# Genetic signatures of high-altitude adaptation in Tibetans

Jian Yang[a,b,1,2], Zi-Bing Jin[b,1,2], Jie Chen[b], Xiu-Feng Huang[b], Xiao-Man Li[b], Yuan-Bo Liang[b], Jian-Yang Mao[b], Xin Chen[b], Zhili Zheng[a,b], Andrew Bakshi[a], Dong-Dong Zheng[b], Mei-Qin Zheng[b], Naomi R. Wray[a], Peter M. Visscher[a], Fan Lu[b,2], and Jia Qu[b,2]

[a]Institute for Molecular Bioscience, Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia; and [b]The Eye Hospital, School of Ophthalmology & Optometry, Wenzhou Medical University, China National Engineering Research Center of Ophthalmology and Optometry, State Key Laboratory Cultivation Base and Key Laboratory of Vision Science, Ministry of Health, Wenzhou 325027, China

Edited by Marcus W. Feldman, Stanford University, Stanford, CA, and approved February 13, 2017 (received for review October 14, 2016)

**Indigenous Tibetan people have lived on the Tibetan Plateau for millennia. There is a long-standing question about the genetic basis of high-altitude adaptation in Tibetans. We conduct a genome-wide study of 7.3 million genotyped and imputed SNPs of 3,008 Tibetans and 7,287 non-Tibetan individuals of Eastern Asian ancestry. Using this large dataset, we detect signals of high-altitude adaptation at nine genomic loci, of which seven are unique. The alleles under natural selection at two of these loci [methylenetetrahydrofolate reductase (*MTHFR*) and *EPAS1*] are strongly associated with blood-related phenotypes, such as hemoglobin, homocysteine, and folate in Tibetans. The folate-increasing allele of rs1801133 at the *MTHFR* locus has an increased frequency in Tibetans more than expected under a drift model, which is probably a consequence of adaptation to high UV radiation. These findings provide important insights into understanding the genomic consequences of high-altitude adaptation in Tibetans.**

high-altitude adaptation | Tibetans | genome-wide association study | mixed linear model | polygenic selection

Genetic adaptation to a novel environment is a fundamental process for the survival and adaptation of a species. In humans, one of the most recent examples is adaptation to high altitude, such as the Tibetan highlands. The Tibetan Plateau (TP; also known as the Qinghai–Tibet Plateau in China) has an average elevation of ∼4,000 m above sea level, where the oxygen concentration is ∼40% lower (1) and UV radiation is ∼30% stronger (2) than at sea level. The indigenous Tibetan people have developed a distinctive set of physiological characteristics to adapt to the extreme environmental conditions in the highlands (1). Previous population-based genetic studies have reported evidence that genetic variants at the *EPAS1* and *EGLN1* loci have been under positive natural selection (3–7). These genetic variants are associated with phenotypic variation of hemoglobin concentration (HGB) in Tibetans (3–5). The *EPAS1* gene, which encodes the hypoxia inducible factor-2α (HIF-2α) subunit of HIF complex, is a transcription factor involved in body response to hypoxia (8, 9). *EGLN1* encodes PHD2, which is a major oxygen-dependent negative regulator of HIFs (10, 11). Apart from these two known genes that have biological relevance to hypoxia adaptation (3–7, 12), several other candidate gene loci (e.g., *PPARA* and *HBB*) have been highlighted in recent studies (3, 4, 13–15). Genetic adaptation to high altitude, however, is likely to be a complex process, with a large number of genes involved in response to not only hypoxia but also, other extreme environmental conditions, such as low temperature, high UV radiation, and insufficient food supply. If the strength of natural selection at these gene loci has been small to moderate, these loci would not be detected in previous studies (3–7) of small sample size (typically $n < 150$). In this study, we perform a large-scale genome-wide study to detect genetic signals of high-altitude adaptation in 3,008 Tibetans and 7,287 non-Tibetan individuals of Eastern Asian (EAS) ancestry. Using this large dataset, we identify signals of genetic adaptation.

## Results

**Genetic Ancestry of Tibetans.** There were 3,717 subjects collected from two sites (Seda and Litang) in the TP in China (*SI Appendix*, Fig. S1). We extracted DNA from blood samples and performed genome-wide SNP genotyping assays using the Illumina CoreExome array, an SNP array with 264,909 tag SNPs with genome-wide coverage and 244,593 exome-focused SNPs (*Materials and Methods*). After standard quality control (QC) filtering of the genotype data, we retained 3,381 subjects and 287,691 SNPs (279,608 on autosomes), most of which were genome-wide tag SNPs.

We performed a principal component analysis (PCA) of the subjects using all 279,608 autosomal SNPs after stringent QCs (*Materials and Methods*). There was no evidence of population stratification between the cohorts recruited from the two sites (*SI Appendix*, Fig. S2*A*), despite the fact that the Seda subjects were recruited from people who came from diverse regions of the TP to study or work at the Seda Larong Wuming Buddhist Institute and the Litang subjects were recruited from nomadic people who have lived in Litang and surrounding areas for many generations (*Materials and Methods*). We, therefore, combined the two cohorts for analysis. We showed by projecting the principal components (PCs) estimated from our samples on those from the 1000 Genome Projects (1000G) that all of our subjects were of EAS

### Significance

The origin of Tibetans and the mechanism of how they adapted to the high-altitude environment remain mostly unknown. We conduct the largest genome-wide study in Tibetans to date. We detect signatures of natural selection at nine gene loci, two of which are strongly associated with blood phenotypes in present day Tibetans. We further show the genetic relatedness of Tibetans with other ethnic groups in China and estimate the divergence time between Tibetans and Han. These findings provide important knowledge to understand the genetic ancestry of Tibetans and the genetic basis of high-altitude adaptation.

