Supporting Information

Beall et al. 10.1073/pnas.1002443107

SI Materials and Methods

Participant Sample 1: Yunnan Tibetans. Data collection took place in June 2007. Samples from the Yunnan province came from two districts (Shangri-La and De Qin) at altitudes from 3,200 to 3,500 m, from four townships (Ben Zi Lan, 28°16'6.04"N 99°12'59.57"E; Jian Tang, 27°50'20.75"N 99°45'49.42"E; Adun Zi, 28°26'53.50"N 98°53'44.21"E; and Ju Shui, 28°26'28.45"N 98°52'35.38"E). The population ethnicity of each of these townships is heavily Tibetan. Participants were recruited based on ethnicity as recorded by the Yunnan province Family Planning Institute. Participants were also asked to declare grandparental ethnicity during the recruitment process. Each participant provided a saliva sample using Oragene DNA collection kits (DNA, Genotek) and basic health phenotype details. We recorded the following for each participant: family details, weight, height, heart rate, blood pressure, heart disease status, respiratory/digestive/urinary/endocrine health, and basic altitude sickness details. A set of 48 individuals was selected for genome scanning. After exclusions for quality control (QC) and relatedness, data from a total of 35 individuals remained for analysis.

Participant Sample 2: Mag Xiang Tibetans. Data collection took place from June to August 2002 in Mag Xiang, a rural agropastoral district of Xigatse Prefecture at 4,200 m, Tibet Autonomous Region. Details on this collection have been reported previously (1, 2). All participants were 18–55 y of age, normotensive, nonanemic, nonsmoking (by self-report and verification of exhaled carbon monoxide level), healthy (by self-report), not pregnant (by self-report), and had normal pulmonary function. They were all born and raised in Mag or nearby villages and were of Tibetan ethnicity. Hemoglobin concentration was determined in duplicate using the cyanmethemoglobin technique (Hemocue Hemoglobinometer; Hemocue AB), immediately after drawing a venous blood sample. DNA samples were collected for each study participant from buccal cells obtained by using the mouthwash method (3). Sufficient DNA was isolated for 57 of the study participants from this procedure. An additional 17 study participants provided DNA samples from buccal and white cells in saliva obtained using the Oragene DNA collection kits in 2007, bringing the total number of subjects with sufficient DNA to 74. After quality control, genotypic and phenotypic data were available for 70 people.

Participant Sample 3: Zhaxizong Xiang Tibetans. Data collection took place in June 2007 in Zhaxizong Xiang, a rural agropastoral district of Xigatse Prefecture, at an altitude of 4,300 m at the base of Mt. Qomolangma. Adults representing about 10% of the total population were recruited for a health survey and collection of phenotypic and genetic information. Study participants were 18–55 y of age, unrelated, high-altitude native Tibetans whose families had lived in the villages for at least two generations. A subset of 92 participants with values of hemoglobin and blood oxygen saturation within normal ranges was selected for inclusion in the present study. Hemoglobin and other blood parameters were measured with an Auto Hematology Analyzer BC-2800 (Shenzhen Mindray Bio-Medical Electronics Co. Ltd). DNA was isolated from blood samples using QIAGEN kits. After quality control, complete genotypic and phenotypic data were available for 91 people.

Participant Sample 4: HapMap Han. Data for the Han population were accessed from the HapMap project (HapMap3, release 2). We selected the HapMap Han as a lowland population with common ancestry, as Y chromosome and mitochondrial DNA studies have illustrated genetic continuity between numerous different Tibetan groups and the Han. Genetic and archaeological evidence suggests that the majority of Tibetan lineages can trace their ancestry to postPaleolithic and Neolithic migrations from Northern China approximately 10,000 y ago (4, 5). Genomic analysis was conducted on the 503,722 SNPs common to both HapMap and Yunnan Tibetan datasets.

Whole Genome SNP Genotyping. Genotyping of the Yunnan Tibetan samples (sample 1) was conducted using the Illumina Veracode platform and 610-Quad high throughput genotyping chips. Forty-eight samples from the Yunnan collection were genotyped (broken down according to township as: Ben Zi Lan, $n = 15$; Jian Tang, $n = 20$; Adun Zi, $n = 6$; Ju Shui, $n = 7$).

