Supporting Information (SI) Appendix

Table S1 Genome-wide mean F_{ST} values between Tibetans and other ethnic groups in China.

Population	Sample size	F_{ST}
Yi	10	0.0042
Tu	10	0.0055
Naxi	9	0.0084
Xibo	9	0.0099
Mongola	10	0.0103
Tujia	10	0.0111
Daur	9	0.0129
Han	34	0.0130
Hezhe	9	0.0135
Miao	10	0.0148
Oroqen	10	0.0182
She	10	0.0190
Dai	10	0.0230
Lahu	10	0.0288
Uygur	10	0.0353

The F_{ST} value between the Tibetan population (n = 3,008) and each of the ethnic populations in HGDP was estimated using all genome-wide SNP data. The F_{ST} value presented in the table is an average across all SNPs. These results are consistent with that demonstrated in **Figure 1a**, i.e. Tibetans show the nearest genetic relatedness to the Yi, Tu and Naxi populations.

Table S2 Simulations to investigate the statistical properties of methods for detecting signals of natural selection.

SNP set	Method	Replicate 1		Replicate 2		Replicate 3	
		Mean	SE	Mean	SE	Mean	SE
SNPs under a drift model	LR-GC	0.800	0.006	0.810	0.006	0.807	0.006
	MLMA-LOCO in BOLT-LMM	1.000	0.007	1.007	0.007	1.014	0.007
SNPs under selection	LR-GC	2.796	0.053	2.708	0.049	2.732	0.049
	MLMA-LOCO in BOLT-LMM	2.885	0.052	2.802	0.048	2.884	0.049

Note: SE represents standard error of the mean estimate. We performed a simulation to investigate the statistical properties of the MLMA-LOCO method implemented in BOLT-LMM. We simulated 60,000 unlinked SNPs with their ancestral allele frequencies (p_0) sampled from a uniform distribution, i.e. $U(0.01, 0.99)$. Note that we chose 60,000 because the estimated effective number of independent markers for common SNPs is about 60,000 (1). For the ease of BOLT-LMM analysis, the SNPs were randomly assigned to 22 chromosomes. The allele frequencies of SNPs in derived populations were simulated from a normal distribution, i.e. $p \sim N[p_0, p_0(1-p_0)F_{ST}]$. We generated from this distribution the allele frequencies of two derived populations (pop1 and pop2) with $F_{ST} = 0.01$. We randomly sampled 10% of the SNPs as the loci under selection. For each of SNPs under selection, we simulated selection signal by increasing the allele frequency difference between pop1 and pop2 by 2-fold, i.e. $p_{1(new)} = p_1 + (p_1 - p_2)$ where p_1 and p_2 represent the allele frequencies in pop1 and pop2 respectively. We generated genotypes of each SNP in pop1 (n = 3000) and pop2 (n = 7000) from a binomial distribution, i.e. $x \sim B(2, p)$. We analysed the simulated data using linear regression followed by GC correction (LR-GC) and MLMA-LOCO respectively. In LR-GC, χ^2 test-statistics were divided by the mean χ^2 value of all SNPs. We repeated the simulation three times. We show in the table below that for SNPs under a drift model the mean χ^2 test-statistic from MLMA-LOCO is close to 1 whereas the test-statistics from LR-GC are deflated consistent with the observation in a previous GWAS study (1). For SNPs under selection, the mean test-statistic from MLMA-LOCO is consistently higher than that from LR-GC across three replicates.

Table S3 *P*-values of the 9 genome-wide significant loci in the MLMA-LOCO analyses of the two subsets of data. Four of the loci are genome-wide significant in the analysis of subset 1.

Chr	SNP	bp	Nearest gene	Subset 1	Subset 2	Combined
1	rs1801133	11,856,378	*MTHFR*	1.4E-05	2.0E-07	6.3E-09
1	rs71673426	112,159,304	*RAP1A*	8.6E-07	8.8E-05	1.5E-08
1	rs78720557	198,096,548	*NEK7*	8.5E-07	6.6E-06	4.7E-08
1	rs78561501	231,448,497	*EGLN1*	1.8E-13	3.7E-11	6.1E-15
2	rs116611511	46,600,030	*EPAS1*	1.8E-17	4.0E-17	3.6E-19
4	rs2584462	100,324,464	*ADH7*	1.9E-09	3.5E-06	3.9E-09
5	rs4498258	44,325,322	*FGF10*	6.5E-07	3.8E-07	1.7E-08
6	rs9275281	32,662,920	*HLA-DQB1*	3.3E-09	5.8E-07	1.1E-10
12	rs139129572	123,178,478	*HCAR2*	2.2E-06	8.9E-06	5.8E-09

Subset 1: Seda-Tibetan (*n* = 2,427) vs. Seda-Han + GERA-EAS (*n* = 5,548);

Subset 2: Litang-Tibetan (*n* = 581) vs. Litang-Han + WZ-Han (*n* = 1,736);

Combined: Tibetans (*n* = 3008) and TP-Han + WZ-Han + GERA-EAS (*n* = 7,284).

Table S4 Replication of candidate genes from previous studies

Study	Chr	Candidate Gene	Top SNP in the gene region[#]	MLMA-LOCO P-value
Yi et al. 2010 Science	2	*EPAS1*	rs116611511	3.60E-19
Yi et al. 2010 Science	11	*HBB/HBG2*	rs10768774	0.0076
Yi et al. 2010 Science	1	*DISC1*	rs77401030	2.00E-10
Yi et al. 2010 Science	16	*FANCA*	rs142556882	0.00012
Yi et al. 2010 Science	1	*PKLR*	rs117968195	0.029
Simonson et al. 2010 Science	10	*CYP2E1*	rs4542321	0.0054
Simonson et al. 2010 Science	4	*EDNRA*	rs75056029	0.018
Simonson et al. 2010 Science	19	*ANGPTL4*	rs12978137	0.0086
Simonson et al. 2010 Science	4	*CAMK2D*	rs12512765	0.00021
Simonson et al. 2010 Science	1	*EGLN1*	rs116912442	3.20E-13
Simonson et al. 2010 Science	16	*HMOX2*	chr16:4478105:D	0.00075
Simonson et al. 2010 Science	10	*CYP17A1*	rs141675337	0.0052
Simonson et al. 2010 Science	22	*PPARA*	rs149670586	9.10E-05
Simonson et al. 2010 Science	10	*PTEN*	rs11202571	0.0051

[#]A gene region is defined as ±50Kb of the UTRs.

The total number of SNPs tested in these gene regions is 2,426. The adjusted p-value threshold is 1.5e-4 = 5e-8 * m / 2426 where m (m = 7,276,846) is the total number of SNPs included in the MLMA-LOCO analysis.

Table S5 Descriptive summary of the 91 quantitative traits in Tibetans.

Trait Name	Description	Men (*n* = 1,064)		Women (*n* = 1,785)	
		mean	SD	mean	SD
AVSLPTIM	Average sleep time per day	7.67	1.87	7.54	1.67
HEIGHT	Height	166.68	7.44	155.76	6.86
WEIGHT	Weight	69.38	12.12	60.08	10.42
BMI	Body mass index	24.62	4.17	24.42	3.97
SEOD	Spherical equivalent of the right eye	-0.49	2.45	-0.39	2.35
SEOS	Spherical equivalent of the left eye	-0.45	2.41	-0.37	2.24
LOGMAROD	Uncorrected visual acuity of the right eye	0.30	1.33	0.37	1.22
LOGMAROS	Uncorrected visual acuity of the left eye	0.35	1.36	0.36	1.20
LOGMARODCC	Corrected visual acuity of the right eye	0.15	1.48	0.29	1.45
LOGMAROSCC	Corrected visual acuity of the left eye	0.02	1.71	0.08	1.52
SBP	Systolic blood pressure	126.91	21.30	118.48	22.75
DBP	Diastolic blood pressure	76.10	13.94	73.31	13.17
XL	Heart rate	81.95	14.15	81.72	12.50
CCTOD	Central cornea thickness of the right eye	511.35	34.52	510.05	38.32
ADOD	Anterior chamber depth of the right eye	2.93	0.46	2.89	0.42
LTOD	Lens thickness of the right eye	3.75	0.39	3.69	0.34
ALOD	Axial length of the right eye	23.40	0.91	23.08	0.91
R1OD	Cornea curvature1 of the right eye	7.85	0.26	7.78	0.28
R2OD	Cornea curvature2 of the right eye	7.67	0.26	7.60	0.27
AXISR1OD	The axis of curvature1 of the right eye	97.85	67.51	104.96	67.27
AXISR2OD	The axis of curvature2 of the right eye	81.99	37.41	80.96	34.04
WTWOD	The width of cornea of the right eye	11.71	0.60	11.60	0.65
PUPILLOD	Pupil diameter of the right eye	6.68	1.67	6.70	1.61
CCTOS	Central cornea thickness of the left eye	511.42	38.11	508.23	35.40
ADOS	Anterior chamber depth of left eye	2.95	0.48	2.92	0.45
LTOS	Lens thickness of the left eye	3.73	0.37	3.67	0.34
ALOS	Axial length of the left eye	23.38	0.95	23.04	0.83
R1OS	Cornea curvature1 of the left eye	7.86	0.30	7.78	0.27
R2OS	Cornea curvature2 of the left eye	7.67	0.29	7.59	0.28
AXISR1OS	The axis of curvature1 of the left eye	80.08	66.79	73.78	68.44
AXISR2OS	The axis of curvature2 of the left eye	94.05	37.78	94.63	32.68
WTWOS	The width of cornea of the left eye	11.73	0.67	11.62	0.63
PUPILLOS	Pupil diameter of the left eye	6.65	1.62	6.65	1.58
FT3	Free triiodothyronine	5.67	1.18	5.38	1.94
FT4	Free thyroxine	18.47	3.57	18.90	8.01
TSH	Thyroid Stimulating Hormone	3.47	3.98	3.71	5.17
PTH	Parathyroid Hormone	28.57	19.53	33.44	21.81
B12	Vitamin B12	473.30	261.62	515.00	284.99
FOLATE	Folate	6.29	2.41	7.36	2.71
ALT	Glutamate pyruvate transaminase	27.45	28.11	18.41	17.38
AST	Glutamic oxalacetic transaminase	25.78	12.92	22.65	12.95
AST2ALT	AST/ALT ratio	1.20	0.65	1.48	0.88
TP	Total protein	77.20	5.08	77.35	5.36
ALB	Albumin	47.62	3.35	46.85	2.85
GLB	Globulin	29.55	4.17	30.45	4.69
A2G	ALB/GLB ratio	1.72	1.57	1.64	1.38
TBIL	Total bilirubin	13.33	7.48	10.38	6.96
DBIL	Direct bilirubin	5.73	2.65	4.54	3.11
IBIL	Indirect bilirubin	7.63	5.16	5.89	4.84
ALP	Alkaline phosphatase	94.70	35.22	85.05	30.95
GGT	Gamma-glutamyl transpeptidase	49.25	50.19	29.98	32.67
GLU	Glucose	5.43	1.79	5.04	1.03
UREA	Urea nitrogen	4.83	1.79	4.56	1.70
CRE	Creatinine	76.68	14.81	59.44	10.86
UA	Uric acid	394.93	82.99	298.10	73.26
TG	Triglyceride	1.18	0.66	1.03	0.47
TCH	Total cholesterol	4.60	1.08	4.51	1.02
HDL	High density lipoprotein	1.20	0.23	1.35	0.28
LDL	Low density lipoprotein	2.89	0.86	2.61	0.76
K	Potassium	4.30	0.47	4.35	0.45