Author contributions: J.Y., Z.-B.J., F.L., and J.Q. designed research; Z.-B.J., J.C., X.-F.H., X.-M.L., Y.-B.L., J.-Y.M., X.C., D.-D.Z., M.-Q.Z., F.L., and J.Q. performed research; J.Y., Z.Z., A.B., N.R.W., and P.M.V. analyzed data; F.L. and J.Q. jointly supervised the study; and J.Y. and Z.-B.J. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The summary-level statistics from all of the mixed linear model-based leave one chromosome out association analyses reported in the paper are available at cnsgenomics.com/data/yang_et_al_2017_pnas.html. The raw genotype and phenotype data of the Tibetan and Han subjects are available through application at https://www.wmubiobank.org.

[1]J.Y. and Z.-B.J. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: jqu@mail.eye.ac.cn, lufan@mail.eye.ac.cn, jinzb@mail.eye.ac.cn, or jian.yang@uq.edu.au.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1617042114/-/DCSupplemental.

ancestry (*SI Appendix*, Fig. S2*B*). On a finer scale, the subjects are stratified along the first PC (*SI Appendix*, Fig. S2*C*), consistent with a few hundred self-reported Han in the sample. We classified our subjects into three groups (Tibetans, Han, and possibly admixed) (*Materials and Methods* and *SI Appendix*, Fig. S2*D*) and removed the possibly admixed subjects. There were 3,008 Tibetans and 373 Han retained for analysis.

We projected the PCs of our subjects on the Chinese subjects from the Human Diversity Genome Project (HGDP) (16) and illustrated the genetic relatedness between Tibetans and other ethnic groups in China (Fig. 1*A*). Our result suggests that Tibetans show the nearest genetic relatedness to Yi, Tu, and Naxi ethnic minority populations (Fig. 1*A* and *SI Appendix*, Table S1), consistent with these populations who reside in the neighboring regions of the TP (Yi and Naxi people are mainly distributed in Yunnan and Sichuan provinces, and most Tu people reside in Qinghai province) (Fig. 1*B*).

We estimated the divergence time between Tibetan and Han populations using the conventional $F_{ST}$-based approach (17) (*SI Appendix*, Text S1). As described above, there were 3,008 Tibetan and 373 Han subjects collected from the TP after QC. We included in this analysis an additional set of 1,726 Han subjects collected from the Eye Hospital of Wenzhou Medical University (WZ) after QC (*Materials and Methods*). We used GCTA-GRM to remove cryptic relatedness in the Tibetan and Han samples (note that the Han sample was a combined set of 373 Han subjects from the TP and 1,726 Han subjects from WZ) at a relatedness threshold of 0.05 and retained 1,998 unrelated Tibetan and 2,059 unrelated Han subjects. There was no genetic difference between WZ-Han and TP-Han as shown by PCA (*SI Appendix*, Fig. S3), probably because most of the Han subjects, collected from either TP or WZ, were originally from diverse regions of China. The genome-wide mean $F_{ST}$ between Tibetans and Han was 0.012 [using the method by Weir and Cockerham (18) implemented in GCTA], consistent with the estimate of the Han subjects from the HGDP (*SI Appendix*, Table S1). Given the genome-wide mean $F_{ST}$ value (*Materials and Methods*), we estimated that the divergence time between Tibetan and Han populations was 189 generations. Assuming an average generation time of 25 y as in previous studies (3, 19), this estimate suggests that Tibetans and Han split about 4,725 y ago, ~2,000 y earlier than that estimated from whole-exome sequencing data (3) but consistent with recent evidence from archeological studies (20, 21).

**Genome-Wide Analysis to Detect Genetic Signals of Adaptation.** To detect genetic signals of high-altitude adaptation, we used a mixed linear model-based leave one chromosome out association (MLMA-LOCO) analysis approach [implemented in the BOLT-LMM software tool (22)] to test for allele frequency difference between Tibetans and non-Tibetans of EAS ancestry (*Materials and Methods*). We investigated the statistical properties of the method using simulations (*SI Appendix*, Table S2). Similar approaches have been used in genome-wide association studies (GWASs) to control for population structure (22, 23). In the MLMA-LOCO model, the target SNP to be tested is fitted as a fixed effect, and all SNPs on the other chromosomes are fitted as random effects (details about the model are in *Materials and Methods*). The underlying assumption is that, under a drift model, the random effects follow a normal distribution with the variance being proportional to $p_0(1-p_0)F_{ST}$, where $p_0$ is the allele frequency in the ancestral population and $F_{ST}$ is the Wright's fixation index between the two derived populations (24). If there are two diverged populations in the sample, even if neither of the populations have been under natural selection, SNPs on different chromosomes will be correlated because of the systematic difference in allele frequency between populations caused by cryptic relatedness in the samples, genetic drift, and/or possibly, admixture with other populations (see below for examples). We, therefore, can correct for the interchromosome correlations by modeling all of the SNPs on the other chromosomes (as random effects, because number of SNPs is usually larger than sample size) when testing for the association of an SNP. To maximize power, we included in the analysis all of the subjects collected from the TP and WZ in China (3,008 Tibetans and 2,099 Han) and an additional set of 5,188 subjects of EAS ancestry from the Genetic Epidemiology Research on Aging (GERA) Study (25) in the United States (*Materials and Methods*). Because the GERA-EAS subjects were genotyped on a different SNP array (Affymetrix Axiom), we imputed all of the genotype data

