A Genome-Wide Search for Signals of High-Altitude Adaptation in Tibetans

Shuhua Xu,*[,1,2] Shilin Li,[3] Yajun Yang,[3] Jingze Tan,[3] Haiyi Lou,[1] Wenfei Jin,[1] Ling Yang,[1] Xuedong Pan,[3] Jiucun Wang,[3] Yiping Shen,[4] Bailin Wu,[3,4] Hongyan Wang,[3] and Li Jin*[,1,2,3]

[1]Chinese Academy of Sciences and Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

[2]Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai, China

[3]State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai, China

[4]Children's Hospital Boston, Harvard Medical School

***Corresponding author:** E-mail: xushua@picb.ac.cn; ljin007@gmail.com.

Associate editor: Naruya Saitou

Abstract

Genetic studies of Tibetans, an ethnic group with a long-lasting presence on the Tibetan Plateau which is known as the highest plateau in the world, may offer a unique opportunity to understand the biological adaptations of human beings to high-altitude environments. We conducted a genome-wide study of 1,000,000 genetic variants in 46 Tibetans (TBN) and 92 Han Chinese (HAN) for identifying the signals of high-altitude adaptations (HAAs) in Tibetan genomes. We discovered the most differentiated variants between TBN and HAN at chromosome 1q42.2 and 2p21. *EGLN1* (or *HIFPH2*, MIM 606425) and *EPAS1* (or *HIF2A*, MIM 603349), both related to hypoxia-inducible factor, were found most differentiated in the two regions, respectively. Strong positive correlations were also observed between the frequency of TBN-dominant haplotypes in the two gene regions and altitude in East Asian populations. Linkage disequilibrium and further haplotype network analyses of world-wide populations suggested the antiquity of the TBN-dominant haplotypes and long-term persistence of the natural selection. Finally, a "dominant haplotype carrier" hypothesis could describe the role of the two genes in HAA. All of our population genomic and statistical analyses indicate that *EPAS1* and *EGLN1* are most likely responsible for HAA of Tibetans. Interestingly, one each but not both of the two genes were also identified by three recent studies. We reanalyzed the available data and found the escaped top signal (*EPAS1*) could be recaptured with data quality control and our approaches. Based on this experience, we call for more attention to be paid to controlling data quality and batch effects introduced in public data integration. Our results also suggest limitations of extended haplotype homozygosity-based method due to its compromised power in case the natural selection initiated long time ago and particularly in genomic regions with recombination hotspots.

Key words: high-altitude adaptation, Tibetan, SNP, natural selection, hypoxia, hypoxia-inducible factor, IPA.

Introduction

The Tibetan Plateau, known as "the roof of the world" and with an average elevation of over 4,500 meters, is the highest plateau in the world (see also supplementary data, fig. S1, Supplementary Material online). It is almost certain that biological adaptability has contributed to the success of the ethnic groups, such as Sherpas and Tibetans in occupying the plateau because traditional technology could not buffer them from the unavoidable environmental stress of severe lifelong high-altitude hypoxia (Beall 2007). Compared with other indigenous highlanders living in the Andean Altiplano and in the Ethiopian Highlands, Tibetans developed a distinctly different physiological characteristics for surviving in the environment of oxygen-thin air (Beall et al. 2002; Beall 2007).

Efforts based on candidate gene approaches so far have not been successful to identify specific genetic loci contributing to high-altitude adaptation (HAA) (Beall 2007). Population genomics offers a potential solution to the limitations of candidate gene studies, where the genome can be surveyed without any a priori assumptions regarding which genes may be under selection and population demographic history and natural selection can be distinguished (Akey 2009). A genome-wide (GW) study by sampling a large number of loci throughout the genome and searching "outliers" in the extreme tail of the empirical distribution, which is the most commonly used population genomics approach (Akey 2009), may provide the clue to the understanding of biological mechanism underlying HAA of Tibetans. In this study, using Affymetrix Genome-Wide Human SNP Array 6.0 which interrogates more than 900 K single nucleotide polymorphisms (SNPs) and 900 K copy number variations (CNVs) probes encompassing the entire genome, we conducted a GW scan to detect the signals of HAAs in genomes of Tibetan people. About 1,000,000 genetic variants (SNPs and CNVs) were

investigated in 46 Tibetans (TBN), and they were compared with those of 92 individuals sampled from Han Chinese (HAN), a population that has been shown to be linguistically and genetically related to TBN but has been living in low-altitude environment. Recent GW SNP data from the International HapMap Project (Frazer et al. 2007) and Human Genome Diversity Panel (HGDP) (Li et al. 2008) were also included in this study for haplotype analyses of worldwide populations. Three recent GW studies (Beall et al. 2010; Simonson et al. 2010; Yi et al. 2010) identified two hypoxia-inducible factor (HIF) related genes as contributing to HAA in Tibetans. One of the two genes, *EPAS1* (or *HIF2A*), showed the top GW signal in both Beall et al. and Yi et al. but did not show significant statistical signals in Simonson et al., in which the other gene, *EGLN1* (or *HIFPH2*), showed the top GW signal. We explored the causes and discussed why the three recent studies (Beall et al. 2010; Simonson et al. 2010; Yi et al. 2010) each missed one of the top two genes in their GW scans, and we also showed how the escaped top GW signal in Simonson et al. can be recaptured with our reanalysis of the available data.