NA	Sodium	140.93	3.90	140.39	3.32
CL	Aluminium	106.85	3.34	107.16	2.83
CA	Calcium	2.39	0.13	2.34	0.12
PHOS	Phosphorus	1.14	0.19	1.22	0.16
FE	Ferrum	18.89	8.94	14.98	8.55
FER	Ferritin	207.85	185.30	83.36	117.91
HCY	Homocysteine	27.17	24.15	18.64	7.86
HBA1C	Glycosylated hemoglobin	5.46	0.83	4.92	0.68
WBC	White blood cell count	6.18	1.89	6.19	1.87
LYMPH	Lymphocyte count	1.87	0.62	1.92	0.67
MID1	Intermediate cell count	0.35	0.15	0.33	0.19
GRAN1	Neutrophile granulocyte	3.96	1.63	3.94	1.60
LYMPH1	Lymphocyte percentage	31.40	8.84	32.13	9.54
MID2	Intermediate cell percentage	5.94	2.11	5.67	2.93
GRAN2	Neutrophile granulocyte percentage	62.66	9.07	62.20	10.25
RBC	Red blood cell count	5.19	0.73	4.75	0.64
HGB	Hemoglobin	171.09	25.87	149.92	26.01
MCHC	Mean Corpuscular Hemoglobin Concentration	349.23	18.05	343.59	13.19
MCV	Mean Corpuscular Volume	94.63	5.94	91.77	8.21
MCH	Mean Corpuscular Hemoglobin	33.04	2.98	31.55	3.55
RDW_CV	Coefficient Of Variation of Red blood Cell	14.64	1.26	14.78	1.67
HCT	Hematocrit	49.04	7.41	43.48	6.85
PLT	Platelets	213.05	57.49	252.74	75.11
MPV	Mean Platelet Volume	8.33	0.87	8.51	0.91
PDW	Platelet Distribution Width	15.97	0.30	15.90	0.31
PCT	Thrombocytocrit	0.18	0.04	0.21	0.05
RDW_SD	Standard deviation of red blood cell distribution	49.91	4.28	48.66	4.48
VITD	Vitamin D	19.75	7.58	17.05	5.99
CDOD	Cup/disc ratio of the right eye	0.43	0.16	0.37	0.15
IOPOD	Intraocular pressure of the right eye	15.23	3.71	15.56	3.66
IOPOS	Intraocular pressure of the left eye	15.36	3.53	15.67	3.68

Table S6 Estimates of heritability for 91 quantitative traits in Tibetans. The estimates are from GCTA-GREML analyses (2, 3) in 3,008 Tibetan subjects. For GWAS data with related individuals, we used the strategy described in Zaitlen et al. (4) to estimate pedigree-based heritability (h^2) and SNP-based heritability (h^2_{SNP}) (i.e. variance explained by all SNPs in unrelated individuals) simultaneously in a model (see http://gcta.freeforums.net/thread/241/gcta-greml-analysis-family-data for details about the GCTA commands used). SE: standard error.

Trait	h^2	SE	h^2_{SNP}	SE
AVSLPTIM	0.119	0.107	0.011	0.147
HEIGHT	0.653	0.076	0.459	0.104
WEIGHT	0.444	0.079	0.296	0.103
BMI	0.331	0.080	0.254	0.107
SEOD	0.044	0.085	0.008	0.109
SEOS	0.211	0.087	0.178	0.115
LOGMAROD	0.155	0.087	0.151	0.116
LOGMAROS	0.128	0.086	0.120	0.116
LOGMARODCC	0.190	0.199	0.007	0.286
LOGMAROSCC	0.414	0.178	0.390	0.265
SBP	0.380	0.075	0.371	0.102
DBP	0.362	0.075	0.338	0.099
XL	0.346	0.084	0.343	0.110
CCTOD	0.317	0.165	0.302	0.222
ADOD	0.024	0.167	0.000	0.222
LTOD	0.469	0.176	0.492	0.262
ALOD	0.366	0.183	0.248	0.239
R1OD	0.394	0.175	0.168	0.243
R2OD	0.489	0.166	0.376	0.237
AXISR1OD	0.410	0.184	0.446	0.262
AXISR2OD	0.385	0.176	0.393	0.257
WTWOD	0.165	0.168	0.000	0.244
PUPILLOD	0.522	0.159	0.273	0.232
CCTOS	0.491	0.166	0.227	0.229
ADOS	0.354	0.163	0.099	0.224
LTOS	0.379	0.194	0.173	0.270
ALOS	0.232	0.174	0.213	0.242
R1OS	0.221	0.177	0.012	0.238
R2OS	0.257	0.169	0.083	0.229
AXISR1OS	0.239	0.180	0.230	0.249
AXISR2OS	0.213	0.175	0.131	0.253
WTWOS	0.000	0.159	0.000	0.247
PUPILLOS	0.520	0.165	0.012	0.225
FT3	0.054	0.084	0.000	0.111
FT4	0.182	0.086	0.074	0.108
TSH	0.291	0.090	0.069	0.112
PTH	0.336	0.082	0.313	0.108
B12	0.544	0.074	0.540	0.104
FOLATE	0.416	0.082	0.408	0.110
ALT	0.112	0.075	0.053	0.102
AST	0.130	0.078	0.129	0.102
AST2ALT	0.211	0.083	0.000	0.110
TP	0.135	0.076	0.067	0.102
ALB	0.185	0.080	0.081	0.103
GLB	0.229	0.079	0.105	0.107
A2G	0.224	0.078	0.122	0.107
TBIL	0.321	0.080	0.270	0.103
DBIL	0.442	0.080	0.315	0.105
IBIL	0.243	0.081	0.195	0.105
ALP	0.120	0.081	0.000	0.106
GGT	0.205	0.076	0.079	0.103

GLU	0.171	0.080	0.141	0.107
UREA	0.248	0.078	0.127	0.102
CRE	0.352	0.080	0.243	0.107
UA	0.267	0.077	0.116	0.101
TG	0.279	0.074	0.258	0.100
TCH	0.383	0.077	0.268	0.106
HDL	0.141	0.077	0.051	0.102
LDL	0.345	0.078	0.237	0.104
K	0.000	0.076	0.000	0.099
NA.	0.350	0.076	0.259	0.106
CL	0.286	0.075	0.159	0.101
CA	0.309	0.080	0.150	0.107
PHOS	0.286	0.079	0.137	0.103
FE	0.203	0.081	0.193	0.104
FER	0.313	0.069	0.305	0.094
HCY	0.219	0.077	0.162	0.103
HBA1C	0.407	0.073	0.109	0.101
WBC	0.321	0.080	0.316	0.113
LYMPH	0.367	0.081	0.206	0.110
MID1	0.431	0.082	0.289	0.111
GRAN1	0.292	0.081	0.240	0.113
LYMPH1	0.304	0.084	0.212	0.111
MID2	0.260	0.086	0.081	0.109
GRAN2	0.389	0.084	0.260	0.112
RBC	0.340	0.082	0.239	0.110
HGB	0.206	0.083	0.150	0.109
MCHC	0.337	0.082	0.218	0.108
MCV	0.184	0.080	0.000	0.107
MCH	0.167	0.081	0.000	0.109
RDW_CV	0.125	0.081	0.052	0.104
HCT	0.234	0.082	0.166	0.108
PLT	0.271	0.082	0.173	0.109
MPV	0.462	0.082	0.332	0.105
PDW	0.252	0.081	0.167	0.111
PCT	0.294	0.079	0.253	0.110
RDW_SD	0.367	0.081	0.255	0.107
VITD	0.446	0.079	0.352	0.112
CDOD	0.345	0.193	0.042	0.262
IOPOD	0.202	0.085	0.180	0.116
IOPOS	0.244	0.087	0.139	0.116

Table S7 Associations of the *MTHFR* and *EPAS1* loci with 5 quantitative traits in Tibetans. A1: the effect allele. A2: the other allele. ***b***: effect size of the SNP on trait in standard deviation units.

