *EGLN1* contributes to the known physiologically adaptive features in Tibetans, we collected data of hemoglobin concentration and degree of blood oxygen saturation from 620 Tibetan individuals (Kangma County of Tibet). The seven tag SNPs and rs12097901 (S127C) were genotyped. Because the hemoglobin levels are different between males and females, we conducted genetic association analyses separately for each sex (age and altitude were considered as covariate). As presented in table 2, in Tibetan males, two SNPs showed nominally significant association with hemoglobin levels ($P = 0.0067$ for rs186996510, $P = 0.022$ for

Table 2. Association of Eight *EGLN1* SNPs with Levels of Hemoglobin and Blood Oxygen Saturation in Tibetans.

	SNP	Position	Allele[a]	Hemoglobin Level					Degree of Blood Oxygen Saturation				
				N	*β*	SE	*P*	Corrected *P*	*N*	*β*	SE	*P*	Corrected *P*
Male	rs2486745	231505717	C	261	0.25	0.26	0.33	/	262	−1.16	1.12	0.30	/
	rs2491406	231518665	A	261	0.29	0.25	0.26	/	262	−1.12	1.10	0.31	/
	rs2491407	231518928	A	261	0.29	0.25	0.26	/	262	−1.12	1.10	0.31	/
	rs2739506	231529183	G	261	0.29	0.25	0.26	/	262	−1.12	1.10	0.31	/
	rs2244986	231537452	T	261	0.28	0.26	0.27	/	262	−1.37	1.11	0.22	/
	rs11805033	231546452	T	261	0.25	0.26	0.32	/	262	−1.09	1.10	0.33	/
	rs12097901	231557255	C	261	0.61	0.27	0.022	0.061	262	−1.93	1.16	0.097	/
	rs186996510	231557623	G	261	0.62	0.23	0.0067	0.023	262	−1.67	0.99	0.092	/
Female	rs2486745	231505717	C	355	−0.10	0.25	0.69	/	356	−0.55	0.85	0.52	/
	rs2491406	231518665	A	355	−0.08	0.25	0.73	/	356	−0.61	0.85	0.47	/
	rs2491407	231518928	A	355	−0.08	0.25	0.73	/	356	−0.61	0.85	0.47	/
	rs2739506	231529183	G	355	−0.11	0.24	0.65	/	356	−0.52	0.85	0.54	/
	rs2244986	231537452	T	355	−0.13	0.25	0.61	/	356	−0.58	0.86	0.50	/
	rs11805033	231546452	T	355	−0.10	0.25	0.69	/	356	−0.55	0.85	0.52	/
	rs12097901	231557255	C	355	0.19	0.25	0.43	/	356	−0.40	0.85	0.64	/
	rs186996510	231557623	G	355	0.20	0.22	0.37	/	356	−1.01	0.76	0.19	/
All	rs2486745	231505717	C	616	0.08	0.18	0.66	/	618	−0.85	0.68	0.21	/
	rs2491406	231518665	A	616	0.10	0.18	0.56	/	618	−0.87	0.67	0.20	/
	rs2491407	231518928	A	616	0.10	0.18	0.56	/	618	−0.87	0.67	0.20	/
	rs2739506	231529183	G	616	0.08	0.18	0.63	/	618	−0.82	0.67	0.23	/
	rs2244986	231537452	T	616	0.09	0.18	0.63	/	618	−0.97	0.68	0.15	/
	rs11805033	231546452	T	616	0.08	0.18	0.65	/	618	−0.82	0.67	0.22	/
	rs12097901	231557255	C	616	0.37	0.18	0.039	0.091	618	−1.04	0.69	0.14	/
	rs186996510	231557623	G	616	0.38	0.16	0.018	0.049	618	−1.29	0.61	0.034	0.086

NOTE.—Significant results are underlined. *β*, the difference per copy increase of the effect allele; SE, standard error; *N*, number of tested individuals for the SNP.
[a]Effect allele.