Materials and Methods

Populations and Samples

DNA samples of 51 Tibetan unrelated individuals were collected from the five regions in Tibet (Shannan, Rikaza, Linzhi, Lasha and Changdu); DNA samples of 101 Han Chinese unrelated individuals were collected from Gansu, Hebei, Shandong, Zhejiang, Guangdong, Yunnan, thus represent both northern and southern Han Chinese populations (Xu et al. 2009). All procedures followed were in accordance with the ethical standards of the ethics committee of Fudan University and the Helsinki Declaration of 1975, as revised in 2000. Four HapMap population samples (60 YRI, Yoruba from Ibadan, Nigeria; 60 CEU, Utah residents with ancestry from northern and western Europe; 45 CHB, Han Chinese in Beijing; and 44 JPT, Japanese in Tokyo) (Frazer et al. 2007) were also included in this study. Recent GW SNP data (Li et al. 2008) of 17 East Asian population samples (Yakut, Oroqen, Mongola, Daur, Hezhen, Xibo, Han-NChina, Japanese, Tu, Han, Tujia, Yi, Miao, She, Naxi, Lahu, Dai) in HGDP (Cann et al. 2002) were also included in haplotype-based analyses.

Genotyping, CNV Detection, and Data Quality Control

Genotyping with the Affymetrix Genome-Wide Human SNP Array 6.0 was performed on the Affymetrix genotyping platform at Fudan University, Shanghai, according to the "48 Sample Protocol" (Affymetrix, *Genome-Wide Human SNP Nsp/Sty 6.0 User Guide, Rev. 3, 2008*, P/N 702504). *.CEL files containing raw intensity data were analyzed with Birdsuite version 1.5.3 (Korn et al. 2008). Five Tibetan and nine Han Chinese samples with a call-rate below 93% were excluded from further analyses. Total genotyping rate in remaining 46 Tibetan and 92 Han Chinese samples is 98.6%. There are 1,989 SNPs of the total 934,968 SNPs in Affymetrix 6.0 chip with genomic positions unknown, 2,977 duplicate SNPs of the remaining 932,979 SNPs according to RS number, one of each pair with less missing genotypes was kept for further data filtration. All together, 47,417 heterozygous haploid genotypes were found and considered as missing data; 9,542 SNPs with missing data $>$ 20% samples were excluded; SNPs failed Hardy–Weinberg equilibrium (HWE) test ($P < 0.0001$) within Tibetan population (195 SNPs) or Han Chinese (530 SNPs) were excluded; In addition, 147,814 SNPs were monomorphic in both Tibetan and Han Chinese samples were not included in further analysis, of which 137,153 were autosomal SNPs. The above data filtering procedure finally yielded 777,262 SNPs were common to both Tibetan and Han Chinese populations, of which 748,994 were autosomal SNPs.

CNV detection was also performed by Birdsuite, version 1.5.3 (Korn et al. 2008). We filtered segment with length less than 1 kb or number of probes less than 3 or logarithm [base 10] of odds score less than 5. All probe coordinates were mapped to the human genome assembly build 36 (hg18). We apply the widely used term to delineate the characteristics. We integrated both Canary and Birdseye results to generate Copy Number Variable Region map. CNV sharing analysis was conducted by counting the number of sharing regions between populations. The *P* value was calculated by testing on a contingency table with two populations and two types of allele counts (loss-allele and nonloss-allele or gain-allele and nongain-allele). Population-specific gene enrichment analysis was performed between Tibetan and Han Chinese and for deletion, duplication separately. Then, we made gene enrichment analysis for those CNV region overlapping genes by DAVID (http://david.abcc.ncifcrf.gov/).

Determination of Ancestral Alleles

A Perl script was written to obtain ancestral allele status of SNPs from NCBI dbSNP database when both ancestral allele and strand are available. The information of ancestral state of a SNP was not used if it is mapped on more than one genomic region. Altogether 4,180,813 SNPs were annotated with ancestral allele information. In addition, we also downloaded an annotation file (snp130OrthoPt2Pa2Rm2.txt) from UCSC website which contains the orthologous status of chimpanzee, orangutan, and macaque for SNPs. We identified the shared allele status among the three species as ancestral allele of each SNP. Altogether, 11,797,184 SNPs were annotated with ancestral allele information. Among the 3,875,650 SNPs overlapping between NCBI and UCSC with ancestral allele information, 3,870,940 (99.9%) SNPs showed identical ancestral allele status.

Statistical and Population Genetic Analyses

Principal components analysis (PCA) was performed at the individual level using EIGENSOFT version 3.0 (Patterson et al. 2006). Expected heterozygosity was calculated based on 748,994 autosomal SNPs and averaged for each population. Genetic difference between populations was measured using F_{ST} following Weir and Cockerham (1984)

which accounts for differences in the sample size in each population. The SNP-specific F_{ST} between Tibetan and Han Chinese for every SNP was also calculated using the same formula. Haplotypes of 22 autosomes were estimated for each individual from its genotypes with fastPHASE (Scheet and Stephens 2006) version 1.2. "Population labels" were applied during the model fitting procedure to enhance accuracy. The number of haplotype clusters was set to 2, the number of random starts of the EM algorithm (-T) was set to 20, and the number of iterations of EM algorithm (-C) was set to 50. This analysis was used to generate a "best guess" estimate of the true underlying patterns of haplotype structure (Scheet and Stephens 2006). In this study, r^2 was calculated following Hill and Weir (1994) to measure linkage disequilibrium (LD) between SNPs. Networks of haplotypes were constructed using the Network package, version 4.5.1.0 (Fluxus Engineering). Networks were calculated by the median-joining method ($\epsilon = 0$) (Bandelt et al. 1995) after having processed the data with the reduced-median method. The iHS and XP-EHH scores (Sabeti et al. 2007) were calculated with code downloaded from the Pritchard lab webpage (http://hgdp.uchicago.edu). The obtained iHS statistics were normalized in 20 derived allele frequency bins, each spanning 5%. The collection of XP-EHH log-ratios is standardized, such that the resultant distribution has zero mean and unit variance. The XP-CLR scores (Chen et al. 2010) were estimated using code provided by Hua Chen (Department of Genetics, Harvard Medical School). A set of grid points as the putative selected allele positions are placed along the chromosome with a spacing of 2 kb, the window size was chosen to be 0.5 cM, and the number of SNPs in each window was fixed equal to be 100.