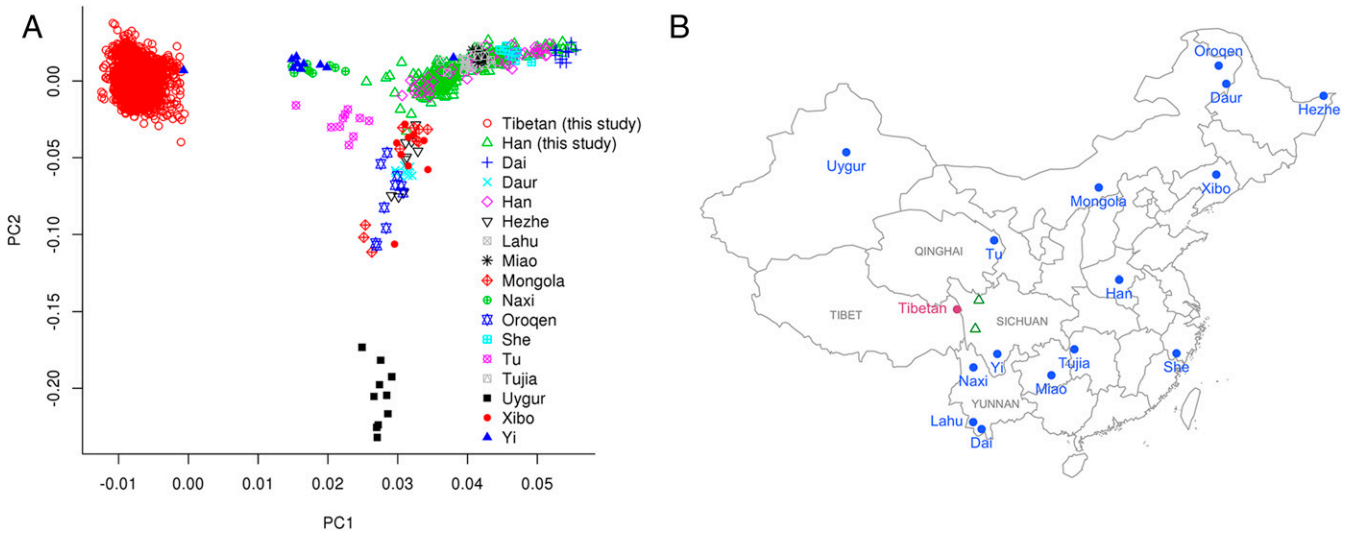


**Fig. 1.** PCA of genetic ancestry in Chinese populations using genome-wide SNP data. (*A*) Result from a PCA in a combined sample of 3,381 genetically confirmed Tibetan and Han from this study and 180 Chinese subjects (multiple ethnic groups) from the HGDP. PC1 and PC2 represent the first two eigenvectors from PCA. Note that one of the Yi subjects from the HGDP seems to be of Tibetan ancestry. (*B*) Distribution of the ethnic groups in China. The blue circles represent the main distribution areas of the ethnic populations in the HGDP, and the red circle represents the Tibetan population. Note that many of the populations, such as Han, Mongola, Tibetan, and Uygur, are distributed widely in a range of regions rather than the specific areas labeled on the map. The green triangles represent the two areas (Seda and Litang) from which our Tibetan subjects were recruited (*SI Appendix*, Fig. S1).

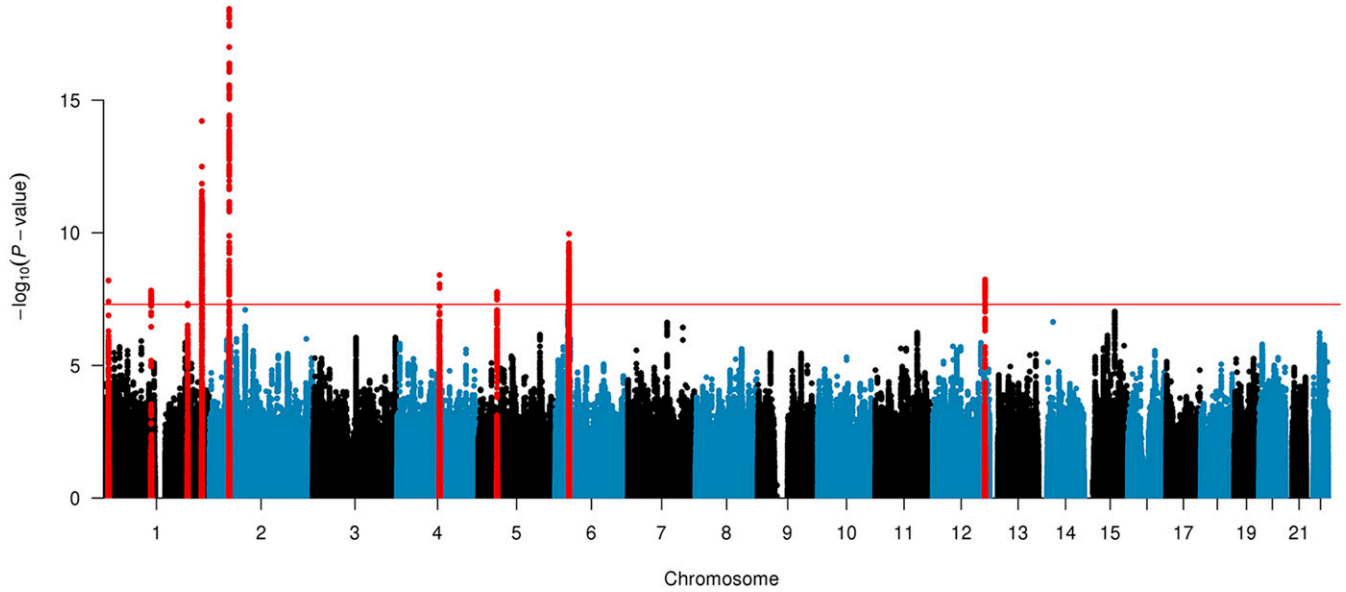


**Fig. 2.** Genome-wide scan for genetic signatures of adaptation. Shown on the *y* axis are −log10 of *P* values from the tests of allele frequency difference between Tibetan Chinese ($n$ = 3,008) and EASs ($n$ = 7,287). The analysis was performed using the MLMA-LOCO method, which tests for difference in allele frequency between populations taking into account the difference caused by random drift. SNPs at the genome-wide significant loci are highlighted in red.