Trait	Chr	SNP	Nearest gene	bp	A1	A2	*b*	SE	P-value
FOLATE	1	rs1801133	*MTHFR*	11856378	A	G	-0.34	0.032	6.5E-27
HCY	1	rs1801133	*MTHFR*	11856378	A	G	0.54	0.031	1.1E-69
RBC	2	rs116611511	*EPAS1*	46600030	G	A	-0.15	0.027	2.4E-08
HGB	2	rs116611511	*EPAS1*	46600030	G	A	-0.11	0.027	7.7E-05
HCT	2	rs116611511	*EPAS1*	46600030	G	A	-0.12	0.027	5.6E-06

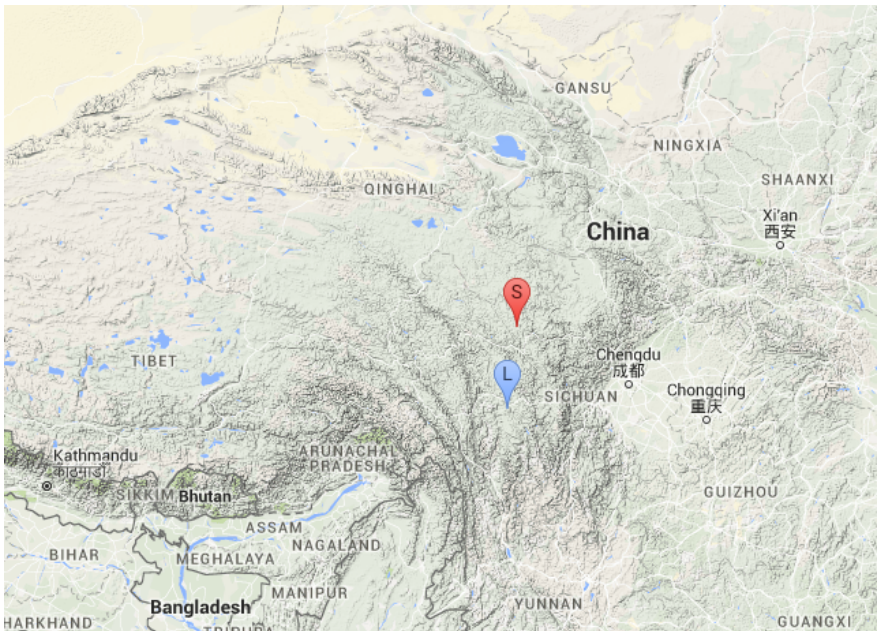


Figure S1 A map of western China. The figure is generated using the R package *RgoogleMaps*. The two sites labeled in red and blue are Seda (S) and Litang (L), respectively. Both sites are ~4100m above sea level.

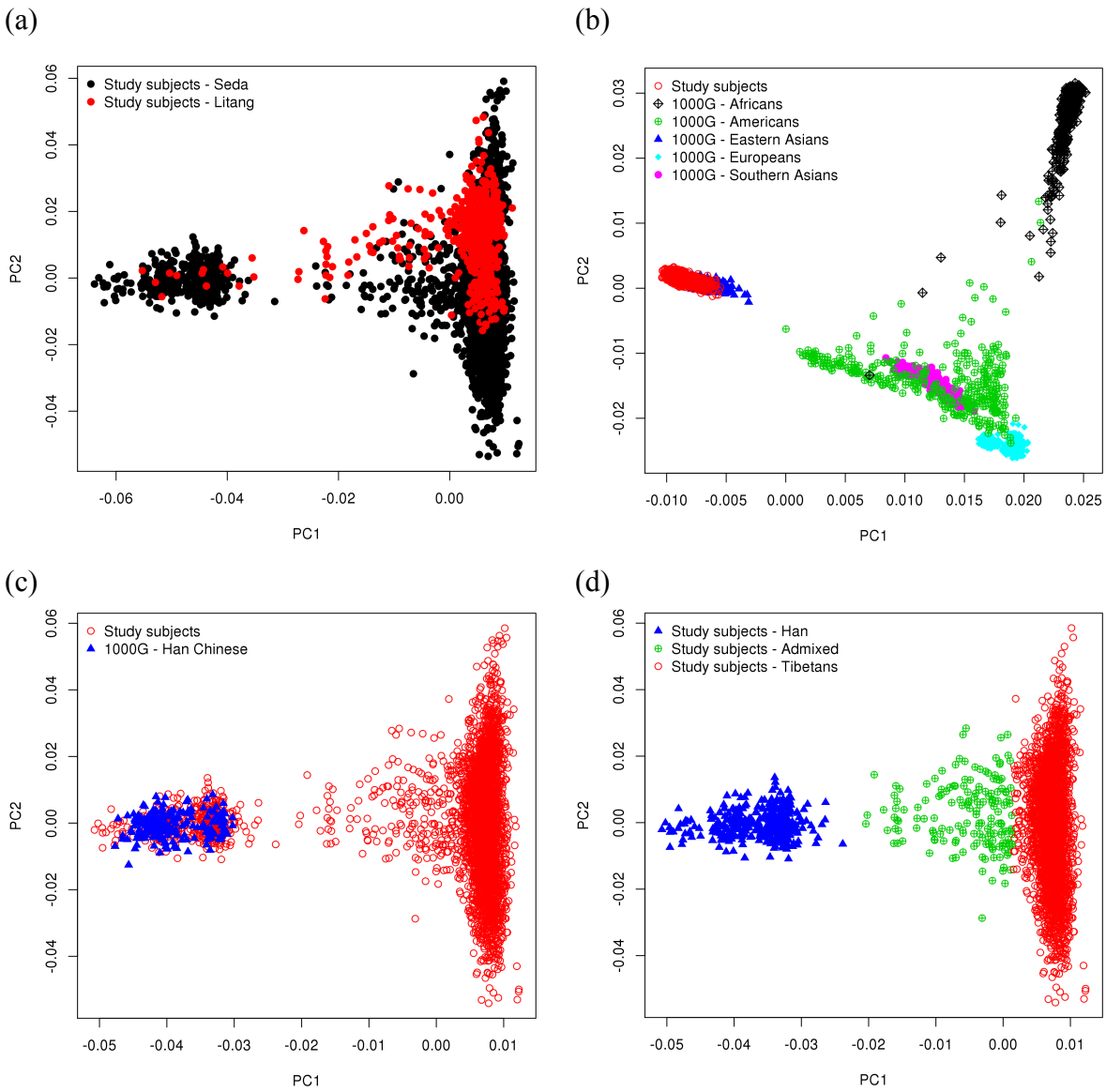


Figure S2 Genetic ancestry of the subjects collected from the Tibetan Plateau (TP). Shown are the results from principal component (PC) analyses using all autosomal genotyped SNPs after QC (**Materials & Methods**). PC1 and PC2 represent the first and the second PCs, respectively. In panel (a), shown are the PC plots for subjects collected from two sites, Seda (in black) and Litang (in red). In panel (b), PCs from our study subjects are projected to those from the 1000 Genomes Project (1000G). The result shows that Tibetans are genetically distinct from Southern Asians (Indians, Pakistanis, Bangladeshis and Sri Lankan Tamils) and Americans (Mexican, Puerto Ricans, Colombians and Peruvians), consistent with Tibetans sharing the most recent common ancestor with one of the Eastern Asian populations. In panel (c), PCs for all the study subjects from TP are projected to the Han Chinese subjects from 1000G (1000G-Han). In panel (d), the study subjects are stratified to Han, "admixed" and Tibetans using the method described in **Materials & Methods**.

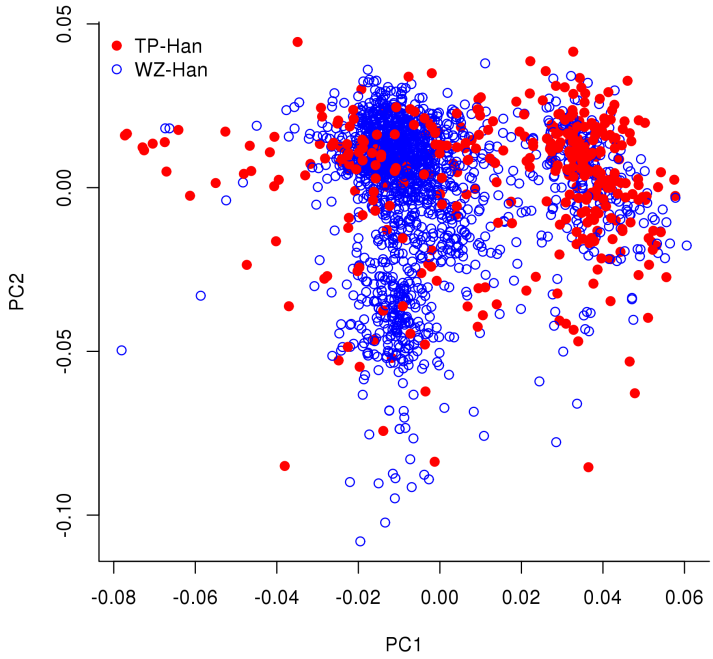


Figure S3 Principal component analysis of Han subjects collected from TP and Wenzhou. The analysis was performed using all genotyped SNPs in all the Han subjects. TP-Han: Han subjects collected from the Tibetan Plateau. WZ-Han: Han subjects collected from Wenzhou.

a) Before removing the ancestry outliers b) After removing the ancestry outliers

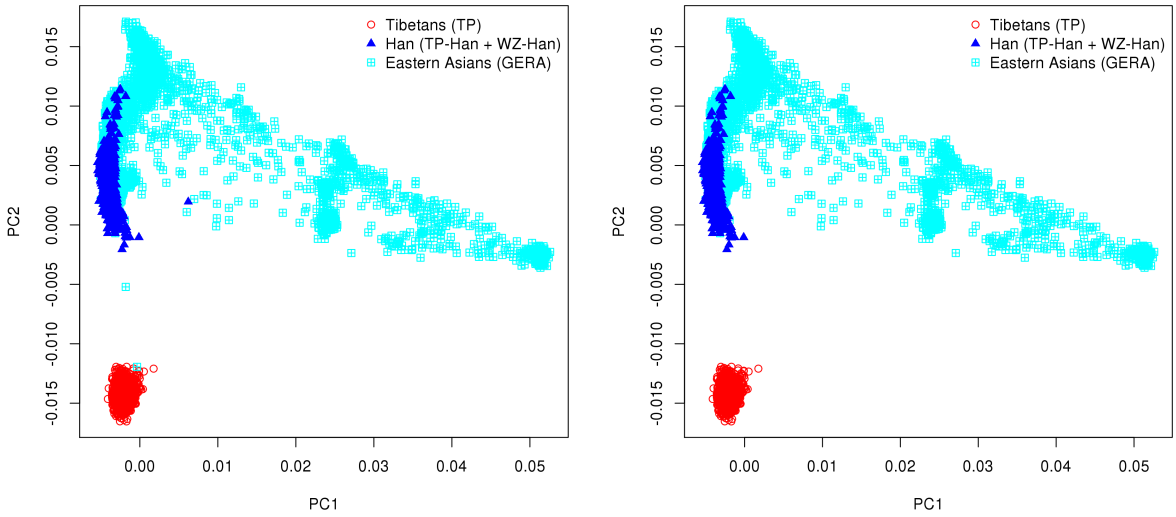


Figure S4 Genetic ancestry of Tibetans and Han in China and Eastern Asians in the US. The principal component (PC) analysis was performed using ~1 million HapMap3 SNPs (MAF ≥ 0.01) from the 1000G-imputed data after QC (**Materials & Methods**). TP: Tibetan Plateau. WZ: Wenzhou. GERA: Genetic Epidemiology Research on Aging study in the US.