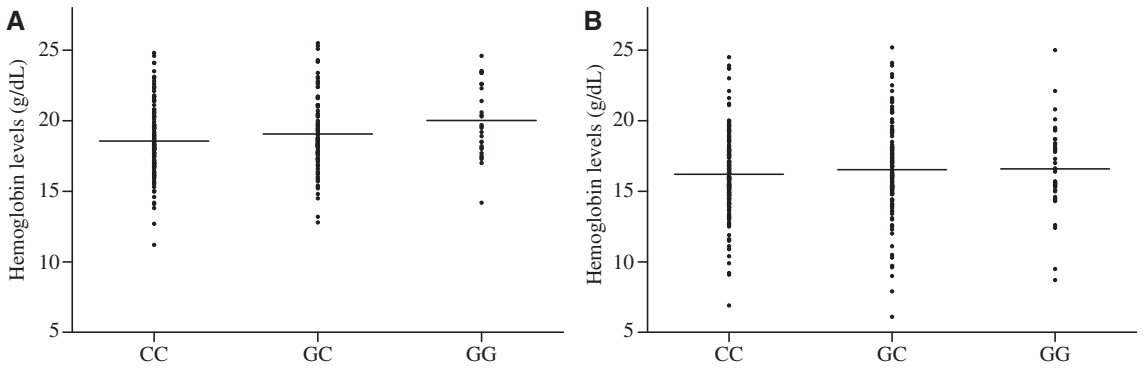

FIG. 5. Comparison of hemoglobin levels among three different genotypes (CC, CG, and GG) at rs186996510 (D4E) in Tibetan males (*A*) and females (*B*).

rs12097901). After multiple test corrections, only rs186996510 remained significant (corrected $P = 0.023$). On average, there is about 7% reduction of hemoglobin level for the homozygotes (CC) of the presumably adaptive allele of rs186996510 compared with the level of the other homozygotes (GG) (fig. 5*A*). In contrast, none of the SNPs showed significant association in females ($P > 0.3$) though the trend was the same as in males (fig. 5*B*). When males and females were combined together (age, sex, and altitude were considered as covariates), the association with hemoglobin was still significant for rs186996510 (corrected $P = 0.049$) (table 2). Similar pattern was observed for the association with degree of blood oxygen saturation though no significance was detected after multiple test corrections (table 2). Additionally, an SNP–SNP interaction analysis combining the two nonsynonymous SNPs (rs186996510, D4E and rs12097901, S127C) did not reveal significant association either for hemoglobin or for blood oxygen saturation ($P > 0.5$).

Discussion

EGLN1 is one of the key genes functioning as an upstream regulator in the HIF pathway. Although previous studies have proposed its potential involvement in the genetic adaptation to high-altitude hypoxia in Tibetans, without resequencing data of the entire gene region of *EGLN1* and an extensive survey of Tibetan populations, it is hard to test the signal of selection and identify causal sequence variations. In this study,

we conducted resequencing of the entire region of *EGLN1* (59.4 kb) and generated a complete list of sequence variations (SNPs and in-dels) in Tibetans. Based on the analyses of between-population allelic divergence (F_{ST}), neutrality test, and haplotype network construction, we showed that *EGLN1* has been under Darwinian positive selection, leading to the enrichment of a Tibetan-specific mutation (rs186996510, D4E), a promising candidate SNP explaining *EGLN1*'s contribution to high-altitude hypoxic adaptation in Tibetans.