to 1000G reference panels using IMPUTE2 (26). There were three ancestry outliers, which were excluded from analysis (*SI Appendix*, Fig. S4). To exclude SNPs with allele frequency differences between cohorts caused by potential batch effects, we performed a "control–control" analysis using the MLMA-LOCO approach to test for difference in allele frequency between TP-Han and a combined set of WZ-Han and GERA-EAS and removed SNPs with $P$ value < $1 \times 10^{-6}$. We then performed a "case–control" analysis using the MLMA-LOCO approach to test for difference in allele frequency between Tibetans ("cases": $n$ = 3,008) and EAS subjects ("controls": TP-Han, WZ-Han, and GERA-EAS; $n$ = 7,287) and identified nine loci that passed the genome-wide significance level ($P_{\text{MLMA-LOCO}}$ < 5e-8) (Fig. 2 and *SI Appendix*, Fig. S5). Of nine loci, two loci, *EPAS1* and *EGLN1*, which show the strongest signals in our analysis, are known (3–7), and the other seven loci are unique (Table 1 and *SI Appendix*, Fig. S6). Note that *FGF10* was one of a set of genes that showed large population branch statistic (PBS) values (Tibetans vs. Han vs. Europeans) in a recent study (15). We show by linkage disequilibrium (LD) score regression analysis (*SI Appendix*, Text S2) that there is no inflation in the test statistic (an estimate of the regression intercept of 0.99 with an SE of 0.01, which is not significantly different from 1), suggesting that the sample structure has been well-controlled in the MLMA-LOCO analysis as expected from theory (23). We further divided the data into the Seda and Litang subsets and reran the analysis in each subset (*Materials and Methods*). Although all of nine loci remained highly significant, not all of them passed the genome-wide significance level in either subset (*SI Appendix*, Table S3). This analysis shows the gain of power for detecting genetic signals of natural selection in a dataset of large sample size. In addition, we performed conditional analyses (27, 28) at nine genome-wide significant loci and did not find evidence of multiple signals at any of these loci. We also performed the MLMA-LOCO analysis to detect signatures of genetic adaptation on the mitochondrial genome and did not observe any significant signal (*SI Appendix*, Fig. S7). We replicated a number of candidate gene loci as reported in previous studies (3, 4). The replication rate after correcting for multiple testing was ∼35.7% (= 5/14), much higher than expected by chance (*SI Appendix*, Table S4).

**Associations of the Loci Under Natural Selection with Phenotypes in Tibetans.** Having identified nine genetic loci that have been under natural selection, we then asked whether these loci are associated

**Table 1. Nine genetic loci with signals of natural selection**

| | | | | | Frequency of A1 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Chromosome | SNP | bp | A1 | A2 | Tibetan | EAS | *P* value | $F_{ST}$ | Nearest gene |
| 1 | rs1801133 | 11,856,378 | A | G | 0.238 | 0.333 | 6.3E-09 | 0.021 | *MTHFR* |
| 1 | rs71673426 | 112,159,304 | C | T | 0.102 | 0.013 | 1.5E-08 | 0.100 | *RAP1A* |
| 1 | rs78720557 | 198,096,548 | A | T | 0.498 | 0.201 | 4.7E-08 | 0.191 | *NEK7* |
| 1 | rs78561501 | 231,448,497 | A | G | 0.599 | 0.135 | 6.1E-15 | 0.414 | *EGLN1* |
| 2 | rs116611511 | 46,600,030 | G | A | 0.447 | 0.003 | 3.6E-19 | 0.570 | *EPAS1* |
| 4 | rs2584462 | 100,324,464 | G | A | 0.211 | 0.549 | 3.9E-09 | 0.203 | *ADH7* |
| 5 | rs4498258 | 44,325,322 | T | A | 0.586 | 0.287 | 1.7E-08 | 0.171 | *FGF10* |
| 6 | rs9275281 | 32,662,920 | G | A | 0.095 | 0.365 | 1.1E-10 | 0.162 | *HLA-DQB1* |
| 12 | rs139129572 | 123,178,478 | GA | G | 0.316 | 0.449 | 5.8E-09 | 0.036 | *HCAR2* |

*P* value indicates the *P* value from the MLMA-LOCO analysis. $F_{ST}$ is the $F_{ST}$ value between Tibetans and EASs. Nearest gene indicates the nearest annotated gene to the top differentiated SNP at each locus except *EGLN1*, which is known to be associated with high-altitude adaptation; rs139129572 is an insertion SNP with two alleles: GA and G. A1, allele 1; A2, allele 2.