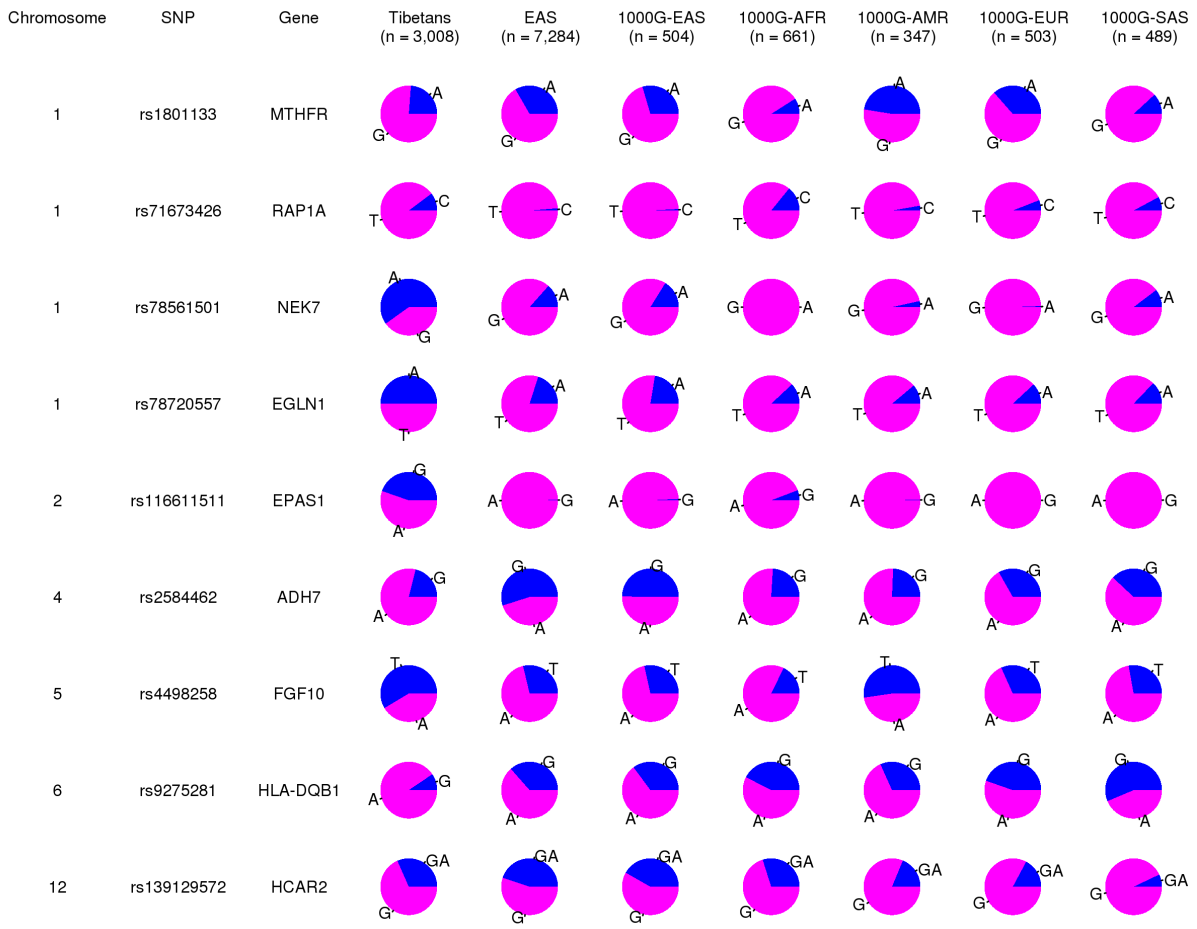
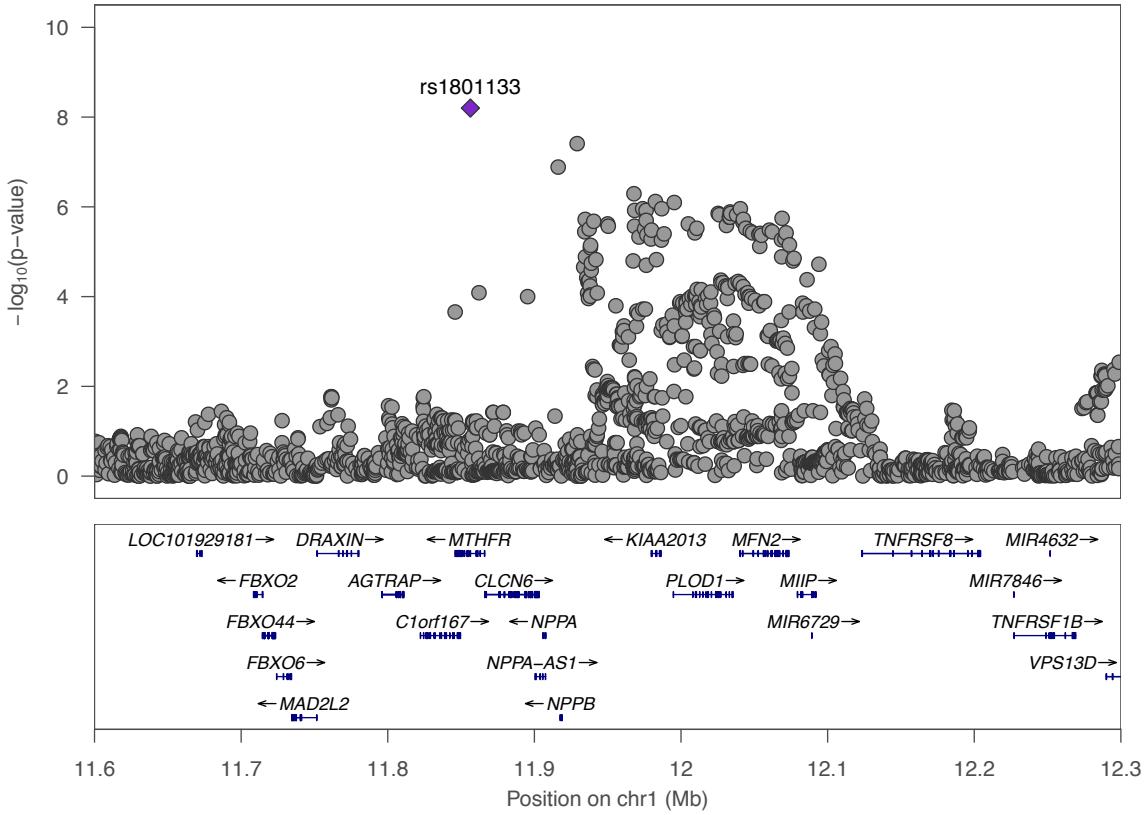
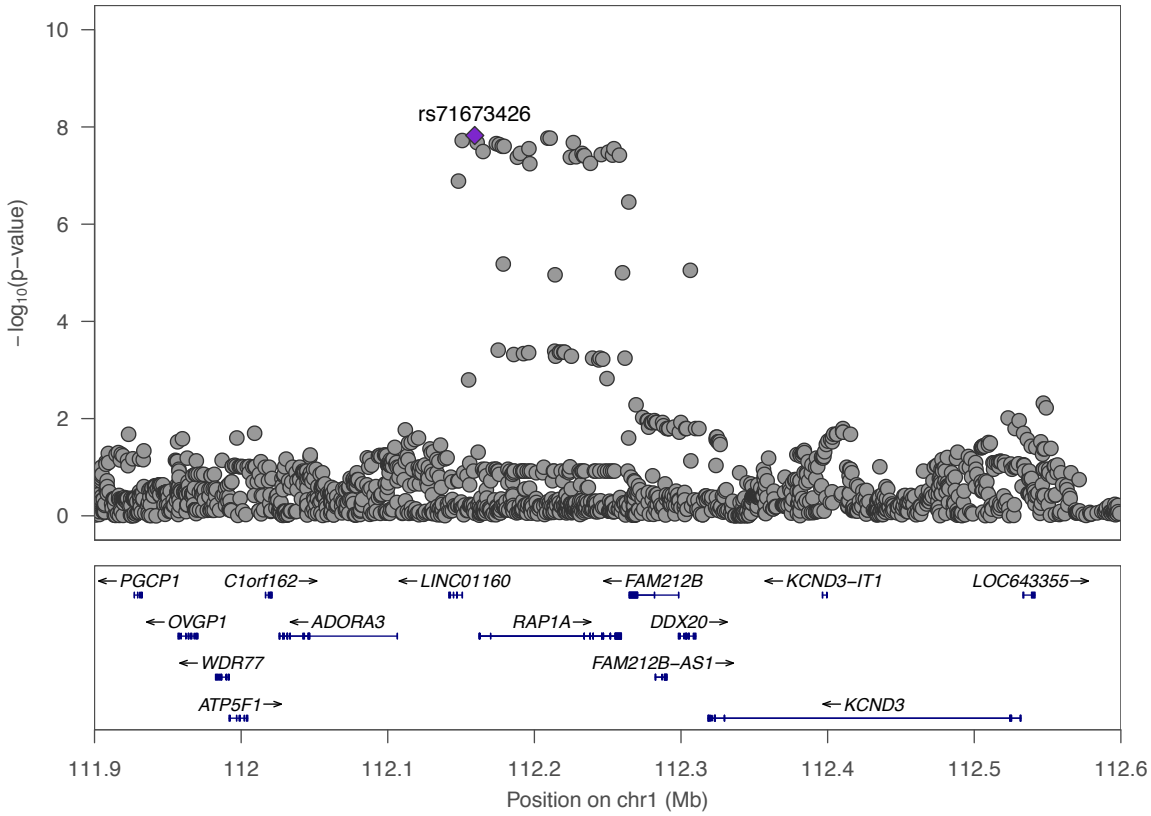


Figure S5 Allele frequencies of the top SNPs at the 9 genome-wide significant gene loci in different populations. The 9 SNPs were identified from the MLMA-LOCO analysis of testing for difference in allele frequency between Tibetans and EAS. Gene: the nearest gene to the top SNP at each locus except *EGLN1*. AFR: Africans. AMR: Americans. EUR: Europeans. SAS: Southern Asians.