As an amino acid changing mutation, rs186996510 (D4E) is located at the N-terminal of the EGLN1 protein (fig. 6) but not in the three known functional domains (McDonough et al. 2006). The amino acid change from aspartic acid to glutamic acid (D4E) is not a drastic change in view of molecular size and charge property. Also, it is hard to predict how this amino acid change would affect the three-dimensional structure of the EGLN1 protein because the X-ray-decoded crystal structure of EGLN1 is not available. However, a protein sequence alignment among distantly related species revealed that the ancestral amino acid (D) is a phylogenetically conserved residue (fig. 6), and no mutation was observed among the seven representative mammalian species with their most recent common ancestor traced back to about 200 Ma (Meredith et al. 2011), suggesting that this amino acid site is functionally important for EGLN1, and the D to E mutation may cause protein functional change. In contrast, the other amino acid changing site S127C is not conserved (fig. 6), and the allelic divergence between Tibetans and non-Tibetans at this site is much smaller ($F_{ST} = 0.160$ between Tibetans and Han Chinese) compared with the D4E site ($F_{ST} = 0.709$). However, we cannot rule out the possibility that S127C may also be functional.

Genetic association analysis of the physiological features in Tibetans also supports *EGLN1*'s contribution to high-altitude hypoxic adaptation. The Tibetan-specific mutation (rs186996510, D4E) showed the strongest association with hemoglobin levels, consistent with a previous report showing haplotypic association of three SNPs 37 kb upstream of *EGLN1* (Simonson et al. 2010). The homozygotes of the presumably adaptive allele of rs186996510 lead to a reduced level of hemoglobin, which explains the adaptively much lower occurrence of excessive polycythemia in Tibetans compared with the acclimatized lowlanders at high altitude (Wu and Kayser 2006). The different results in association analysis with hemoglobin levels between male and female Tibetans are probably caused by sex hormones as previous studies reported that testosterone stimulates and ovarian hormones blunt hemoglobin production at altitude (Fried 1995; Leon-Velarde et al. 2001). Alternatively, population stratification, for example, a sex-biased pattern of migration, could also cause the sex differences observed in the hemoglobin association results, which is yet to be tested.

As far as gene function is concerned, *EGLN1* is a negative regulator of *EPAS1*, a gene directly regulating the expression of erythropoietin (*EPO*) and eventually influencing the level of hemoglobin in the blood. Because the relatively lower hemoglobin level compared with acclimatized lowlanders is one of the key adaptive features in Tibetans, we speculate that the D to E mutation at rs186996510 may enhance the enzyme activity of the EGLN1 protein, leading to an increased degradation of EPAS1, and therefore relatively less production of *EPO* under hypoxic conditions.

It should be noted that *EPAS1* also underwent strong Darwinian positive selection in Tibetans (Beall et al. 2010; Bigham et al. 2010; Yi et al. 2010; Peng et al. 2011; Xu et al. 2011), and the functional mutation(s) of *EPAS1* may also have a contribution to hemoglobin level. Additionally, there are also other reported hemoglobin-regulating genes showing signal of positive selection in Tibetans, for example, *HMOX2* (hemeoxygenase 2 that cleaves the heme ring at the alpha methene bridge to form biliverdin), *HBB*, and *HBG2* (the subunit of β globin and γ globin, respectively) (Peng et al. 2011). Consequently, the adaptively low hemoglobin levels in Tibetans are likely the outcome of adaptive changes of multiple genes and gene–gene interactions in the HIF pathway.

Intriguingly, the estimated age of selection on the C allele of rs186996510 falls in the early Neolithic (~8,400 years ago), which is much younger than the estimated selection age of *EPAS1* (~18,000 years ago) (Peng et al. 2011). There are two possible scenarios explaining the difference. First, the adaptive mutations of *EGLN1* and *EPAS1* were brought onto the Himalayas at different times. According to the recent population data of mtDNA and Y chromosome diversity, there had been two major migratory waves into the Himalayas, an early one during the Upper Paleolithic about 30,000 years ago and a latter one about 10,000–7,000 years ago (Shi et al. 2008). The selection age of *EGLN1* coincides with the second migration onto the Himalayas during the early Neolithic. Second, the adaptive mutations of *EGLN1* and *EPAS1* might be brought onto the Himalayas at the same time during the Upper Paleolithic, but selection might have occurred at different times on these two genes. *EPAS1* was likely selected first when people were living at relatively low altitude (2,700–3,000 m) during the initial colonization. With the Neolithic population expansion, people began to explore higher altitude (>3,000 m) and selection kicked in for *EGLN1* that helped to adapt to higher altitudes. This notion is consistent with the observation that the frequency of the C allele (presumably the adaptive allele) of rs18688510 is positively correlated with altitude (fig. 4), and no such correlation was detected for SNPs in *EPAS1* (Peng et al. 2011). More data are needed to test these two hypotheses.