with any phenotypes in Tibetans ($n$ = 2,849). There were 91 quantitative traits measured on the Tibetan subjects (*Materials and Methods*), mainly morphological, blood biochemistry, and optometric measures (*SI Appendix*, Table S5). The phenotypic correlation matrix of these traits is shown in *SI Appendix*, Fig. S8. Most of the traits were highly heritable, with a substantial proportion of phenotypic variance explained by all SNPs in unrelated individuals (*SI Appendix*, Fig. S9 and Table S6). We then performed GWAS analysis in Tibetans using the MLMA-LOCO approach described above to control for population structure. We found that the methylenetetrahydrofolate reductase (*MTHFR*) and *EPAS1* loci were associated with multiple traits (*SI Appendix*, Fig. S10), and five of these traits were significant after correcting for multiple testing ($P_{GWAS} < 1.5 \times 10^{-4}$) (*SI Appendix*, Table S7). The *MTHFR* locus was strongly associated with folate ($b = -0.34$, $P_{GWAS} = 6.5 \times 10^{-27}$) and homocysteine ($b = 0.54$, $P_{GWAS} = 1.1 \times 10^{-69}$), where $b$ is the effect size in SD units. This locus is known to be associated with homocysteine in Europeans (29). MTHFR is a key enzyme involved in the metabolic pathway of homocysteine and folate (30). The frequency of homocysteine-increasing allele was lower in Tibetans (0.238) than that in EAS (0.333) (Table 1), in line with homocysteine level in Tibetans (mean = 21.8, SE = 0.3) being lower than in Han (mean = 25.5, SE = 1.4), where SE represents SEM estimate. *EPAS1* is known to be associated with HGB (3, 5). Our results suggest that *EPAS1* is strongly associated with HGB, red blood cell count, and hematocrit (*SI Appendix*, Table S7) and that the HGB-decreasing allele is under very strong positive selection in Tibetans, with a frequency of 0.45 in Tibetans vs. 0.003 in EASs (Table 1). It is also interesting to note that the *ADH7* locus is associated with weight and body mass index (BMI) in Tibetans ($P_{GWAS} = 7.1 \times 10^{-4}$ and $P_{GWAS} = 4.9 \times 10^{-4}$, respectively), although *ADH7* is not a known BMI-associated locus in Europeans (31). However, the associations are not significant after correcting for multiple testing. The *EGLN1* locus has been previously reported to be associated with HGB (4). We found that the association between *EGLN1* and HGB was very weak ($P_{GWAS} = 0.02$,, not significant after correcting for multiple testing), and the effect size was stronger in males (–0.112 in SD unit, SE = 0.046) than in females [–0.037, SE = 0.036, $P_{difference} = 0.01$, consistent with the result from a previous study (12)].

## Discussion

We have performed a large-scale genetic study in 3,008 Tibetans and 7,287 non-Tibetans of EAS ancestry. We showed the genetic relatedness between Tibetans and a number of other ethnic groups in China and found that Yi, Tu, and Naxi people are genetically intermediate between Han and Tibetans (Fig. 1*A*). These people are also geographically distributed between major residential areas of Han and Tibetans (Yi, Tu, and Naxi people reside at the eastern border of the TP) (Fig. 1*B*), suggesting potential routes of people migrating from the east to the TP. There has not been a consensus on the divergence time between Tibetans and Han (32). The estimates from different genetic studies are often inconsistent [varying from 2,750 (3) to ∼8,000 (11, 12) and ∼30,000 y B.P. (33)], even for studies using the same method [9,000–15,000 (34) vs. 20,000–40,000 y B.P. (35)]. Our estimate from, so far, the largest genetic data of Tibetans is that the divergence time between Tibetans and Han was ∼4,725 y B.P., which is consistent with the estimated permanent settlement time of ∼3,750–6,500 y B.P. from archaeological studies (20). Interestingly, a recent study (21) that investigates archaeological crop remains unearthed in the northeastern TP estimated that the first village was established 5,200 y B.P., which is highly concordant with our estimate. However, there is an important caveat in interpreting estimates from population genetics analyses. That is, if there is a constant gene flow from the founder population to the TP after initial settlement, then the estimate of divergence from a population genetic analysis will be biased downward. Therefore, our estimate should be interpreted as a lower limit of the permanent settlement time, implying that the actual settlement time of people in the TP is likely earlier than 4,725 y B.P.

We applied the MLMA-LOCO method (27) as implemented in BOLT-LMM (22) to detect genetic signals of selection. Compared with the prevailing methods (3, 5, 6), the MLMA-LOCO performs statistical tests at a genome-wide significance level, controlling for locus-specific population differentiation and potential relatedness in the sample. It is expected that the analysis using unrelated individuals was, on average, less powerful than using all of the individuals, but overall, the results are highly consistent (*SI Appendix*, Fig. S11). We show below an example of how the MLMA-LOCO controls for locus-specific population differentiation. There were three SNPs (on chromosomes 9, 20, and 22) that showed strong signals in linear regression (5), $F_{ST}$ (18, 24, 36, 37), or PBS (3) analysis but did not reach the genome-wide significance level in the MLMA-LOCO analysis (*SI Appendix*, Fig. S12) because the three SNPs are located in regions with strong locus-specific population differentiation (*SI Appendix*, Fig. S13). Using the MLMA-LOCO method, we identified nine gene loci that have been under selection as a consequence of adaptation to the high altitude (Fig. 2 and *SI Appendix*, Fig. S5), seven of which are unique. It is noteworthy that there are surprisingly few loci that have been identified given the large sample size of this study, consistent with a model of polygenic adaptation (38). The two known loci (*EPAS1* and *EGLN1*) showed the strongest signals in our analysis. The top signal (*EPAS1*) remained highly significant in the analysis of a small subset of data (150 Tibetans vs. 150 Han) (*SI Appendix*, Fig. S14), which explains why the *EPAS1* locus can be detected in previous studies of small sample size (3, 5, 6). We further found that genetic variants at three of these loci (*MTHFR*, *EPAS1*, and *ADH7*) were associated with several phenotypes in Tibetans, with *MTHFR* being associated with folate and homocysteine levels and *EPAS1* being associated with HGB and hematocrit at an experimentwise significance level (*SI Appendix*, Fig. S10). In addition, it was suggested in a previous study (4) that the *PPARA* gene locus is associated with high-altitude adaptation and HGB level in Tibetans. In our study, we found that the signal of selection at the *PPARA* was not genome-wide significant ($P_{MLMA\text{-}LOCO} = 9.1 \times 10^{-5}$ at the top SNP rs149670586) and did not find any evidence that rs149670586 is associated with HGB ($P_{GWAS} = 0.20$). There is a caveat in interpreting the MLMA-LOCO results. We found evidence of natural selection at nine gene loci by comparing the allele frequencies between Tibetans and EASs under the null hypothesis that there is no natural selection but a population differentiation caused by genetic drift and possibly, admixture with other populations. This result, however, does not necessarily mean that the selection has to relate to hypoxia. It could be adaptation to any of the extreme environmental or pathological conditions in the TP. For example, the folate-increasing allele of the SNP rs1801133 at the *MTHFR* locus (*SI Appendix*, Table S7) has an increased frequency in the Tibetan population, more than expected under a drift model (Table 1), which is possibly a consequence of adaptation to high UV radiation, because the degradation of folate could be accelerated by UV exposure (39).