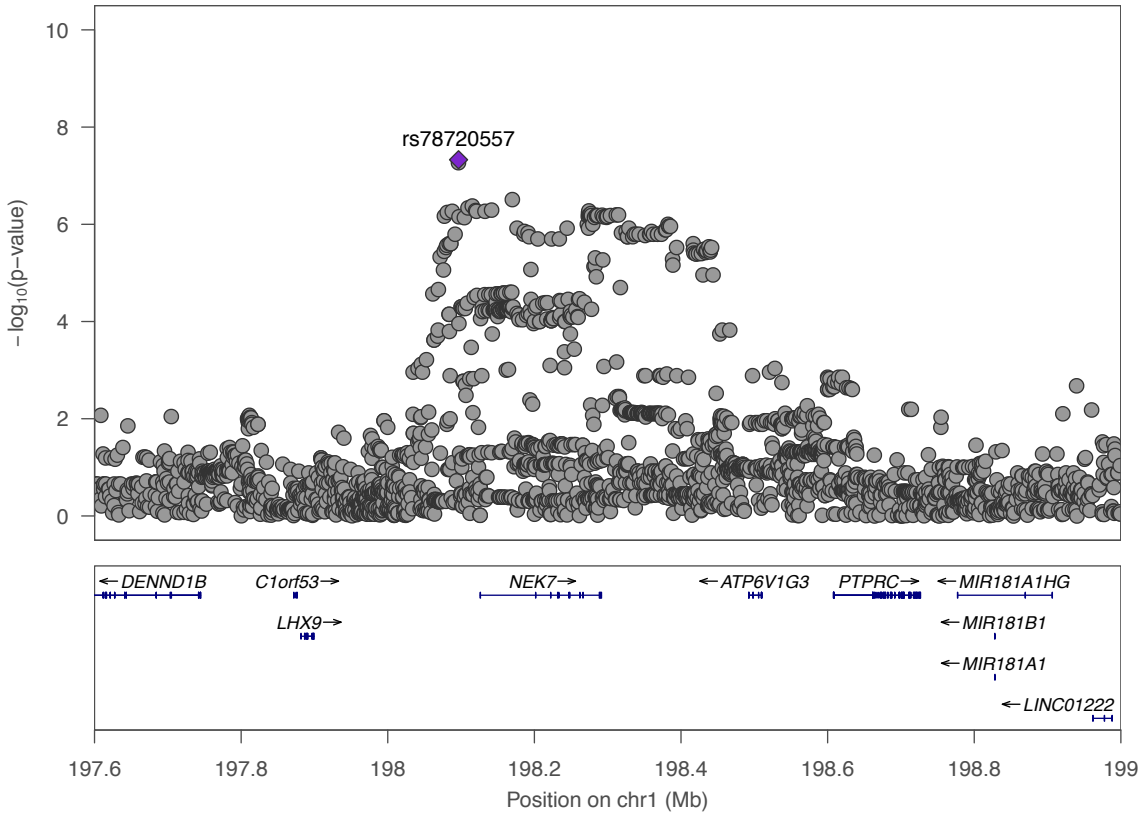
a)


rs1801133
−log10(p−value)
10
8
6
4
2
0
LOC101929181→
DRAXIN→
←MTHFR
←KIAA2013
MFN2→
TNFRSF8→
MIR4632→
←FBXO2
AGTRAP→
CLCN6→
PLOD1→
MIIP→
MIR7846→
FBXO44→
C1orf167→
←NPPA
MIR6729→
TNFRSF1B→
FBXO6→
NPPA−AS1→
VPS13D→
←MAD2L2
←NPPB
11.6
11.7
11.8
11.9
12
12.1
12.2
12.3
Position on chr1 (Mb)

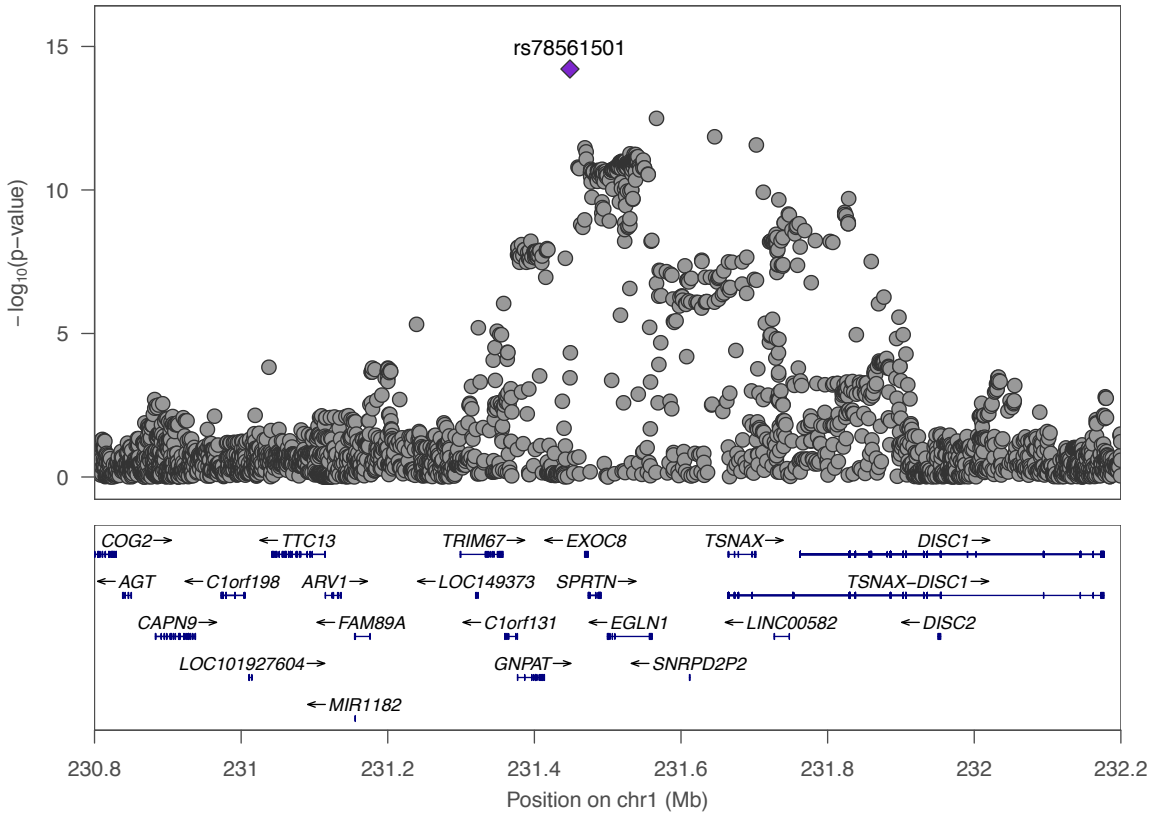
b)


rs71673426
−log10(p−value)
10
8
6
4
2
0
←PGCP1
C1orf162→
←LINC01160
←FAM212B
←KCND3−IT1
LOC643355→
←OVGP1
←ADORA3
RAP1A→
DDX20→
←WDR77
FAM212B−AS1→
ATP5F1→
←KCND3
111.9
112
112.1
112.2
112.3
112.4
112.5
112.6
Position on chr1 (Mb)

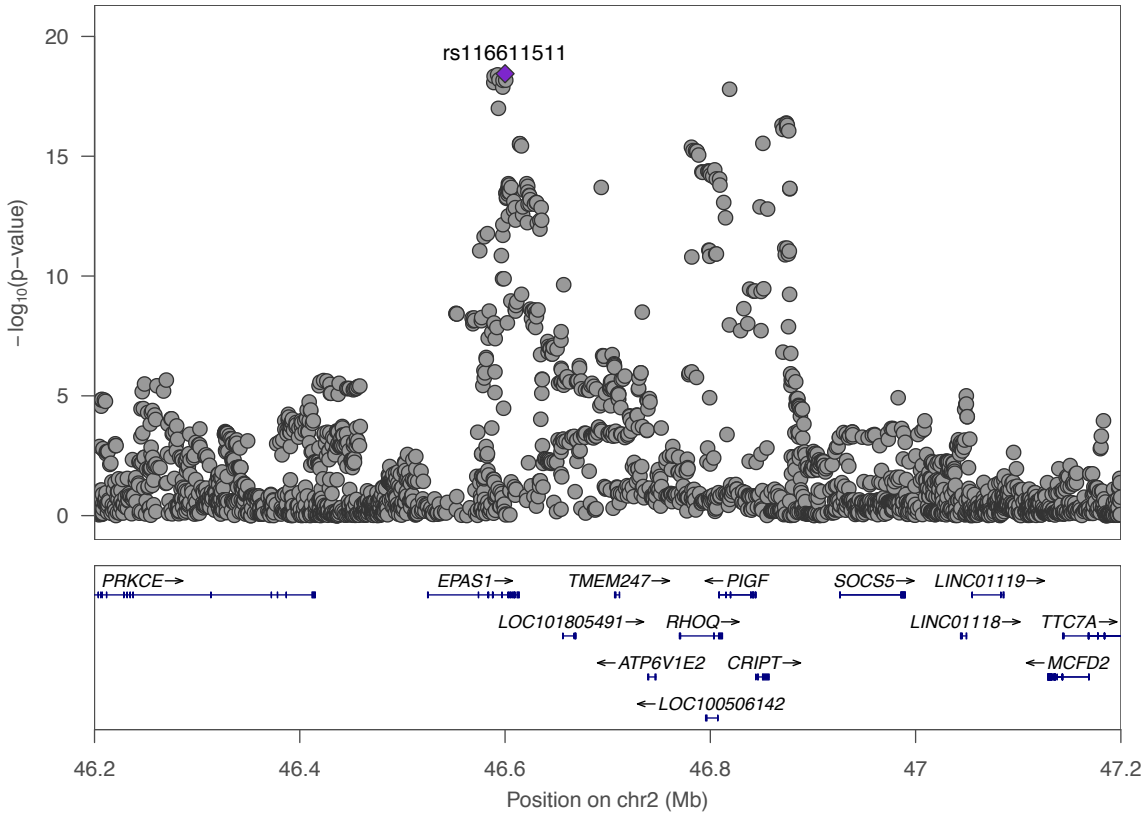
c)


rs78720557
−log10(p−value)
10
8
6
4
2
0
←DENND1B
C1orf53→
NEK7→
←ATP6V1G3
PTPRC→
←MIR181A1HG
LHX9→
←MIR181B1
←MIR181A1
←LINC01222
197.6
197.8
198
198.2
198.4
198.6
198.8
199
Position on chr1 (Mb)