Materials and Methods

Tibetan Samples

A total of 46 Tibetan individuals were used for resequencing, covering the entire genomic region of *EGLN1* (59.4 kb). These individuals were from the 50 individuals analyzed previously by the Affimatrix Genome-wide Human SNP Array 6.0 who were randomly selected from multiple Tibetan populations (Peng et al. 2011). The sample size (92 chromosomes) should be sufficient to detect SNPs showing unusual allelic differences between Tibetan highlanders and reference lowlander populations (e.g., Han Chinese). For population survey of seven *EGLN1* tag SNPs, we collected a total of 803 Tibetan

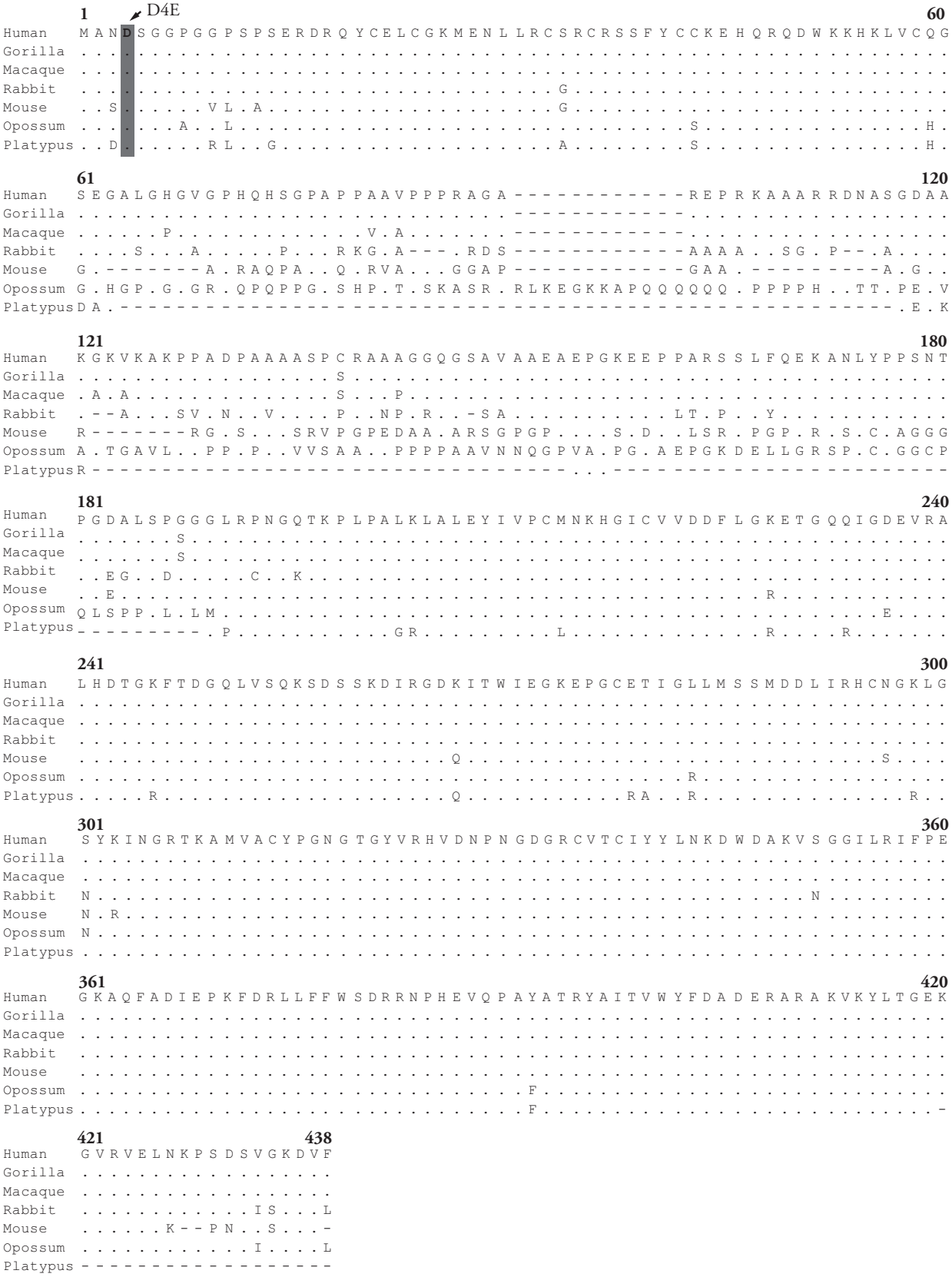

Fig. 6. Protein sequence alignment of EGLN1. The nonsynonymous site (rs186996510, D4E) in humans is highly conserved among representative mammalian species. The dots indicate the same sequences compared with the human copy, and the dashes indicate deletions.

individuals, representing eight populations from all seven prefectures of Tibet (Nyingchi, Lhasa, Chamdo, Xigaze, Shannan, Nagri, and Nagqu) (supplementary fig. S1, Supplementary Material online). For genetic association analysis with hemoglobin levels, we collected 620 Tibetan samples (262 men and 358 women, age: 40.9 ± 13.4 years) from Kangma County of Tibet (elevations ranging from 4,500 to 5,100 m). The blood samples were obtained from each individual with written informed consents. The protocol of this study was reviewed and approved by the Internal Review Board of Kunming Institute of Zoology, Chinese Academy of Sciences (approval no.: SWYX-2012008).

Resequencing and between-Population Divergence Analysis