All samples were brought into a single BeadStudio file using the standard Illumina cluster file. Samples were checked for very low intensity and for call rates <95%, and all samples passed this first QC step. All SNPs that had a call frequency below 100% were then reclustered. Samples were then checked for call rates <98% after the reclustering, and again all samples passed this second QC step. Next, a "1% rule" was applied where all SNPs that had a call frequency below 99% were deleted. Any SNPs where more than 1% of samples were not called or were ambiguously called were deleted (SNPs with many samples not called (or potentially miscalled) can lead to false positives in statistical associations). We introduced the following procedure to prevent errant calls (from SNP reclustering) from entering the final report. The SNP data are screened within BeadStudio by looking at two criteria. First, all SNPs with a cluster separation value below 0.3 are manually checked to ensure correct calls. Many of these SNPs can be manually fixed, but some have to be deleted. Next, all SNPs (excluding X chromosome SNPs) with a Het Excess value between −1.0 to −0.1 and 0.1 to 1.0 are evaluated to determine if the raw and normalized data show a clean call. Any SNP cluster that doesn't appear normal is deleted. This includes SNPs that appear to show a deletion (hemizygotes and homozygous deletion). The rationale behind this is to avoid artifacts from either the chemistry or an interfering SNP during hybridization. These procedures resulted in a 97.5% success rate of genotyping calls.

To minimize problems arising from hidden population and family structure in the Yunnan sample, we removed 13 individuals using the following quality control steps. First, we computed pairwise identity-by-descent calculations for all pairs of individuals (Pi_Hat calculation in PLINK; ref. 6). We identified six families defined by members sharing greater than third degree relatedness. Four of these involved two individuals, one involved five individuals, and one involved six individuals. We removed the fewest number of individuals consistent with breaking all first and second degree relationships (Pi_Hat < 0.17). Where a choice of individuals was possible, we picked the sample with the highest inbreeding coefficient (a measure of average genome-wide homozygosity) for removal. This resulted in 10 exclusions. Next, we removed two individuals with inbreeding coefficient values greater than 0.09 (appearing as outliers on a histogram of inbreeding values). Finally, we removed one individual who appeared as a clear outlier from PCA analysis of the genotype data. We then removed 139 SNPs with less than 95% coverage across the reduced sample of 35 individuals. The genotype dataset complied with that expected of a population in Hardy Weinberg equilibrium.

For the GWADS analysis described later, we downloaded HapMap Phase III Han Chinese data (release 2, accessed 2 Feb 2009; CHB sample, $n = 84$, typed on both Illumina 1M and

Affymetrix 6.0 platforms) from www.hapmap.org. We merged this dataset with our post-QC Yunnan dataset and restricted to those SNPs common to both datasets (n = 504,450). We removed 726 symmetric SNPs (A/T or C/G alleles) and 2 SNPs with ambiguous map positions. Finally, we removed 94 SNPs that emerged as singletons from GWADS analysis. These were SNPs with a very high apparent allelic frequency difference between the Yunnan Tibetan and HapMap Han samples but with no support for this signal from neighboring SNPs in the region.

Candidate Gene Genotyping in the Mag Xiang Discovery Sample. A GoldenGate assay was used for all 74 samples from Mag Xiang (Sample 2). *EPAS1* was 1 of 11 candidate genes in a customer-designed Illumina GoldenGate assay (384 SNP plex). For SNP determination, first, all haplotype-tagging SNPs for the *EPAS1* gene were chosen from genotypes of Asian populations (HapMap Han/Beijing and Japanese/Tokyo) of the HapMap project (Phase II dataset). In addition, nonoverlapping tag SNPs of the HapMap African population (Yoruba) were included to increase genotyping coverage. In total, 103 SNPs across the *EPAS1* gene were selected for genotyping.

BeadStudio was used for genotype determination. A GenCall score of 0.25 and GenTrain score of 0.55 were applied for initial QC screening. Genotypes were then verified by manual clustering; loci whose homo- and heterozygote clusters did not separate clearly were eliminated. Data from 73 individuals passed QC. Of the 103 *EPAS1* SNPs in the GoldenGate assay, 96 SNPs passed QC, of which 49 SNPs were common with a minor allele frequency (MAF) $\geq$ 0.05.