In summary, we performed a large-scale genetic study in Tibetans. We showed the genetic relatedness between Tibetans and other ethnic groups in China and estimated divergence time between Tibetans and Han (4,725 y B.P.). We identified genetic signatures of high-altitude adaptation at seven gene loci. These findings provide important insight into understanding of how the Tibetan genome has changed during high-altitude adaptation.

## Materials and Methods

**Sample Collection and Genotyping.** The subjects were recruited at two separate sites (Seda and Litang) of the TP in Sichuan, western China (*SI Appendix*, Fig. S1). Both sites are ∼4,100 m above sea level. The Seda subjects were

recruited from people who were studying or working at the Seda Larong Wuming Buddhist Institute. These subjects are originally from diverse regions of the TP. The Litang subjects were recruited from nomadic people who have lived in Litang and surrounding areas for many generations. All of the subjects were recruited following the protocol approved by the Ethics Committee of the WZ. An informed written consent was obtained from each subject participating in this study. There were a total of 3,996 subjects with blood samples (3,142 from Seda and 854 from Litang). Peripheral blood sample was obtained from each subject for extraction of genomic DNA (Simgen Blood DNA Mini Kit; Simgen). DNA concentrations were subsequently determined using a NanoDrop 1000 spectrophotmeter (Thermo Scientific). DNA samples were subjected to array genomic hybridization using the HumanCoreExome-12 BeadChip (Illumina Inc.). An iScan Reader was used to scan the array slide (Illumina Inc.). SNP genotypes were called using the GenCall algorithm implemented in GenomeStudio (GenTrain Score threshold = 0.15).

**QC.** There were 3,717 subjects genotyped on 542,585 SNPs (526,123 on autosomes) before QC. We removed SNPs and individuals with missingness rate > 5% and excluded SNPs with minor allele frequency (MAF) < $3.3 \times 10^{-4}$ [equivalent to minor allele count (MAC) < 3] or Hardy–Weinberg Equilibrium (HWE) test $P$ value < $1.0 \times 10^{-6}$. We flipped strand for the SNPs that were not called on the forward strand using the latest annotation file from Illumina (https://support.illumina.com/downloads/humancoreexome-12v1-1_product_support_files.html) and removed SNPs with alleles called from the subjects that were inconsistent with those in the annotation file. We then used GCTA-GRM (27) to estimate the genetic relatedness between pairwise individuals using all of the common (MAF ≥ 0.01) SNPs on autosomes after QC and removed one of each pair of individuals (the one with higher missingness rate) with estimated genetic relatedness > 0.8 (all were duplicated subjects, except one self-reported monozygotic twin pair).

We performed a PCA (40, 41) in the sample using GCTA-PCA (27) on all common autosomal SNPs after QC and projected the PCs to the subjects from the 1000G (42). Our study subjects were stratified along PC1 (*SI Appendix*, Fig. S2*A*), because there were a few hundred self-reported Han Chinese in the sample. This information was confirmed by projecting the PCs estimated from our sample on the Han Chinese subjects from the 1000 Genome Project (1000G-Han) (*SI Appendix*, Fig. S2*C*). We stratified our subjects into three groups (Han, Tibetans, and possibly admixed individuals) using the following method. (*i*) We classified the individuals whose PC1s were less than four SDs away from mean PC1 of 1000G-Han (mean = −0.039 and SD = 0.0044) as Han. (*ii*) We then fitted a two-component mixture normal distribution to PC1 and classified the individuals whose PC1s were less than four SDs away from the mean of the second mixture component (mean = 0.0073 and SD = 0.0015) as Tibetans. (*iii*) The rest of the subjects were classified as possibly admixed individuals (or individuals from the other minority groups) and excluded from subsequent analyses (*SI Appendix*, Fig. S2*D*). After all of the QC filters, we retain 3,381 individuals (3,008 Tibetan and 373 Han) and 287,691 SNPs (279,608 autosomal SNPs). We then imputed the cleaned genotype data to 1000G reference panels (phase 1) (42) using IMPUTE2 (43). After imputation, we removed monomorphic SNPs, SNPs with HWE $P$ value < $1 \times 10^{-6}$, SNPs with IMPUTE-INFO < 0.3, or SNPs with MAF < 0.01 in Tibetans and retained ∼7.8 million SNPs for analysis.