Ingenuity Pathway Analysis

SNPs showing substantial differences among clusters were investigated for network and functional interrelatedness using the Ingenuity Pathway Analysis (IPA 8.5, release date: 13 February 2010; content version: 2802, release date: 16 January 2010) software tool Ingenuity Systems, www.ingenuity.com). This web-based entry tool can help the search look for information on genes and/or chemicals, their impact on diseases and cellular processes, and their role in pathways. IPA scans data generated by various large-scale technologies, including gene expression and SNP microarrays, proteomics experiments, and small-scale experiments that generate gene lists to identify networks using information in the Ingenuity Pathways Knowledge Base, a repository of molecular interactions, regulatory events, gene to phenotype associations, and chemical knowledge that pulled from the full text of the peer-reviewed life sciences literatures, and is continuously updated. A molecular network of direct or indirect physical, transcriptional, and enzymatic interactions between mammalian orthologs was computed from this knowledge base. With IPA, we can analyze data in the context of molecular mechanisms, identify key mechanistic differences between subpopulations, and further relate molecular events to higher order cellular and disease processes.

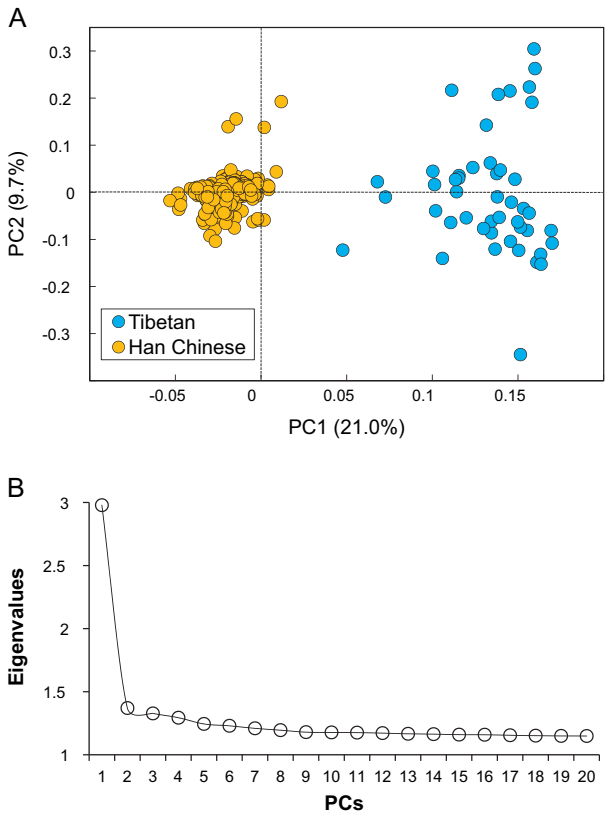

FIG. 1. PCA of Tibetan and Han Chinese individuals. (*A*) Analysis of the first two principal components. (*B*) Distribution of eigenvalues for the first 20 PCs.

Results

Genetic Difference between TBN and HAN

After controls for genotype quality (see Materials and Methods), the 46 TBN samples and 92 HAN samples could be classified as two distinctive clusters in the PCA performed on GW autosomal SNPs (fig. 1). The average heterozygosity of TBN (0.249 ± 0.187) is similar to that of HAN (0.248 ± 0.187), and both populations have similar frequency spectrums based on 748,994 autosomal SNPs (supplementary fig. S2, Supplementary Material online). The genetic difference between TBN and HAN measured by unbiased F_{ST} (Weir and Cockerham 1984) was 0.011 (standard deviation [SD] = 0.00004 based on 1,000 bootstrapping), based on all 917,694 SNPs with missing data less than 20% of samples in both TBN and HAN.

The Most Differential Genomic Regions

A GW distribution of F_{ST} was shown in figure 2A. Among 98 (top 0.0001%) SNPs with $F_{ST} > 0.30$, six of them are located in *EGLN1* (*HIFPH2*) gene region and 25 are in *EPAS1* (*HIF2A*) gene region. Information of the other SNPs were shown in supplementary table S1 (Supplementary Material online). IPA showed that a considerable number of genes associated with them are involved in cardiovascular system

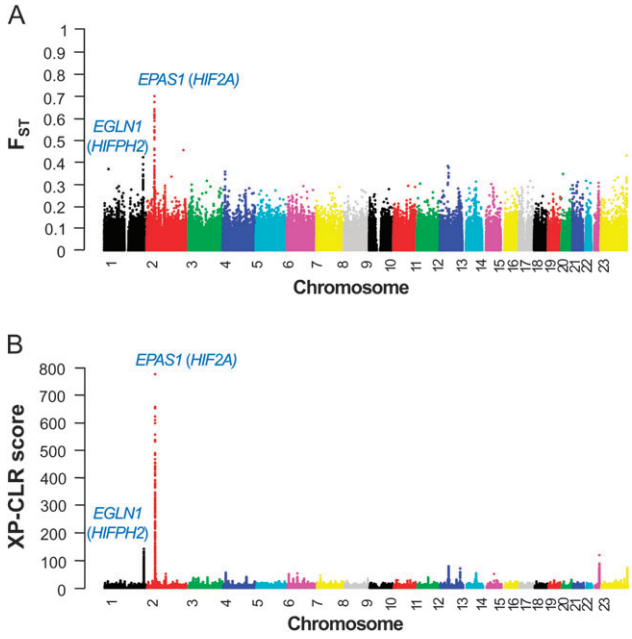

FIG. 2. Genomic distribution of F_{ST} (*A*) and XP-CLR score (*B*). SNP-specific F_{ST} statistic between Tibetan and Han Chinese populations was calculated for every SNP that passes QC. XP-CLR score was calculated following Chen et al. (2010). *EGLN1* on chromosome 1 and *EPAS1* on chromosome 2 were identified in regions showing significant high F_{ST} values (>0.30, top 0.01%) and high XP-CLR scores (>100, top 0.01%).

development and function ($P = 9.58 \times 10^{-5}$), tissue morphology ($P = 1.51 \times 10^{-3}$), and respiratory disease ($P = 1.51 \times 10^{-3}$).