Sequencing of the *EGLN1* genomic region was conducted using an ABI 3730 sequence analyzer (Applied Biosystems, Foster City, CA). We first amplified 13 overlapping fragments of 2.5–7.5 kb using long polymerase chain reaction (PCR) amplification, covering the entire 59.4 kb of the *EGLN1* genomic region. The inner primers were then used to conduct sequencing. The long PCR and inner sequencing primers are listed in supplementary table S3, Supplementary Material online. We obtained high-quality sequences of the entire 59.4-kb region in those 46 Tibetan individuals. The GenBank accession numbers for the 46 *EGLN1* gene sequences are KC554019–KC554064.

The *EGLN1* sequence variation data of the reference populations (Han Chinese from Beijing, CHB; Japanese, JPT; Europeans, CEU; and Africans, YRI) were retrieved from the 1000 Human Genomes Project (1000 Genomes Project Consortium 2012). There are 163 polymorphic sites (150 SNPs and 13 in-dels) shared between Tibetans and the reference populations. To estimate genetic divergence between Tibetans and the reference populations, we calculated F_{ST} described by Akey et al. (2002).

The LD map of the *EGLN1* sequence variations (166 SNPs) in the 46 Tibetan individuals was constructed by the software Haploview using the r^2 algorithm (Barrett et al. 2005).

Neutrality Test, Haplotype Network Analysis, Allele Age, and Selection Intensity Estimation

A neutrality test of *EGLN1* was conducted by the method of Fay and Wu (2000), a commonly used method to detect recent Darwinian positive selection on a population and is not affected by the false-positive signal due to recent population expansion. The sequence data of a 15-kb fragment of *EGLN1* were used, flanking the SNP (rs186996510) (3.6-kb upstream plus 12.4 kb downstream sequences) showing the largest genetic divergence between Tibetans and Han Chinese ($F_{ST} = 0.709$).