Of the 73 individuals whose genotyping passed quality control, phenotypic data were missing in one individual and measurements of hemoglobin concentration in a further two individuals. Therefore the discovery analysis was conducted on 49 common SNPs of the *EPAS1* gene in 70 individuals of the Mag Xiang cohort.

Candidate Gene Genotyping in the Zhaxizong Xiang Replication Sample. GoldenGate and MassARRAY assays were used to genotype the *EPAS1* gene on a total of 92 individuals from Zhaxizong Xiang (sample 3). With the same quality control procedure as described above, individual GoldenGate assays were carried out on a set of 44 samples, and another 48 samples were genotyped by GoldenGate assay as pooled groups to increase sample usage. Minor allele frequency was then calculated from the genotype data from all 92 samples, and this gave a total of 58 SNPs of the *EPAS1* gene with a MAF $\geq$ 0.05. These SNPs were selected for further genotyping on each individual sample from the 48 previously pooled samples using MassARRAY analysis. Typer 4.0 software (Sequenom, Inc.) was used to collect genotype data after clustering. A specificity of 0.90 and sensitivity of 0.95 were set to cluster all loci. After clustering, all samples with call rates less than 80% or loci with call frequencies less than 95% were eliminated. Next, SNPs with homo- and heterozygote clusters not clearly separated were removed manually. One entire sample and six SNP loci were excluded by QC; this gave 52 SNPs in 47 individuals with a call rate of 99.67%.

Combining the data from the 47 individuals genotyped by MassARRAY with the data from the 44 individuals individually genotyped by GoldenGate assay resulted in genotypes for 52 SNPs from 91 samples from Zhaxizong Xiang. Further alignment of the 52 SNPs with the successful SNPs of Mag Xiang samples in discovery analysis yielded 48 overlapping SNPs. Of these 48 SNPs, 45 with common loci (MAF $\geq$ 5%) were analyzed in 91 Zhaxizong Xiang samples in the replication study.

Following the quality control procedure outlined above, the eight significant SNPs identified in the genome-wide allelic differentiation scan (see below) were also genotyped using the MassARRAY system. A subset of 30 individuals from Mag Xiang and all 92 individuals from Zhaxizong Xiang were analyzed. One sample from Mag Xiang and three samples from Zhaxizong Xiang were eliminated after quality control analysis. Call rates were 100% and 99.72% in the 29 and 89 samples analyzed from Mag Xiang and Zhaxizong Xiang, respectively.

SI Statistical Analysis and Results

Genome-Wide Allelic Differentiation Scan (GWADS). Here, we present the method used for conducting differential scans, which is based on approaches originally developed for genome-wide association studies. Existing methods for detecting differential selection rely on the calculation of statistics such as F_{ST} (7) or iHS (8), which are intuitive but lack any simple distributional theory. This prevents the reliable testing of genome-wide significance of these statistics. Typically, one must either resort to genome-wide simulation, which limits the conclusions to the demographic model used with its particular assumptions, or on reporting the extreme tail values of a genome-wide distribution, which abrogates genome-wide significance testing altogether because there will always be, for example, a 5%, 1%, or 0.1% tail to every distribution.

In particular, we use SNP-by-SNP allelic χ^2 statistics, calculated between two population samples, as an appropriate differentiation statistic, and we correct for background levels of differentiation using genomic control. This allows for unusually differentiated loci to be declared as genome-wide significant. Allelic (or allele-counting) χ^2 statistics are not usually used for genetic association testing because departures from the null under simple genetic risk models typically result in equal power for the allelic and Cochran Armitage trend tests, and the latter enjoys the advantage that the assumption of Hardy Weinberg equilibrium is not required (9). However, allelic tests are entirely appropriate for tests of population differentiation because departures from the null under simple population genetic models for this scenario typically result in greater power for allelic tests over trend tests, and departures from the assumption of Hardy Weinberg equilibrium (which need only hold within each population for the test to be valid) can in any case be corrected for by genomic control.