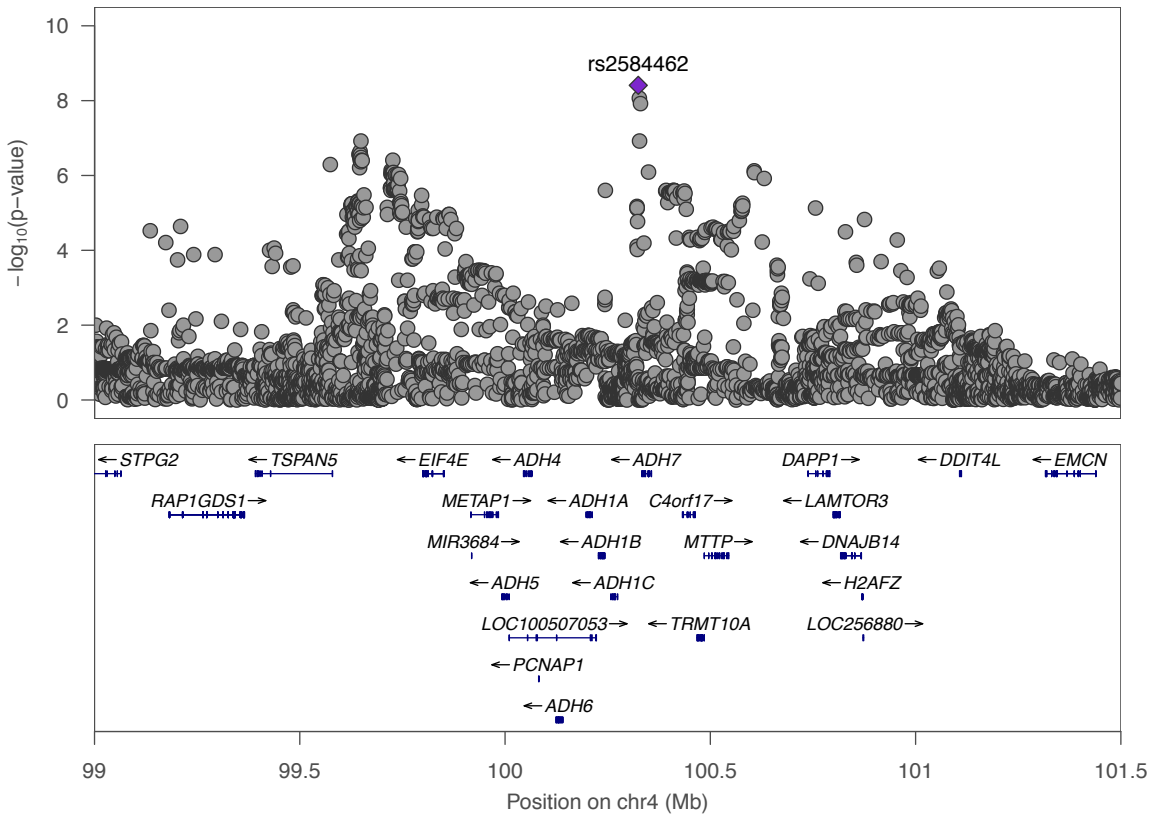
d)


rs78561501
−log10(p−value)
15
10
5
0
COG2→
←TTC13
TRIM67→
←EXOC8
TSNAX→
DISC1→
←AGT
←C1orf198
ARV1→
←LOC149373
SPRTN→
TSNAX−DISC1→
CAPN9→
←FAM89A
←C1orf131
←EGLN1
←LINC00582
←DISC2
LOC101927604→
GNPAT→
←SNRPD2P2
←MIR1182
230.8
231
231.2
231.4
231.6
231.8
232
232.2
Position on chr1 (Mb)

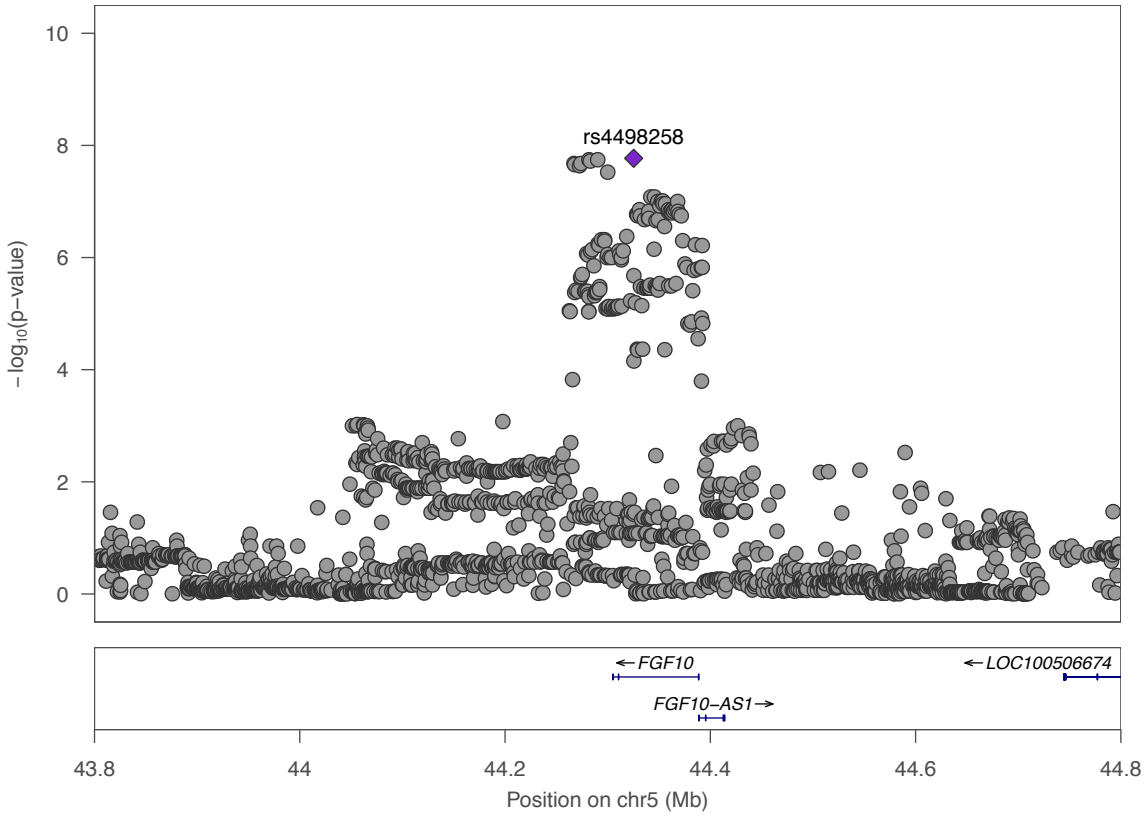
e)


rs116611511
20
15
10
5
0
– log10(p–value)
PRKCE→
EPAS1→
TMEM247→
←PIGF
SOCS5→
LINC01119→
LOC101805491→
RHOQ→
LINC01118→
TTC7A→
←ATP6V1E2
CRIPT→
←MCFD2
←LOC100506142
46.2
46.4
46.6
46.8
47
47.2
Position on chr2 (Mb)

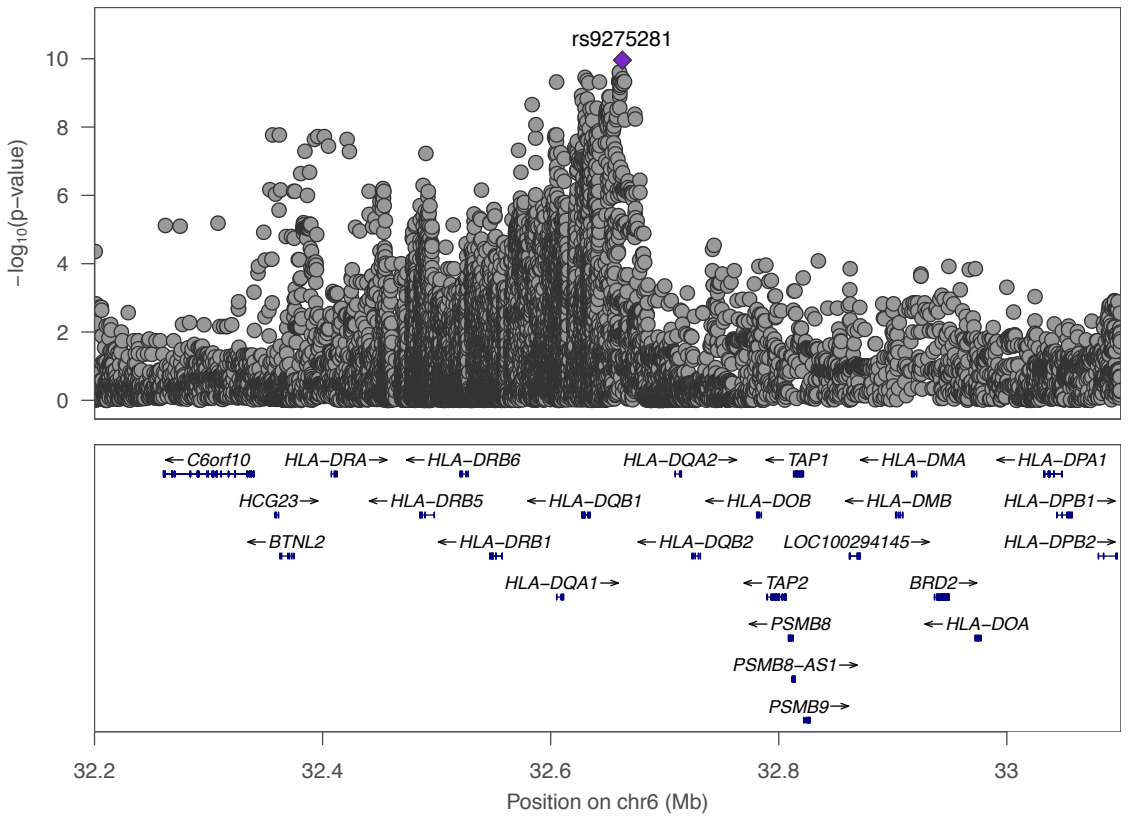
f)


rs2584462
10
8
6
4
2
0
– log10(p–value)
←STPG2
←TSPAN5
←EIF4E
←ADH4
←ADH7
DAPP1→
←DDIT4L
←EMCN
RAP1GDS1→
METAP1→
←ADH1A
C4orf17→
←LAMTOR3
MIR3684→
←ADH1B
MTTP→
←DNAJB14
←ADH5
←ADH1C
←H2AFZ
LOC100507053→
←TRMT10A
LOC256880→
←PCNAP1
←ADH6
99
99.5
100
100.5
101
101.5
Position on chr4 (Mb)

g)


rs4498258
10
8
6
4
2
0
−log10(p−value)
←FGF10
←LOC100506674
FGF10−AS1→
43.8
44
44.2
44.4
44.6
44.8
Position on chr5 (Mb)

h)


rs9275281
10
8
6
4
2
0
−log10(p−value)
←C6orf10
HLA−DRA→
←HLA−DRB6
HLA−DQA2→
←TAP1
←HLA−DMA
←HLA−DPA1
HCG23→
←HLA−DRB5
←HLA−DQB1
←HLA−DOB
←HLA−DMB
HLA−DPB1→
←BTNL2
←HLA−DRB1
←HLA−DQB2
LOC100294145→
HLA−DPB2→
HLA−DQA1→
←TAP2
BRD2→
←PSMB8
←HLA−DOA
PSMB8−AS1→
PSMB9→
32.2
32.4
32.6
32.8
33
Position on chr6 (Mb)

i)