**GWAS Data of Han from Wenzhou and EASs from the Database of Genotypes and Phenotypes.** We had access to a GWAS dataset of 2,043 Han Chinese subjects who were recruited at the WZ using the same protocol. The genotyping experiments were performed on the same genotyping platform (Illumina CoreExome array) as described above. We removed SNPs and subjects with missingness rate > 5% and excluded SNPs with HWE $P$ value < $1 \times 10^{-6}$ or MAC < 3. After QC, there were 1,726 subjects and 270,630 SNPs (263,345 autosomal SNPs). The cleaned genotype data were imputed to 1000G reference panels (phase 1) using IMPUTE2. After imputation, monomorphic SNPs or SNPs with HWE $P < 1 \times 10^{-6}$ or IMPUTE-INFO score < 0.3 were excluded from analysis. Although the subjects were recruited at the WZ, the individuals came from various different regions of China as confirmed by the genetic data (*SI Appendix*, Fig. S3).

We also had access to GWAS data of the GERA Study (25) in the United States through the database of Genotypes and Phenotypes (accession no. phs000674.v2.p2). There were 5,188 EAS subjects genotyped on Affymetrix Axiom arrays. QCs of the genotype data have been detailed elsewhere (25). We followed the QC protocol provided by the GERA Study, and we further removed SNPs and individuals with missingness rate > 2% and excluded SNPs with HWE $P$ value < $1 \times 10^{-6}$ or MAC < 3. We imputed the genotype data to all of the 1000G reference panel (Phase 1) using the same imputation pipeline as described above and removed SNPs with HWE $P$ value < $1 \times 10^{-6}$ or IMPUTE-INFO < 0.3 postimputation.

**Genome-Wide Analysis to Detect Genetic Signals of Selection.** We performed a genome-wide analysis to detect signals of selection by comparing allele frequency of each SNP between two diverged populations (e.g., Tibetan and Han) based on the following model:

$$y = \mu + xb + \sum_k w_{ik} u_k + e,$$

where $y$ is coded as one or zero to indicate population for an individual (e.g., one for Tibetan and zero for Han), $\mu$ is a fixed mean term, $x$ is a genotype variable (coded as zero, one, or two for the three genotypes) for an SNP, $b$ is a fixed effect that is a function of allele frequency difference between the two populations, $\sum_k w_{ik} u_k$ is a term that fits all of the SNPs on the other chromosomes in the model to control for population differentiation, $w_k$ is the standardized genotype variable for an SNP $k$, $u_k$ is the corresponding effect size [$u_k$ follows a normal distribution with variance being proportional to $p_0(1 - p_0)F_{ST}$, where $p_0$ is the allele frequency in the ancestral population], and $e$ is the residual. In GWASs, this analysis is called the MLMA-LOCO (23). We have shown previously by theory and simulations that this method not only controls for population structure but also, gains power compared with linear regression (23). The MLMA-LOCO method has been implemented in the GCTA-MLMA (23, 27) and BOLT-LMM (22) software tools. We used BOLT-LMM in this study, because it is computationally more efficient than GCTA-MLMA when sample size is large. If we do not include the random effect term in the model; it will become a linear regression analysis (i.e., $y = \mu + xb + e$), analogous to an $F_{ST}$ analysis (*SI Appendix*, Fig. S12 has a comparison between linear regression and $F_{ST}$ analysis). The test statistics from linear regression will be inflated because of the population differentiation caused by genetic drift. Post hoc correction approaches, such as "Genomic Control" (44), can be used to correct for the inflation; however, Genomic Control correction could lead to overcorrection (23) (*SI Appendix*, Table S2) and is not able to account for potential locus-specific differentiation (*SI Appendix*, Fig. S13).

For data analysis, we used all of the Tibetan and Han subjects collected from the TP and WZ and all of the EAS subjects from the GERA cohort. All of the genotype data have been imputed to the 1000G and passed the QC steps as described above. We included only the SNPs in common among the datasets and SNPs with MAF ≥ 0.01 in Tibetans and the combined sample. There were ∼7.3 million SNPs on 10,292 individuals (3,008 Tibetans, 2,098 Han, and 5,186 EAS) included in analysis. The WZ-Han and GERA-EAS subjects were genotyped and imputed separately from the TP subjects, and the GERA-EAS subjects were genotyped on a different type of SNP array (the GERA-EAS subjects were genotyped on Affymetrix Axiom arrays, and all of the other subjects were genotyped on Illumina CoreExome arrays). To control for difference in allele frequency caused by genotyping or imputation artifacts, we performed a control–control analysis using the MLMA-LOCO method as described above to test for allele frequency (AF) difference between TP-Han ($n$ = 373) and WZ-Han + GERA-EAS ($n$ = 6,911) and removed SNPs at $P < 1 \times 10^{-6}$. To detect genetic signals of high-altitude adaptation, we then performed an MLMA-LOCO analysis to test for AF difference between Tibetans ($n$ = 3,008) and TP-Han + WZ-Han + GERA-EAS ($n$ = 7,284). We further reran the analyses in two subsets of data to show the gain of power by combining all of the available samples [i.e., (*i*) Seda-Tibetan ($n$ = 2,427) vs. Seda-Han + GERA-EAS ($n$ = 5,548); (*ii*) Litang-Tibetan ($n$ = 581) vs. Litang-Han + WZ-Han ($n$ = 1,736), where Seda and Litang are two sites in the TP as described above]. Summary-level statistics from all of the MLMA-LOCO analyses are available at cnsgenomics.com/data/yang_et_al_2017_pnas.html. The raw genotype and phenotype data of the Tibetan and Han subjects are available through application at https://www.wmubiobank.org.