Genomic control was proposed by Devlin and colleagues (10), but the distributional theory that justifies it came in a later paper by Devlin and colleagues (11). They showed that, under very general conditions, one can expect the allelic χ^2 statistic of a SNP typed in two population samples to be inflated in the presence of population stratification by an amount that depends on sample size, F_{ST} and inbreeding coefficients, but not on the allele frequency of the SNP in question. The amount of χ^2 statistic inflation in other SNP loci can therefore be used as a yardstick to gauge the significance of the inflation seen in any one SNP of interest. In a genome-wide context, averaging all SNPs in a genome-wide panel allows for the calculation of an inflation factor, lambda, which represents the "background" population level of differentiation between two populations. The corresponding inflated χ^2 distribution (or equivalently the standard χ^2 distribution, if the test statistic is first divided by lambda) represents an appropriate distribution from which to test the genome-wide significance of any one observed χ^2 statistic. Because the same lambda is applied to all SNPs, genomic control is only capable of correcting for background differentiation, and this property is seen as a failing of the method in the context of genetic association because it is possible to envisage highly differentiated SNPs that may still confound an association test (12). However, in a GWADS it is precisely these unusual, highly differentiated SNPs that one wishes to detect, as selection represents the best hypothesis to account for such cases.

The distributional theory presented by Devlin et al. (11) makes certain normality assumptions that do not hold exactly for the case of only two populations. Furthermore, genomic control has not been widely investigated under a combination of high F_{ST} (envisioned for GWADS but unlikely to hold for well designed association studies) and the extremely large χ^2 values necessary for testing genome-wide significance. We therefore confirmed the

effectiveness of genomic control via simulation (Figs. S1–S3), as described below.

Fig. S1 is a quantile-quantile (QQ) plot summarizing the results of GWADS, applied to all autosomal SNP loci, for the Yunnan Tibetan versus HapMap Phase 3 Han samples. The red line is where values are expected to fall if there is no population differentiation between these two samples. The green line is where values are expected to fall if all loci were subject to the same level of background differentiation between the two samples, using genomic control theory and using the robust median method of Devlin and Roeder (10) to estimate lambda (there are several methods for calculating lambda, but all typically converge to the same value as the number of SNPs used for estimation grows large, as is the case here). The two blue lines represent a check on the validity of the genomic control result based on simulation. These are χ^2 values obtained from two sets of 500,000 SNPs that were simulated under a Balding-Nichols model (13) using an F_{ST} value of 0.0088 (equal to that observed between the Yunnan Tibetan and HapMap Han samples) and using a β (0.888,0.888) distribution to draw the ancestral allele frequency of each SNP (again derived from the empirical allele frequency distribution observed in the data). The distinct departure of observed points (black x-marks) above the green and blue lines represents a significant departure of some SNPs from the background level of population differentiation. This can be seen by their departure from 95% "concentration bands" (14) indicated in gray, which mark out the approximate zone within which one would expect QQ plots to reside under the null hypothesis (which here has been adjusted to reflect expected genomic control inflation).

The majority of SNPs at the top end of the QQ plot are found to come from the genomic region around *EPAS1* (chr2:46400000. 0.46900000; Fig. S4 and Fig. S5). When SNPs from this genomic region are removed (Fig. S2), the QQ plot for the remaining autosomal SNPs is more in line with the 95% concentration band, although with enough departure above it to suggest further investigation and study would be warranted (*SI Enrichment of HIF Pathway Genes in Subgenome Wide Significant Results*).

We analyzed the X chromosome separately, and the resulting QQ plot is shown in Fig. S3. We treat this chromosome separately because genomic control theory predicts the lambda to be less than that for autosomal SNPs. Following the distributional theory provided by Devlin et al. (11), we expect $\lambda \approx 1 + nc$, where n is the number of chromosomes sampled and c is a constant that depends on F_{ST} and inbreeding coefficients. If n_A is the number of autosomal chromosomes sampled, then $n_X = n_A(g_f + 0.5g_m)$, where n_X is the number of X-chromosomes sampled, and g_m and g_f are, respectively, the frequencies of males and females sampled. Thus, if we assumed that the underlying F_{ST} and inbreeding coefficients were the same for both autosomal and X-chromosome loci, we would then predict $\lambda_X \approx 1 + (g_f + 0.5g_m)(\lambda_A - 1)$, where λ_A and λ_X are the respective lambda values for autosomal and X-chromosome loci, and we could use that to construct an appropriately weighted joint analysis. However, it is possible for sex-specific migration patterns to result in different background levels of F_{ST} for these two sets of loci. For this reason, we have chosen to consider the X-chromosome separately. Fig. S3 shows that there are no large signals of selection on the X-chromosome indicated by our data.