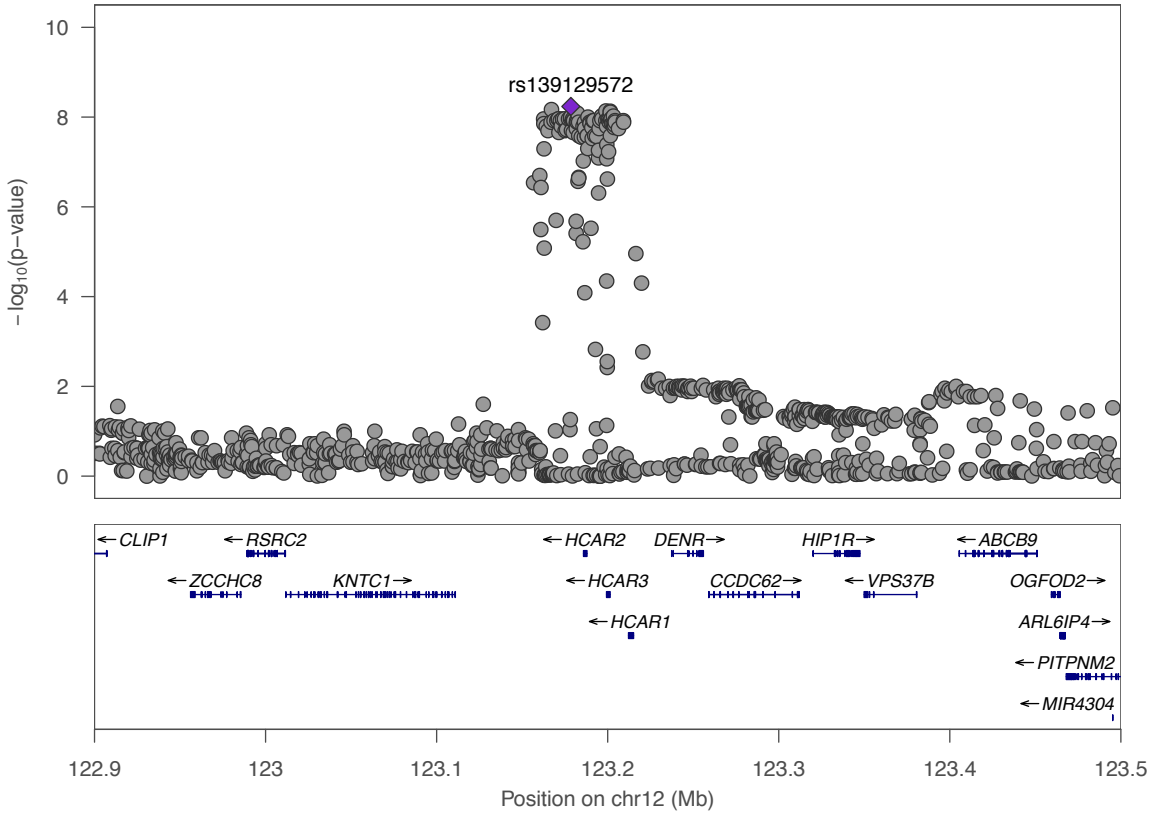


Figure S6 Plot of –log10(p-value) against physical distance at each of the 9 loci. The p-values were calculated from the MLMA-LOCO analysis of testing for difference in allele frequency between Tibetans and Eastern Asians. The plot was generated using the online tool LocusZoom (5) (http://locuszoom.sph.umich.edu/locuszoom/).

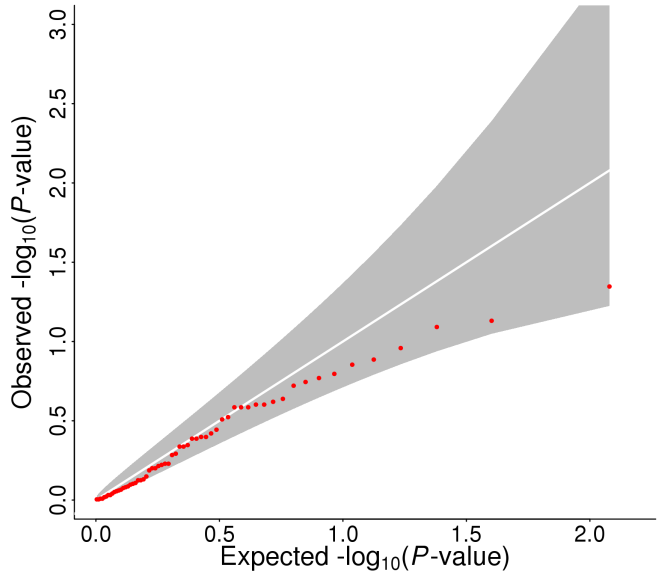


Figure S7 MLMA-LOCO analysis to detect genetic signatures of high-altitude adaptation on the mitchondrial genome. The analysis was performed in Han (WZ_Han + TP_Han, *n* = 2,099) and Tibetans (*n* = 3,008) using 60 genotyped mitchondrial SNPs after QC. All the autosomal SNPs were fitted in the model as random effects to control for population structure (**Materials & Methods**).

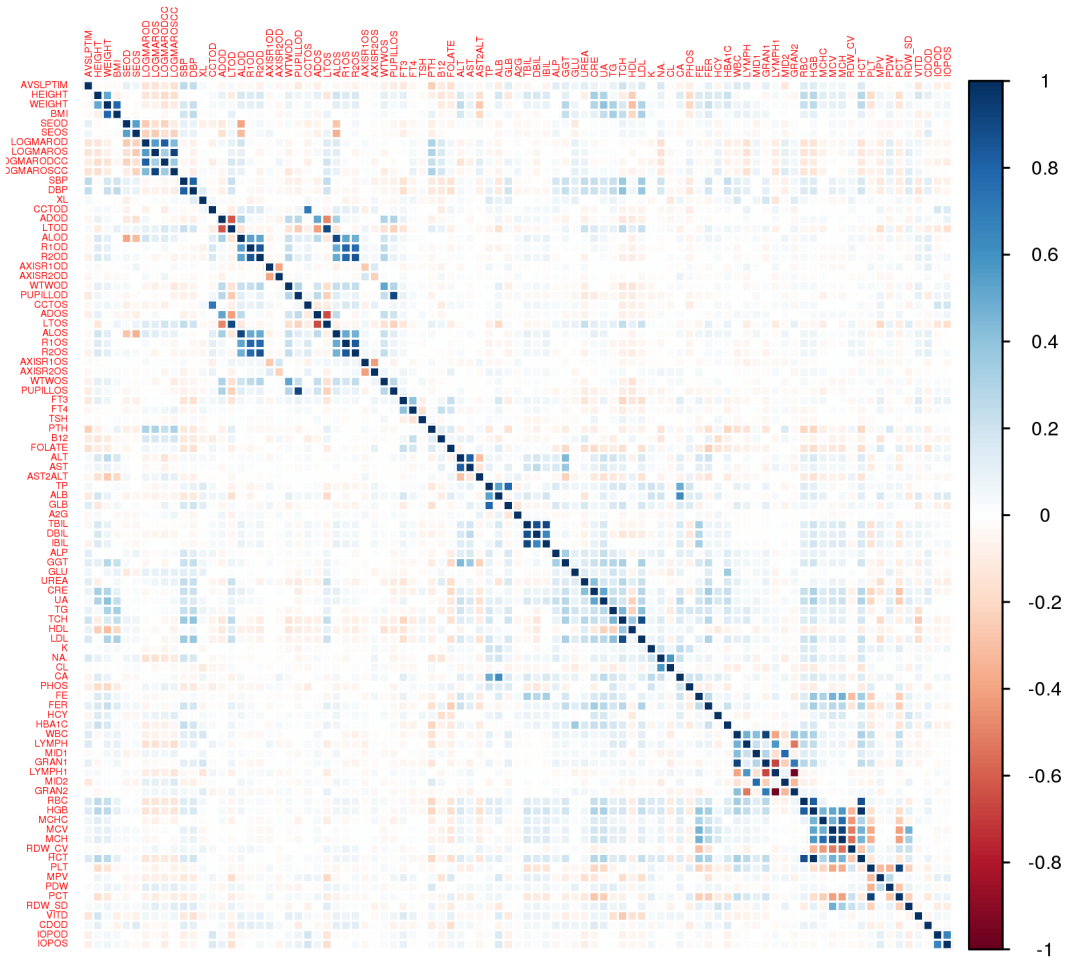


Figure S8 Phenotypic correlation matrix for 91 quantitative traits in Tibetans. A list of full names of the traits can be found in **Table S5**.