Two HIF Genes

EPAS1 (located in chromosome 2p21) and *EGLN1* (located in chromosome 1q42.2) showed highest differentiation between TBN and HAN as ranked by F_{ST} (Weir and Cockerham 1984) (fig. 2*A*) and XP-CLR scores (Chen et al. 2010) (fig. 2*B*). Interestingly, both genes are associated with HIF which is a transcriptional complex playing a central role in mammalian oxygen homeostasis (Sarkar et al. 2003; Smith et al. 2008). Particularly, *EGLN1* encodes a protein, HIF-prolyl hydroxylase 2, which catalyzes the posttranslational formation of 4-hydroxyproline in HIF-α proteins, whereas *EPAS1* encodes HIF-2α, a transcription factor involved in the induction of genes regulated by oxygen, which is induced as oxygen levels fall. IPA network analysis clearly showed a direct connection between the two genes (supplementary fig. S3, Supplementary Material online). In addition, no selective event was observed in the two gene regions in the comparisons of neighboring low-altitude Asian populations based on GW SNP data of HGDP panel (Li et al. 2008; Pickrell et al. 2009). We therefore hypothesized that the two genes are associated with the HAA of Tibetans and could have subjected to natural selection. We also performed a GW analysis on CNV (see Materials and Methods, supplementary tables S2–S5, fig. S4, Supplementary Material online), but none of the CNVs investigated in this study were implicated.

Global LD, Local LD, EHH, and Recombination Hotspot

The magnitudes of chromosome-wide LD of TBN and HAN are very similar (supplementary fig. S5, Supplementary Material online). However, TBN showed an increased LD (supplementary figs. S6 and S7, Supplementary Material online) as well as reduced haplotype diversity (supplementary fig. S8, Supplementary Material online) in the regions encompassing *EPAS1* and *EGLN1*, consistent with the presence of a dominating haplotype in each of the genes. The extended haplotype homozygosity (EHH) analysis showed significant signal in *EGLN1* gene region ($P < 10^{-5}$) but did not yield significant indication of positive selection in *EPAS1* gene region (see Materials and Methods, supplementary figs. S9 and S10, Supplementary Material online). This result is consistent with that of Simonson et al.. However, this does not necessarily preclude the role of *EPAS1* in HAA if the action of selection initiated long time ago (Pickrell et al. 2009). The long history (>21 ka) of Tibetan population (Shi et al. 2008; Zhao et al. 2009) and selection initiated long time ago may compromise the power of commonly used methods, such as EHH in detecting the signatures of natural selection (Sabeti et al. 2002, 2007), as we observed in our data (supplementary fig. S10, Supplementary Material online). This is also one possible reason besides the data quality per se that Simonson et al. based on EHH approaches failed to identify *EPAS1*, the top candidate gene in our study, Yi et al., and Beall et al. The lack of significantly increased LD in TBN compared with that of the other populations suggests antiquity of the natural selection on the two genes because LD decays with time, whereas the increased LD within both genes indicate the selection is long-term persistent and ongoing. Furthermore, LD decays quickly and haplotype homozygosity could not extend to long range in case natural selection occurred in genes with recombination hotspots. For example, according to the estimation based on HapMap data (McVean et al. 2004), there are 11 hotspots in the genomic region where *EPAS1* located with 200 kb extended on both sides, that is, 2.4 hotspots/100 kb, which is much higher than the average hotspot density across the chromosome 2 (1.1 hotspots/100 kb). In contrast, there are only 6 hotspots in the genomic region where *EGLN1* located region with 200 kb extended on both side, that is, 1.3/100 kb, which is just slightly higher than the average hotspot density in chromosome 1 (1.1 hotspots/100 kb). These estimations from the HapMap populations could apply to Tibetan population because we observed the comparable LD magnitudes across the entire chromosome between Tibetan and Han Chinese populations (supplementary fig. S5, Supplementary Material online). Therefore, the power of EHH-based methods could be compromised in case of *EPAS1* because the natural selection initiated long time ago considering the long history (>21 ka) of Tibetan population (Shi et al. 2008; Zhao et al. 2009) and recombination hotspots in the genomic regions where *EPAS1* located.