Enrichment of HIF Pathway Genes in Subgenome Wide Significant Results. In view of the role of *EPAS1* as a core gene of the hypoxia-inducible factor transcription activation pathway, we investigated the other core genes of this pathway to determine whether they associate more than expected by chance with other regions of potential interest from the GWADS analysis. The set of core candidate genes was defined as: *HIF1α*, *EPAS1*, *HIF3α*, *ARNT*, *EGLN1*, *EGLN2*, *EGLN3*, *HIF1AN* and *VHL*. We then defined a weak GWADS P value cutoff of 5×10^{-4}, corresponding approximately to the point in Fig. S2 at which the observed GWADS χ^2 values among non-*EPAS1* autosomal SNPs showed some evidence of additional signals by dint of rising above the 95% Q-Q plot concentration band. This procedure produced a set of 375 non-*EPAS1* SNPs, as well as an expanded set of 31 SNPs in the *EPAS1* region that spanned chr2:46420780–46863250 in Build 36 coordinates and covered most of the *EPAS1* gene (chr2:46378067–46467340). See Table S1 for a full list of these SNPs. We excluded *EPAS1* SNPs from further analysis to ask whether there was additional evidence of coincidence of GWADS signals with the other eight genes in the core candidate gene set. We defined a "GWADS signal region" to be ±50 kb around each of the 307 non-*EPAS1* SNPs with a GWADS P value $< 5 \times 10^{-4}$. SNPs within 50 kb of each other were grouped in one region, and 180 regions were defined in this way, ranging in size from the defined minimum of 100 kb up to 302 kb and covering 22.2 Mb in total. We found that two of the eight genes—*HIF1α* and *EGLN3*—lay within a GWADS signal region. The P value for this coincidence, constructed from a null that assumes random placement of GWADS signal regions throughout the genome, is 0.0014. We conclude that additional evidence exists for differential selection among other genes in the HIF pathway.

1. Hoit BD, et al. (2005) Nitric oxide and cardiopulmonary hemodynamics in Tibetan highlanders. *J Appl Physiol* 99:1796–1801.
2. Erzurum SC, et al. (2007) Higher blood flow and circulating NO products offset high-altitude hypoxia among Tibetans. *Proc Natl Acad Sci USA* 104:17593–17598.
3. Lum A, Le Marchand L (1998) A simple mouthwash method for obtaining genomic DNA in molecular epidemiological studies. *Cancer Epidemiol Biomarkers Prev* 7: 719–724.
4. Zhao M, et al. (2009) Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. *Proc Natl Acad Sci USA* 106: 21230–21235.
5. Shi H, et al. (2008) Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. *BMC Biol* 6:45.
6. Purcell S, Cherny SS, Sham PC (2003) Genetic Power Calculator: Design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 19:149–150.
7. Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: Defining, estimating and interpreting F(ST). *Nat Rev Genet* 10:639–650.
8. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4:e72.
9. Sasieni PD (1999) Statistical analysis of the performance of diagnostic tests. *Cytopathology* 10:73–78.
10. Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55: 997–1004.
11. Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 60:155–166.
12. Campbell CD, et al. (2005) Demonstrating stratification in a European American population. *Nat Genet* 37:868–872.
13. Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96:3–12.
14. Quesenberry CP, Hales C (1980) Concentration bands for uniformity plots. *J Stat Comput Simul* 11:41–53.