**ACKNOWLEDGMENTS.** We thank Jonathan Pritchard for constructive and helpful comments on an earlier version of the manuscript. This study was supported by National Key Basic Research Program Grant 2013CB967502; National Natural Science Foundation of China Grants 81522014, 81371059, 81470659, 81500741, 81271039, and 81570880; Ministry of Science and Technology (MOST) Projects Grant 2012YQ12008004; National Key Clinical Specialty (Ophthalmology); the project of Zhejiang Provincial Top Key Discipline and Key Construction Discipline of Medicine; Wenzhou Science and Technology Innovation Team Project Grant C20150004; Zhejiang Provincial Natural Science Foundation of China Grants LR13H120001 and LQ14B020005; National Health and Family Planning Commission (NHFPC) Grant-in-Aid for Medical and Health Science 201472911; Australian National Health and Medical Research Council Grants 1078037, 1078901, 1107258, and 1113400; Australian Research Council Grant 160101343; and the Sylvia & Charles Viertel Charitable Foundation.

1. Beall CM (2007) Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *Proc Natl Acad Sci USA* 104(Suppl 1):8655–8660.
2. Dahlback A, Gelsor N, Stamnes JJ, Gjessing Y (2007) UV measurements in the 3000–5000 m altitude region in Tibet. *J Geophys Res Atmos* 112:D09308.
3. Yi X, et al. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329:75–78.
4. Simonson TS, et al. (2010) Genetic evidence for high-altitude adaptation in Tibet. *Science* 329:72–75.
5. Beall CM, et al. (2010) Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. *Proc Natl Acad Sci USA* 107: 11459–11464.
6. Xu S, et al. (2011) A genome-wide search for signals of high-altitude adaptation in Tibetans. *Mol Biol Evol* 28:1003–1011.
7. Peng Y, et al. (2011) Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Mol Biol Evol* 28:1075–1081.
8. Tian H, McKnight SL, Russell DW (1997) Endothelial PAS domain protein 1 (EPAS1), a transcription factor selectively expressed in endothelial cells. *Genes Dev* 11:72–82.
9. Tian H, Hammer RE, Matsumoto AM, Russell DW, McKnight SL (1998) The hypoxia-responsive transcription factor EPAS1 is essential for catecholamine homeostasis and protection against heart failure during embryonic development. *Genes Dev* 12: 3320–3324.
10. Lee FS, Percy MJ (2011) The HIF pathway and erythrocytosis. *Annu Rev Pathol* 6: 165–192.
11. Lorenzo FR, et al. (2014) A genetic mechanism for Tibetan high-altitude adaptation. *Nat Genet* 46:951–956.
12. Xiang K, et al. (2013) Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. *Mol Biol Evol* 30:1889–1898.
13. Wang GD, et al. (2014) Genetic convergence in the adaptation of dogs and humans to the high-altitude environment of the Tibetan plateau. *Genome Biol Evol* 6:2122–2128.
14. Wuren T, et al. (2014) Shared and unique signals of high-altitude adaptation in geographically distinct Tibetan populations. *PLoS One* 9:e88252.
15. Jha AR, et al. (2016) Shared genetic signals of hypoxia adaptation in Drosophila and in high-altitude human populations. *Mol Biol Evol* 33:501–517.
16. Rosenberg NA, et al. (2002) Genetic structure of human populations. *Science* 298: 2381–2385.
17. Nei M (1987) *Molecular Evolutionary Genetics* (Columbia Univ Press, New York).
18. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370.
19. McEvoy BP, Powell JE, Goddard ME, Visscher PM (2011) Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res* 21:821–829.
20. Aldenderfer M (2011) Peopling the Tibetan plateau: Insights from archaeology. *High Alt Med Biol* 12:141–147.
21. Chen FH, et al. (2015) Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 B.P. *Science* 347:248–250.
22. Loh PR, et al. (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 47:284–290.
23. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* 46:100–106.
24. Bhatia G, Patterson N, Sankararaman S, Price AL (2013) Estimating and interpreting FST: The impact of rare variants. *Genome Res* 23:1514–1521.
25. Banda Y, et al. (2015) Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. *Genetics* 200:1285–1295.
26. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 44:955–959.
27. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* 88:76–82.
28. Yang J, et al.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 44:369–375.
29. van Meurs JB, et al. (2013) Common genetic loci influencing plasma homocysteine concentrations and their effect on risk of coronary artery disease. *Am J Clin Nutr* 98: 668–676.
30. Jacques PF, et al. (1996) Relation between folate status, a common mutation in methylenetetrahydrofolate reductase, and plasma homocysteine concentrations. *Circulation* 93:7–9.
31. Locke AE, et al.; LifeLines Cohort Study; ADIPOGen Consortium; AGEN-BMI Working Group; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GLGC; ICBP; MAGIC Investigators; MuTHER Consortium; MIGen Consortium; PAGE Consortium; ReproGen Consortium; GENIE Consortium; International Endogene Consortium (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518: 197–206.
32. Qiu J (2015) Who are the Tibetans? *Science* 347:708–711.
33. Qi X, et al. (2013) Genetic evidence of paleolithic colonization and neolithic expansion of modern humans on the tibetan plateau. *Mol Biol Evol* 30:1761–1778.
34. Lu D, et al. (2016) Ancestral origins and genetic history of Tibetan Highlanders. *Am J Hum Genet* 99:580–594.
35. Jeong C, et al. (2014) Admixture facilitates genetic adaptations to high altitude in Tibet. *Nat Commun* 5:3281.
36. Shriver MD, et al. (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum Genomics* 1:274–286.
37. Raj SM, Pagani L, Gallego Romero I, Kivisild T, Amos W (2013) A general linear model-based approach for inferring selection to climate. *BMC Genet* 14:87.
38. Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: Hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol* 20:R208–R215.
39. Borradale DC, Kimlin MG (2012) Folate degradation due to ultraviolet radiation: Possible implications for human health and nutrition. *Nutr Rev* 70:414–422.
40. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909.
41. Galinsky KJ, et al. (2016) Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *Am J Hum Genet* 98:456–472.
42. The 1000 Genomes Project Consortium, et al. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.
43. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5: e1000529.
44. Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55: 997–1004.