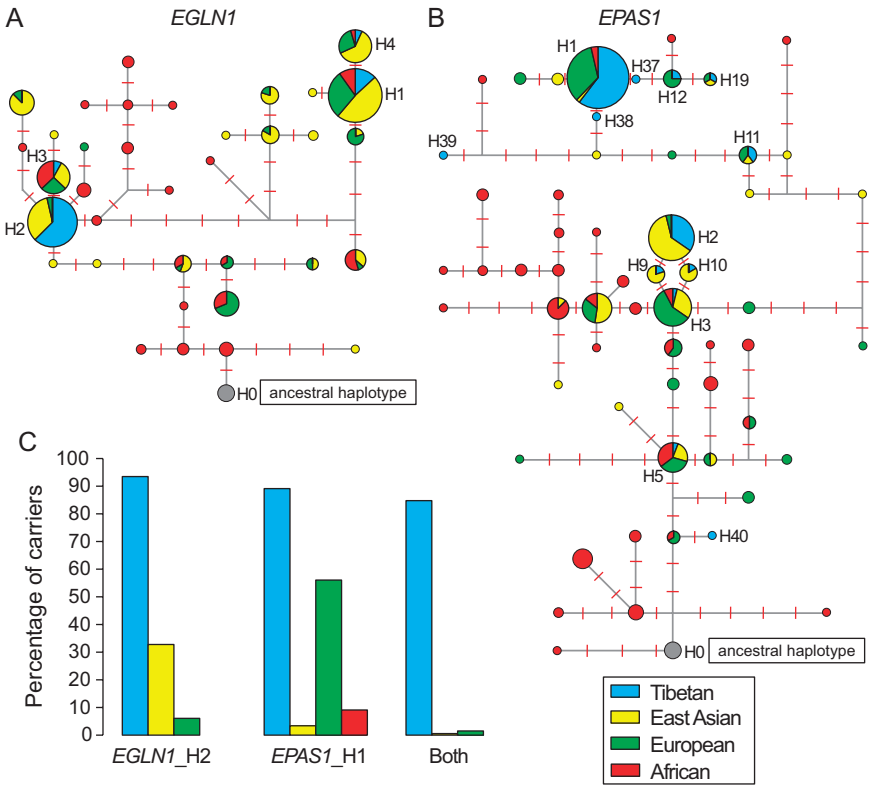

Fig. 3. Network of haplotypes in the two genes and distribution of haplotype carrier in Tibetan and worldwide populations. (*A*) Network of *EGLN1* haplotypes which were reconstructed based on 18 SNPs in a 55 kb "haplotype block" region, the entire gene region is about 59 kb. (*B*) Network of *EPAS1* haplotypes which were reconstructed based on 20 SNPs in a 58 kb haplotype block region, the entire gene region is about 89 kb. (*C*) Comparison of the percentages of the Tibetan-dominant haplotype carriers among worldwide populations. *EGLN1*_H2 is the H2 haplotype showing in (*A*), *EPAS1*_H1 is the H1 haplotype showing in (*B*).

Haplotype Networks and TBN-Dominant Haplotypes

A haplotype network can reveal evolutionary relationship and frequency distribution of haplotypes among populations. Haplotype networks of *EGLN1* (fig. 3*A*) and *EPAS1* (fig. 3*B*) were reconstructed based on SNPs in the most differential genomic regions (SNP information shown in supplementary table S6, Supplementary Material online). Each gene has one haplotype overrepresented in Tibetans (see also supplementary fig. S8, Supplementary Material online) (93.5% for haplotype H2 in *EGLN1* and 89.1% for haplotype H1 in *EPAS1*), and both are significantly less prevalent in African (0% for *EGLN1* and 9.1% for *EPAS1*), European (6.1% for *EGLN1* and 56.1% for *EPAS1*), and other East Asian populations (32.8% for *EGLN1* and 3.4% for *EPAS1*). A strong correlation of the altitude with the frequency of H2 for *EGLN1* ($R^2 = 0.60$, $P = 1.53 \times 10^{-4}$) and with the frequency of H1 for *EPAS1* ($R^2 = 0.92$, $P = 5.69 \times 10^{-6}$) in East Asian populations (fig. 4) were observed, respectively, suggesting the possible roles of these two haplotypes in HAA in TBN. Because these dominating haplotypes derived from them were also found in African and the other worldwide populations (fig. 3), they are most likely old; but it could be also the case that the TBN-specific variation was not identified in our data.

A "Dominant Haplotype Carrier" Model

The observation of significant overrepresentation of the carriers of the dominating haplotypes of *EGLN1* and *EPAS1* (fig. 3*C*), respectively, in Tibetans led to a dominant haplotype carrier model on a possible role of the two genes underlying adaptation to altitude. In other words, a copy of one of the haplotypes would render a better chance to survive. Furthermore, following the lead of the IPA analysis which indicated an interaction of *EGLN1* and *EPAS1*, we observed a significantly higher proportion of the carriers

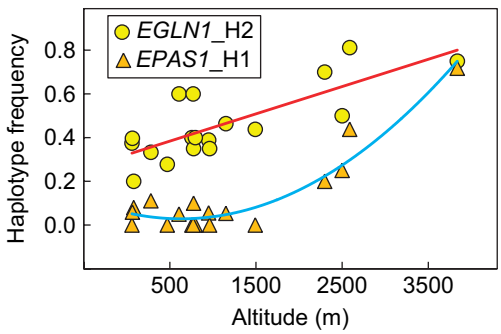

Fig. 4. Correlation of haplotype frequency in East Asian populations and altitude. R^2 for the regression of haplotype frequency on altitude is 0.60 ($P = 1.53 \times 10^{-4}$) and 0.92 ($P = 5.69 \times 10^{-6}$) for *EGLN1* haplotype (H2) and *EPAS1* haplotype (H1), respectively.

of both dominating haplotypes in TBN (84.8%) than that in African (0%), European (1.5%), and other East Asian (0.6%) populations (Fisher exact test, $P < 10^{-21}$) (fig. 3C), respectively. Several tests further showed that the carriers of both dominating haplotypes in TBN are overrepresented when compared with the expectation assuming independent contribution of the two haplotypes (Fisher exact test, $P = 10^{-4} \sim 10^{-10}$) (supplementary table S7, Supplementary Material online). A strong correlation between the proportion of such carriers and the altitude in East Asian populations ($R^2 = 0.88$, $P = 4.88 \times 10^{-6}$) (supplementary fig. S11, Supplementary Material online) further supports a possible interaction of the two haplotypes in HAA in Tibetan and other high-altitude Asian populations.