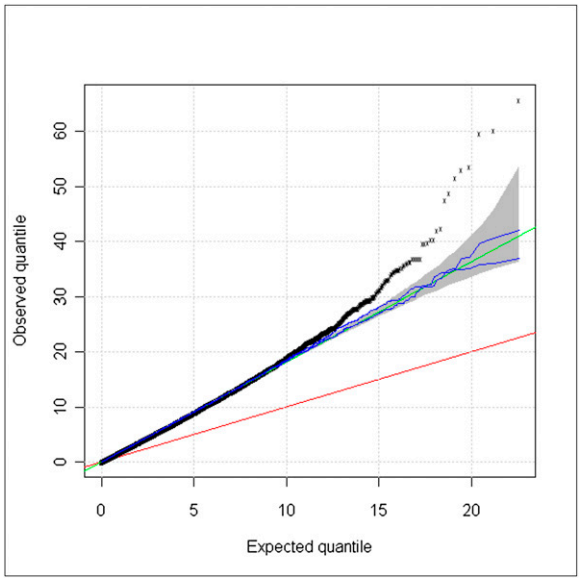

Fig. S1. GWADS (Yunnan Tibetan sample vs. HapMap Phase 3 Han sample) quantile–quantile (QQ) plot applied to all autosomal loci (crossed symbols). The red line indicates where values are expected to fall if there is no population differentiation between these two samples. The green line indicates where values are expected to fall if all loci were subject to the same level of background differentiation between the two samples, using genomic control theory and using the robust median method of Devlin and Roeder (10) to estimate lambda (here, 1.817). The gray area indicates the 95% concentration band around the green line. The two blue lines represent a check on the validity of the genomic control result based on simulation. See *SI Materials and Methods* for further details.

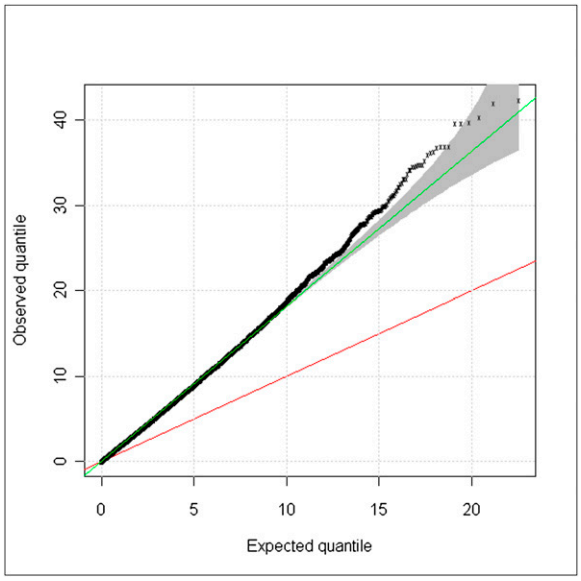

Fig. S2. GWADS (Yunnan Tibetan sample vs. HapMap Phase 3 Han sample) applied to all autosomal loci minus the *EPAS1* region. See Fig. S1 for more information.

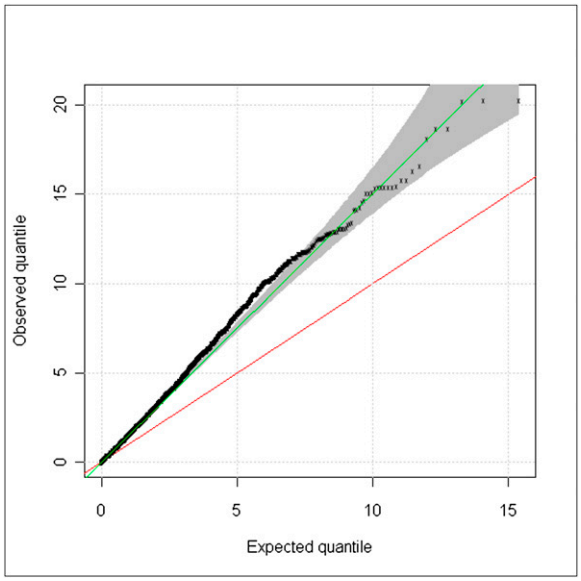

Fig. S3. GWADS (Yunnan Tibetan sample vs. HapMap Phase 3 Han sample) applied to X-chromosome loci only. See Fig. S1 for more information.