For haplotype network analysis, the 150 shared SNPs among Tibetan and non-Tibetan populations were used, and haplotype inferring was conducted by PHASE embedded in DnaSP Version 5 (Librado and Rozas 2009). A median-joining network was constructed following the method described in Bandelt et al. (1999). To simplify the network, a maximum parsimony calculation was performed to eliminate superfluous links between haplotypes with the default settings.

The allele age and selective intensity were estimated following previous method in Voight et al. (2006) (see supplementary data from Peng et al. [2011] for details). We first estimated the allele age by the decay of ancestral haplotype sharing as described in Voight et al. (2006). After the point estimate of the allele age was obtained, we further estimated selection intensity by fitting the allele frequency of the selected mutant in the Tibetan population with a deterministic logistic sweep model, in which we assumed the current allele frequency was increased from an initial allele frequency as a function of selection intensity and time (see supplementary data of Peng et al. [2011]). It should be noted that the methods for estimating selection intensity and allele age are crude because we make strong assumptions in the model. In the estimation of allele age, we assumed that haplotypes are independent of each other ("star genealogy") and thus recombination histories along the genealogy are independent. In the estimation of selection intensity, we assumed a deterministic selective sweep model ignoring the randomness of selected allele frequency trajectories. Because of the above strong assumptions, the seemingly narrow confidence intervals estimated by the method of Voight et al. (2006) are likely underestimates of the true variance.

Measurement of Levels of Hemoglobin and Blood Oxygen Saturation

Arterial oxygen saturation at rest (SaO2rest) was recorded after the subjects laid down for 2–3 min using a hand-held pulse oximeter (Nellcor NPB-40, CA). For hemoglobin concentration, a HemoCue Hb 201+ analyzer (Ängelholm, Sweden) was applied to measure Hb for finger tip capillary blood. Among the sampled 620 Tibetan individuals, 616 and 618 individuals were recorded for hemoglobin concentration and arterial oxygen saturation, respectively.

Genotyping of *EGLN1* Tag SNPs in Multiple Tibetan Populations

We selected seven SNPs (rs186996510, rs11805033, rs2244986, rs2739506, rs2491407, rs2491406, and rs2486745) for population diversity survey, representing the major LD blocks of *EGLN1* in Tibetans (fig. 1). Genotyping was conducted using the SNaPshot method on an ABI 3730 sequencer (Applied Biosystems, Foster City, CA), and we designed a series of primers for covering these genomic regions. The SNaPshot results were confirmed by sequencing method. For the correlation analysis between the allele frequencies of the seven tag SNPs and altitudes, we calculated Spearman's ρ using the program in SPSS version 20 (SPSS Inc.).

Genetic Association Analysis

We collected a total of 620 individuals for genetic association analysis of hemoglobin levels and arterial oxygen saturation levels. DNA samples were extracted from blood following the standard method. Association analysis was performed using linear regression with an additive model in PLINK v1.07

(Purcell et al. 2007). Age, sex, and altitude were considered as covariates for statistical assessment. To test potential epistatic effect of two SNPs (e.g., SNP1 and SNP2), the effect of SNP1 was assessed separately in the subgroups divided by distinct SNP2 genotypes, and the χ^2 test with the software PLINK v1.07 was used.

Because the SNPs may not be totally independent due to LD, to avoid type II error when applying the traditional Bonferroni correction, we corrected the *P* values with a max (*T*) permutation procedure implemented in PLINK, and the "–mperm" option was used ($n = 1{,}000$), which takes a single parameter, the number of permutations to be performed. This is achieved by comparing each observed test statistic against the maximum of all permuted statistics (i.e., over all SNPs) for each single replicate.