Discussion

In this study, we showed a possible involvement of two hypoxia-related genes, *EGLN1* and *EPAS1*, in HAA in Tibetans. Interestingly, one each but not both of the two genes were also identified by three recent studies published in "Science" (Simonson et al. 2010; Yi et al. 2010) and "PNAS" (Beall et al. 2010), hereafters refer to Simonson et al., Yi et al., and Beall et al., respectively. We reanalyzed the available GW genotype data from Simonson et al. and explored why the top signal was missed in its GW scan; we also compared and discussed the possible reasons why Yi et al. and Beall et al. missed the top second signal. Notably, we found *EPAS1*, which ranked the top signal in Beall et al., Yi et al., and our present study but escaped in Simonson et al. can be recaptured with carefully reprocessed data and reanalysis (supplementary fig. S12*C*–*F*, Supplementary Material online). In fact, *EPAS1* is ranking the top signal in the data of Simonson et al. no matter whether the comparison is between Tibetan samples collected from Qinghai (TBN-QH, which were originally used in Simonson et al.) and HapMap East Asian samples (CHB + JPT) (supplementary fig. S12*C*–*D*, Supplementary Material online), or it is between TBN-QH and HAN in our present study (supplementary fig. S12*E*–*F*, Supplementary Material online).

Since no significant difference of Tibetan samples was observed between Simonson et al. and our study according to the PCA based on GW data (supplementary fig. S13, Supplementary Material online), we explored the possible reasons that could have affected the local estimation, including: 1) genotyping quality, 2) batch effect, strand (±) switch or other annotation problems in public data integration, and 3) statistical methods adopted for detecting natural selection signatures.

First, we noted there were some obvious signatures of poor genotype data quality. Simonson et al. did not provide any description on the procedures of data filtration and quality controls conducted, but there are 22,515 (2.53%) autosomal SNPs missing in the released data compared with the total 890,661 nonredundant autosomal SNPs in Affymetrix 6.0. Missing such a large proportion of markers seemed not due to filtration of monomorphic markers because there are still 180,713 (20.8%) SNPs with single allele presenting in the data. Data integration would not either loss many markers because almost all the Affymetrix markers are present in HapMap data. In addition, there are still 703 SNPs with missing data >50% and 141 SNPs without any genotypes (100% missing data) for the remaining 868,146 SNPs (supplementary table S8, Supplementary Material online); there are 486 SNPs in slight Hardy–Weinberg disequilibrium (HWD) ($P < 10^{-3}$) and 22 SNPs in significant HWD ($P < 10^{-6}$) (supplementary table S9, Supplementary Material online). These problems could have been resulted from the genotyping per se, genotype calling, or both but could not be fully evaluated due to lack of the raw intensity data.

Secondly, there are many problems seem to be associated with SNP genotype annotation of Simonson et al. It is obvious that Simonson et al. used a very old version of annotation file, there are 454 SNPs with very old rs# and 1,019 SNPs with rs# being merged to new ones in dbSNP130, suggesting that an old version of annotation file was used (supplementary table S10, Supplementary Material online). In addition, the rs# of additional 127 SNPs could not be found in Affymetrix 6.0 annotation files (supplementary table S11, Supplementary Material online), neither in an earlier version (Release 27, 25 November 2008) nor in the latest version (Release 30, 15 November 2009). Surprisingly, there are 48 SNPs which are obviously belonging to horse (*Equus caballus*) genome (supplementary table S12, Supplementary Material online). Furthermore, it is not clear whether the data have been successfully integrated with HapMap data because most of the genotypes did not follow the Affymetrix strand or the HapMap strand.

With the help of the GW data, we generated in our own Tibetan samples (TBN-TB; no significant difference was observed between TBN-QH and TBN-TB according to the PCA based on GW data, see also supplementary fig. S13, Supplementary Material online) using the same technology, we were able to correct the annotation in Simonson et al. using the latest Affymetrix annotation file. A clean data set with 859,252 SNPs were used for further analysis, after removing 8,894 SNPs that show high degree of missing data (less than two successful genotypes), significant deviations from HWE ($P < 10^{-6}$), or uncertain allelic strand of markers with allele status of A/T or G/C. In our further analysis, we randomly selected 31 CHB + JPT samples from HapMap data to match the number of Tibetan samples typed by Simonson et al. We also included 31 Han Chinese samples (HAN) from our own Affymetrix 6.0 data (Xu et al. 2009) for further comparison study.

Thirdly, even though the overall F_{ST} values, using reprocessed data, between TBN-QH and CHB (0.01140 ± 0.00005) or HAN (0.01154 ± 0.00005) are comparable with that between TBN-TB and HAN (0.01132 ± 0.00004) (table 1), the variance of the estimation of F_{ST} among loci is still high based on the reprocessed data of Simonson et al. In particular, the SD of F_{ST} among loci between TBN-QH and CHB + JPT (SD = 0.0299) is much higher than that (SD = 0.0228) between TBN-TB and HAN ($P < 10^{-33}$, see also supplementary fig. S12, Supplementary Material online),

Table 1. Pairwise F_{ST} between Populations.

	TBN-TB	TBN-QH	HAN	CHB
TBN-QH	0.00226 ± 0.00003			
HAN	0.01132 ± 0.00004	0.01154 ± 0.00005		
CHB	0.01135 ± 0.00004	0.01140 ± 0.00005	0.00031 ± 0.00001	
JPT	0.01744 ± 0.00005	0.01728 ± 0.00005	0.00715 ± 0.00002	0.00691 ± 0.00002

NOTE.—TBN-TB, Tibetan samples collected from Tibet (this study); TBN-QH, Tibetan samples collected from Qinghai (Simonson et al. 2010). SD of F_{ST} over loci was calculated by bootstrap resampling with 1,000 replications.

indicating a possible batch effect due to the data on TBN-QH and that on CHB + JPT were generated separately in different laboratories, which highlights the importance of researchers generating their own control data.