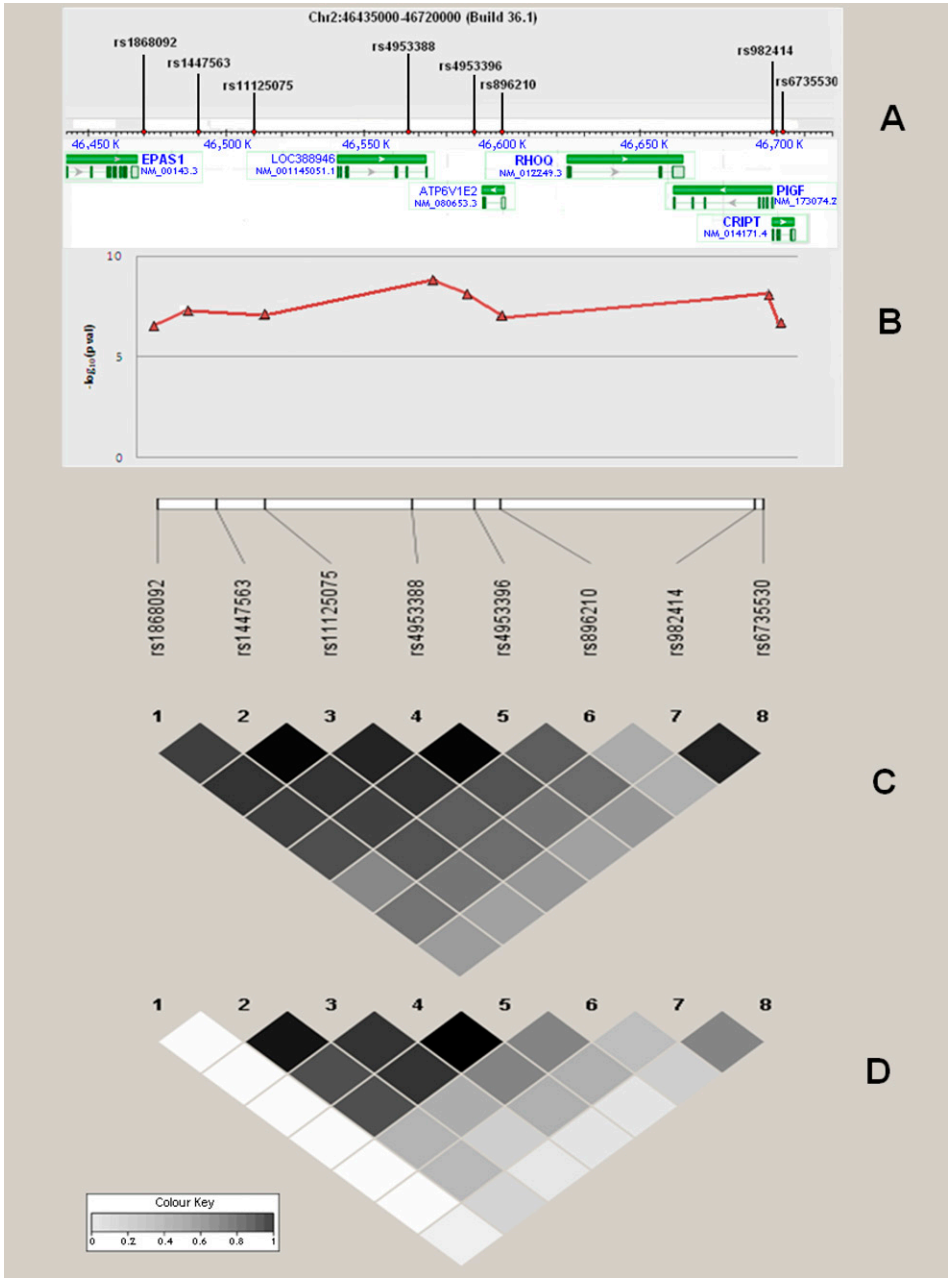

Fig. S4. Overview of chromosome 2 region containing the 8 GWADS significant SNPs. (*A*) Annotation of region showing genomic context. (*B*) $-\log_{10}$ of *P* values from GWADS between Han and Yunnan Tibetan populations. (*C*) Pairwise r^2 disequilibrium heat map in Yunnan sample.