Supplementary Material

Supplementary figure S1 and tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to the volunteers who donated their samples for this study. They thank all participating staff from High Altitude Medical Research Center, School of Medicine, and Tibetan University for their assistance during sample collection. This work was supported by the National 973 Program of China (2012CB518202 to T.W. and X.Q., 2011CB512107 to Ou.), the National Natural Science Foundation of China (91231203 and 31123005 to B.S., and 91131001 to H.S.), State Key Laboratory of Genetic Resources and Evolution (GREKF11-02 to Ou., GREKF12-04 to Bi., GREKF13-04 to T.W. and GREKF10-02 to C.C.), and the Natural Science Foundation of Yunnan Province (2009CD107 and 2010CI044 to H.S.).

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12:1805–1814.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol.* 16:37–48.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265.

Beall CM. 2007. Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *Proc Natl Acad Sci U S A.* 104: 8655–8660.

Beall CM, Cavalleri GL, Deng L, et al. (29 co-authors). 2010. Natural selection on *EPAS1* (*HIF2α*) associated with low hemoglobin concentration in Tibetan highlanders. *Proc Natl Acad Sci U S A.* 107: 11459–11464.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet.* 74:1111–1120.

Bigham A, Bauchet M, Pinto D, et al. (14 co-authors). 2010. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet.* 6:e1001116.

Dupuy D, Aubert I, Duperat VG, Petit J, Taine L, Stef M, Bloch B, Arveiler B. 2000. Mapping, characterization, and expression analysis of the SM-20 human homologue, c1orf12, and identification of a novel related gene, SCAND2. *Genomics* 69:348–354.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.

Fried W. 1995. Erythropoietin. *Annu Rev Nutr.* 15:353–377.

Jaakkola P, Mole DR, Tian YM, et al. (13 co-authors). 2001. Targeting of HIF-*α* to the von Hippel-Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. *Science* 292:468–472.

Leon-Velarde F, Rivera-Chira M, Tapia R, Huicho L, Monge CC. 2001. Relationship of ovarian hormones to hypoxemia in women residents of 4,300 m. *Am J Physiol Regul Integr Comp Physiol.* 280: R488–R493.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.

McDonough MA, Li V, Flashman E, et al. (17 co-authors). 2006. Cellular oxygen sensing: crystal structure of hypoxia-inducible factor prolyl hydroxylase (*PHD2*). *Proc Natl Acad Sci U S A.* 103:9814–9819.

Meredith RW, Janecka JE, Gatesy J, et al. (22 co-authors). 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* 334:521–524.

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.

Peng Y, Yang Z, Zhang H, et al. (15 co-authors). 2011. Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Mol Biol Evol.* 28:1075–1081.

Purcell S, Neale B, Todd-Brown K, et al. (11 co-authors). 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 81:559–575.

Shi H, Zhong H, Peng Y, et al. (13 co-authors). 2008. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. *BMC Biol.* 6:45.

Simonson TS, Yang Y, Huff CD, et al. (12 co-authors). 2010. Genetic evidence for high-altitude adaptation in Tibet. *Science* 329:72–75.

Tishkoff SA, Reed FA, Ranciaro A, et al. (19 co-authors). 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet.* 39:31–40.

Tishkoff SA, Varkonyi R, Cahinhinan N, et al. (17 co-authors). 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* 293: 455–462.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.

Wang B, Zhang YB, Zhang F, et al. (18 co-authors). 2011. On the origin of Tibetans and their genetic basis in adapting high-altitude environments. *PLoS One* 6:e17002.

Wu T, Kayser B. 2006. High altitude adaptation in Tibetans. *High Alt Med Biol.* 7:193–208.

Xu S, Li S, Yang Y, et al. (13 co-authors). 2011. A genome-wide search for signals of high-altitude adaptation in Tibetans. *Mol Biol Evol.* 28: 1003–1011.

Yi X, Liang Y, Huerta-Sanchez E, et al. (67 co-authors). 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329:75–78.