We performed GW scan in this data set based on allele frequency comparison between TBN-QH and lowland East Asian populations (HapMap CHB + JPT and our HAN samples). Interestingly, we recaptured the escaped top signal in this data set (supplementary fig. S12*C*–*F*, Supplementary Material online), across the genome, we observed the highest F_{ST} in *EPAS1* gene region. In contrast, as shown in Simonson et al., the observed F_{ST} was less than 0.5 or even less than 0.4 for most SNPs at *EPAS1* gene (supplementary fig. S14*A*, Supplementary Material online). With our reanalysis, a considerable number of SNPs in this gene showed high F_{ST} values above 0.5 (supplementary fig. S14**B** and S15*A*–*B*, Supplementary Material online), which are comparable with the results obtained from TBN-TB and HAN (supplementary fig. S14*C*, Supplementary Material online). These results indicate that the two data sets are now comparable and data quality is not a serious issue with our data reprocessing and reanalysis.

Although we also expected the population structure (or allele frequency difference) in Tokyo Japanese (JPT) and Beijing Han Chinese (CHB) would have affected the results of Simonson et al., our reanalyses using either combined (JPT + CHB) samples or single population (JPT or CHB) samples did not show significant difference in detecting the top GW signals (e.g., *EPAS1* and *EGLN1*). Because the allele difference between JPT and CHB is very much smaller compared with that between JPT (or CHB) and Tibetans. We did not either observe the results were affected by the known population substructure of Han Chinese, although the differentiation of northern and southern Han Chinese is distinct in GW data (Xu et al. 2009), the allele frequency difference is much smaller compared with that between Han Chinese and Tibetans.

Lastly, we further explored the possible statistical reasons that led to the missing of *EPAS1* in the GW scan of Simonson et al. *EGLN1* was identified as a candidate for adaptation in Tibetan in Simonson et al. because it showed the highest XP-EHH and integrated haplotype score scores in EHH-based analyses (Sabeti et al. 2002, 2007) among a set of a priori functional candidate loci and its known hypoxia-related functions. We made a similar observation (supplementary fig. S10, Supplementary Material online). No significant EHH signals implicated *EPAS1* in the data, but this does not necessarily preclude the role of *EPAS1* in HAA if the action of selection initiated long time ago (Pickrell et al. 2009) which may compromise the power of EHH-based methods in detecting the signatures of natural selection (Sabeti et al. 2002, 2007). This constitutes another reason, in addition to data quality, that Simonson et al. failed to identify *EPAS1* in their GW scan. Furthermore, LD decays quickly and haplotype homozygosity could not extend to long range in case natural selection occurred in genes near recombination hotspots, as we discussed earlier in Results.

With regard to the study of Beall et al., since the GW data are not available, it is not feasible to make a full evaluation of the issues which we did for Simonson et al. However, the obviously larger fluctuation of the GW signals (supplementary fig. S16, Supplementary Material online) which resembles that of Simonson et al. compared with our study indicates there might be similar problems of public data integration or batch effect because Beall et al. also used HapMap data as control. We further compared the Tibetan samples in three studies (Beall et al. 2010; Simonson et al. 2010; present study) based on allele frequencies of SNPs showing highly differentiated alleles between the Tibetan samples collected from Yunnan (TBN-YN) and CHB which are the only available data from Beall et al. so far. The information of 123 shared SNPs and allele frequencies are shown in supplementary table S13 (Supplementary Material online). We observed an obvious deviation of allele frequency in TBN-YN from that in both TBN-TB and TBN-QH (supplementary fig. S17, Supplementary Material online). This result suggests either difference of Tibetan samples (TBN-YN) in Beall et al. or the data quality problems or batch effect caused in public data integration. Another reason that Beall et al. missed *EGLN1*, the top signal in Simonson et al. and the top second signal in our study, could be that different chip used in Beall et al. where some of the top SNPs detected using Affymetrix 6.0 are not present in the Illumina 610-Quad chip.

To summarize, the exclusion of the role of *EPAS1* in HAA in Tibetans in Simonson et al. was most likely due to annotation problems in public data integration as well as the compromised power of EHH-based methods in detecting the signature of ancient adaptation events occurred in genomic regions with dense recombination hotspots; the missing of *EGLN1* in Beall et al. could be due to the large variation of the statistical estimations among loci caused by batch effect in public data integration or the different chip used.

Base on this experience, we suggest: 1) more attention should also be paid to careful analysis of the raw data and upstream population genetic analysis in a GW study, although downstream function studies are currently more

appealed and emphasized. GW data are so valuable, and high accuracy of the analysis is high priority. It is here, with a huge investment, that the basic and ultimate information for further study could be extracted. Careful analysis of GW data per se should not be compromised for a priori functional candidate loci, otherwise the results could be misleading. 2) Considering the much larger variation of locus-specific estimation using public data in both Simonson et al. and Beall et al., we suggest it is necessary that researcher to generate their own control data if batch effects and platform issues could not be confidently excluded in integrating public data. 3) Both haplotype-based and allelic frequency-based methods should be performed in detecting natural selection signatures in GW data, haplotype-based methods are not always powerful especially in case the natural selection initiated long time ago and particularly in genomic regions with recombination hotspots.

The study of Yi et al. was based on exome-sequencing technology and it captured the DNA sequence information mainly in coding regions of the genome. Because almost all the significant signals are from intronic regions where the coverage of the data of Yi et al. is very low, it is no wonder they failed to identify *EGLN1*. As a matter of fact, even for *EPAS1* which they identified as the top candidate gene under natural selection, the strongest (also the only two) signals were from two intronic SNPs.

We also noted Yi et al. estimated Tibetans and Han Chinese diverged 2,750 years ago. On the one hand, we argue this divergence time could be underestimated considering it was based on exome data and a model with several unapproved assumptions of population growth and migration history. On the other hand, we emphasize that this estimate of divergence time of populations should not be taken as the time estimation of genetic history of Tibetans. Recent genetic studies based on Y chromosomal and mitochondrial DNA data suggested a long history (>21 ka) of Tibetan population and its ancestral lineages (Shi et al. 2008; Zhao et al. 2009). There are also evidence from archaeological studies supporting that the Tibetans occupied the Plateau 20,000 years ago (Zhang et al. 2003).