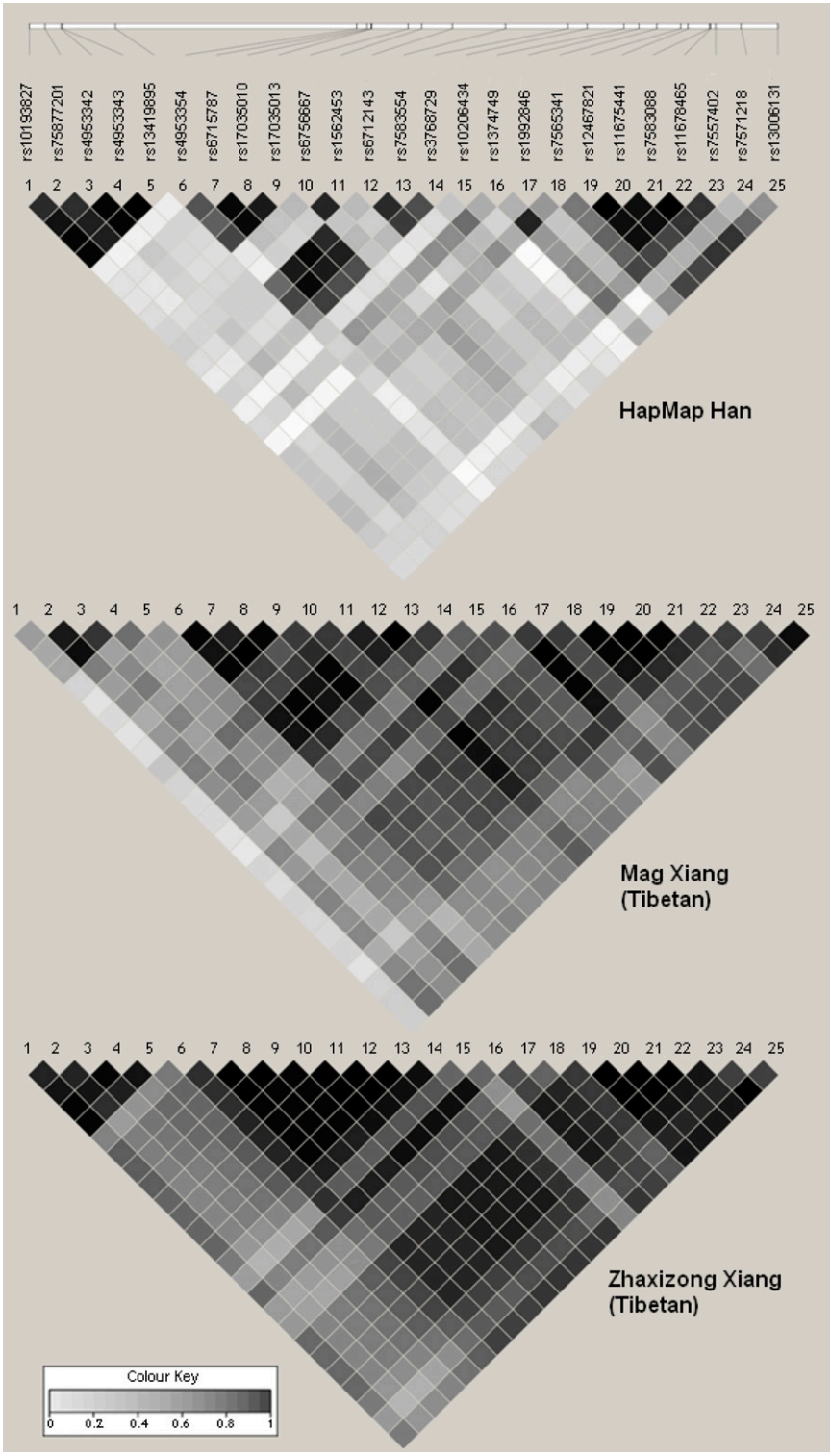

Fig. S5. Pairwise r^2 linkage disequilibrium between *EPAS1* hemoglobin-associating SNPs. Pairwise LD values (r^2) calculated between the 26 variants associating significantly with hemoglobin concentration in both the Mag Xiang and Zhaxizong Xiang populations. Darker shading indicates higher levels of LD.

Other Supporting Information Files

Table S1 (DOC)
Table S2 (DOC)
Table S3 (DOC)
Table S4 (DOC)