The antiquity of the dominating haplotypes in *EGLN1* and *EPAS1* as revealed in network analysis suggests that they may not be selected for until they reached a high-altitude environment. We expect that the selection pressure of hypoxia in high-altitude area has not relaxed since the first group of people arrived. The antiquity of the haplotypes and selection initiated long time ago in addition to the recombination hotspots may compromise the power of commonly used methods, such as EHH in detecting the signatures of natural selection (Sabeti et al. 2002, 2007), as we observed in our data (supplementary fig. S10, Supplementary Material online). This is also one possible reason that Simonson et al. based on EHH approaches failed to identify *EPAS1*, the top candidate gene in Yi et al., Beall et al., and our present study. On the other hand, the presence of the dominating haplotypes in lowlanders could explain the survival of a large number of immigrants who migrated to the Plateau subsequently. A systematic analysis on the proportion of dominating haplotypes in immigrants and their descendents in Tibet is therefore warranted to further examine the dominant haplotype carrier model.

None of the four GW studies observed a strong signal in *HIF1A* gene region, although this gene is often considered as the "master regulator" of oxygen homeostasis (Semenza 2004) and was implicated in HAA in some other populations, such as Sherpa (Beall 2007; Stobdan et al. 2008). In the Tibetan samples studied, the average F_{ST} in *HIF1A* region is only 0.022 (±0.032), most of SNPs in *HIF1A* gene region fall outside of the top 1% high differentiated SNPs ($F_{ST} > 0.1$), only three of 76 SNPs with F_{ST} slightly higher than 0.1 (supplementary fig. S18, Supplementary Material online). However, an association of *HIF1A* with altitude acclimatization in Han Chinese was observed (Zhuoma B, personal communication of unpublished data).

Physiological responses to high-altitude environment are complex and involve a range of mechanisms such as the expression/activation of an array of genes redirecting the metabolic and other cellular mechanisms in a highly coordinated manner to achieve enhanced cell survival under hypoxic environment (Sarkar et al. 2003). Therefore, it may not be surprising that numerous genes are involved in the response to high-altitude environments. The two genes implicated in the current study could be a tip of iceberg. Because our study based on between-population comparison can only detect the most significant signals with confidence, to identify a more comprehensive list of genes, further work is needed with large-scale sampling and within-population study designs. However, a further study on the role of *EGLN1* and *EPAS1* in altitude acclamation (physiological responses) may shed light on the understanding of its connection with altitude adaptation (evolutionary survival).

Supplementary Material

Supplementary figures S1–S18 and tables S1–S13 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Dr Hua Chen for providing us the XP-CLR source code and assistance in running his software. This work was supported by the MoST International Cooperation Base of China. S.X. was supported by the National Science Foundation of China (30971577) and the Science and Technology Commission of Shanghai Municipality (09ZR1436400). L.J. was supported by grants from the National Outstanding Youth Science Foundation of China (30625016), the National Science Foundation of China (30890034), the Shanghai Leading Academic Discipline Project (B111), and the Center for Evolutionary Biology. S.X. also gratefully acknowledges the support of SA-SIBS Scholarship Program and K.C. Wong Education Foundation, Hong Kong.

References

Akey JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* 19:711–722.

Bandelt HJ, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* 141:743–753.

Beall CM. 2007. Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *Proc Natl Acad Sci U S A*. 104(Suppl 1):8655–8660.

Beall CM, Cavalleri GL, Deng L, et al. (29 co-authors). 2010. Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. *Proc Natl Acad Sci U S A*. 107:11459–11464.

Beall CM, Decker MJ, Brittenham GM, Kushner I, Gebremedhin A, Strohl KP. 2002. An Ethiopian pattern of human adaptation to high-altitude hypoxia. *Proc Natl Acad Sci U S A*. 99:17215–17218.

Cann HM, de Toma C, Cazes L, Legrand MF, et al. (41 co-authors). 2002. A human genome diversity cell line panel. *Science* 296:261–262.

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Res*. 20:393–402.

Frazer KA, Ballinger DG, Cox DR, et al. (40 co-authors). 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.

Hill WG, Weir BS. 1994. Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am J Hum Genet*. 54:705–714.

Korn JM, Kuruvilla FG, McCarroll SA, et al. (16 co-authors). 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*. 40:1253–1260.

Li JZ, Absher DM, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–584.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. 2:e190.

Pickrell JK, Coop G, Novembre J, et al. (11 co-authors). 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res*. 19:826–837.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.

Sabeti PC, Varilly P, Fry B, et al. (249 co-authors). 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–918.

Sarkar S, Banerjee PK, Selvamurthy W. 2003. High altitude hypoxia: an intricate interplay of oxygen responsive macroevents and micromolecules. *Mol Cell Biochem*. 253:287–305.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*. 78:629–644.

Semenza GL. 2004. Hydroxylation of HIF-1: oxygen sensing at the molecular level. *Physiology (Bethesda)*. 19:176–182.

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.

Smith TG, Robbins PA, Ratcliffe PJ. 2008. The human side of hypoxia-inducible factor. *Br J Haematol*. 141:325–334.

Stobdan T, Karar J, Pasha MA. 2008. High altitude adaptation: genetic perspectives. *High Alt Med Biol*. 9:140–147.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370.

Xu S, Yin X, Li S, et al. (24 co-authors). 2009. Genomic dissection of population substructure of Han Chinese and its implication in association studies. *Am J Hum Genet*. 85:762–774.

Yi X, Liang Y, Huerta-Sanchez E, et al. (65 co-authors). 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329:75–78.

Zhang DD, Li SH, He YQ, Li BS. 2003. Human settlement of the last glaciation on the Tibetan plateau. *Curr Sci*. 84:701–704.

Zhao M, Kong QP, Wang HW, et al. (14 co-authors). 2009. Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. *Proc Natl Acad Sci U S A*. 106:21230–21235.