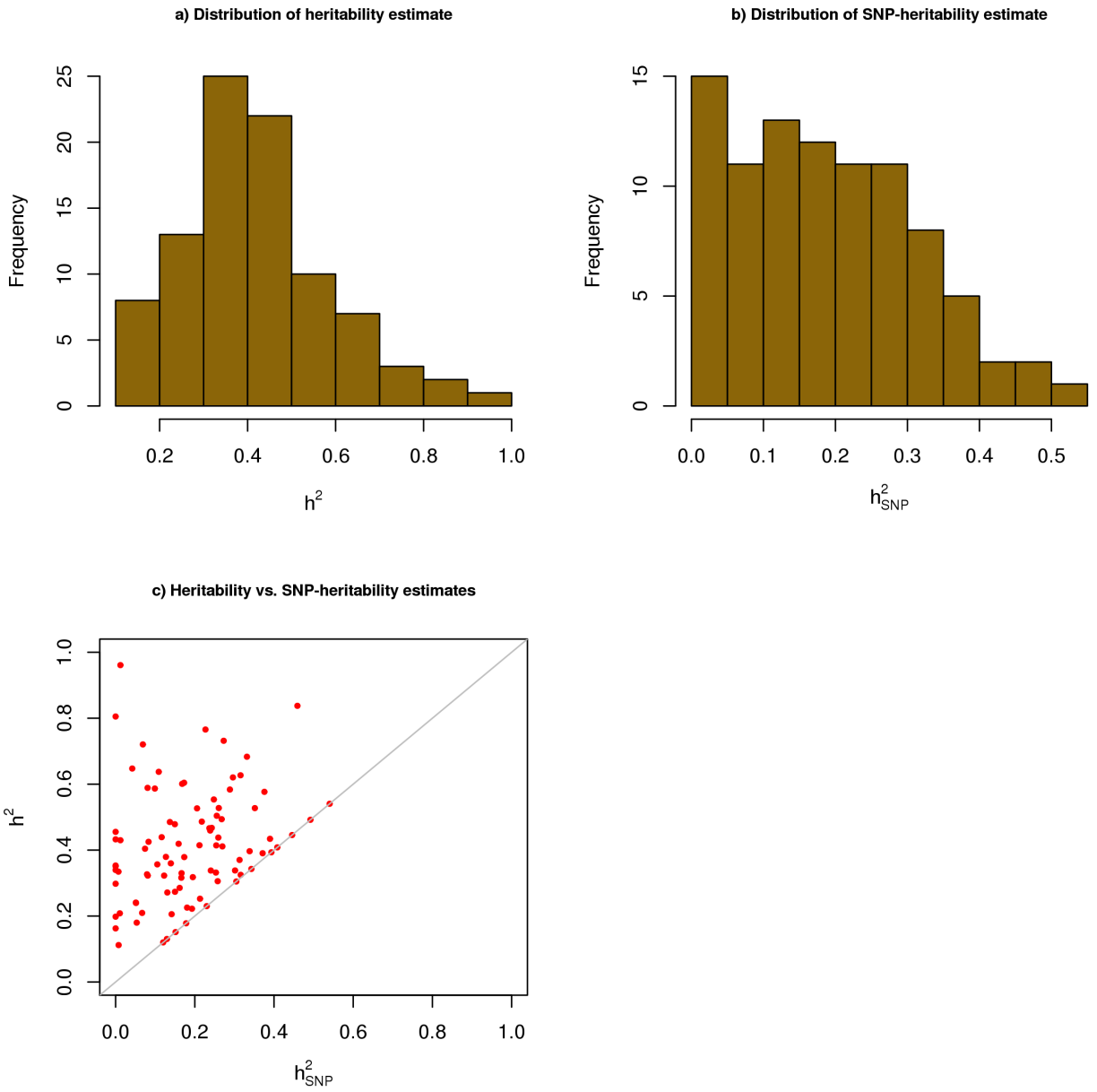
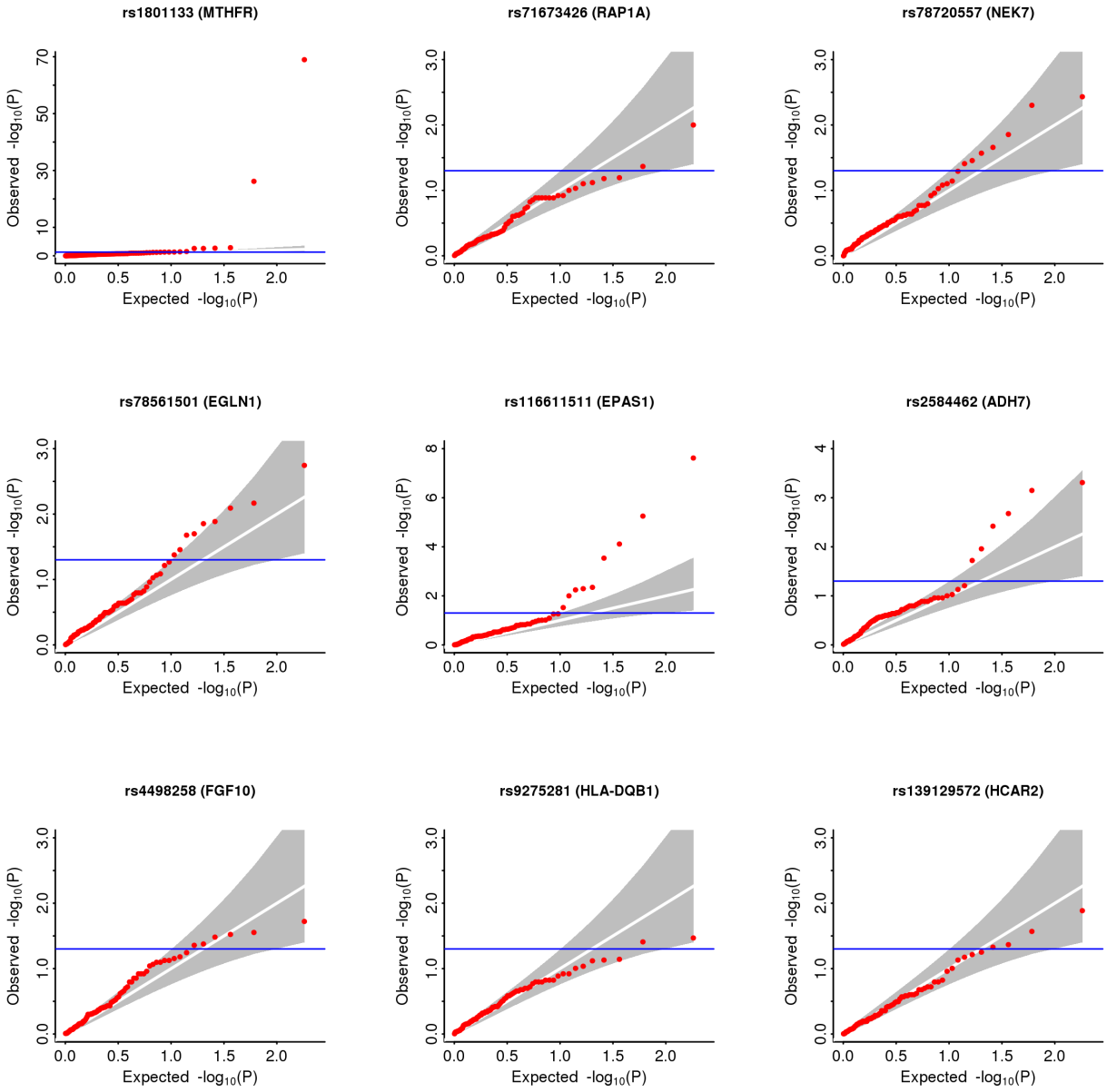


Figure S9 Estimation of heritability for 91 quantitative traits in Tibetans. The estimates are from GCTA-GREML analyses (2, 3) in 3,008 Tibetan subjects. For GWAS data with related individuals, we used the strategy described in Zaitlen et al. (4) to estimate pedigree-based heritability (h^2) and SNP-based heritability (h^2_{SNP}) (i.e. variance explained by all SNPs in unrelated individuals) simultaneously in a model (see http://gcta.freeforums.net/thread/241/gcta-greml-analysis-family-data for details about the GCTA commands used). Panel (a): distribution of the estimates of pedigree-based heritability for the 91 traits. Panel (b): distribution of the estimates of SNP-based heritability. Panel (c): plot of the estimate of pedigree-based heritability against that of SNP-based heritability.

a) QQ plots


rs1801133 (MTHFR)
rs71673426 (RAP1A)
rs78720557 (NEK7)
rs78561501 (EGLN1)
rs116611511 (EPAS1)
rs2584462 (ADH7)
rs4498258 (FGF10)
rs9275281 (HLA-DQB1)
rs139129572 (HCAR2)
Observed -log10(P)
Expected -log10(P)

b) Associations at $p < 0.05$

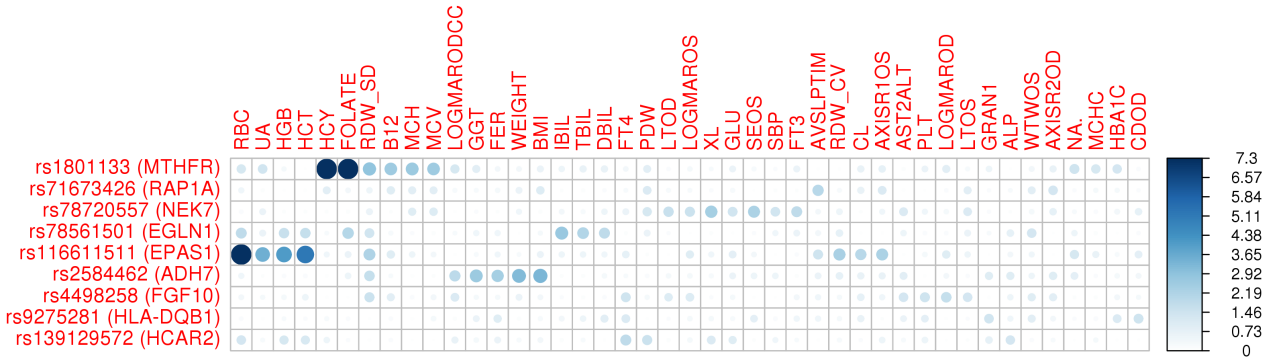


Figure S10 Associations between the 9 differentiated loci and 91 quantitative traits in Tibetans. Panel (a): QQ-plot of associations between each of the 9 loci and the 91 traits. The horizontal line in blue represents p-value at 0.05. The results seem to suggest that there is inflation in test-statistics at the *MTHFR*, *EPAS1* and *ADH7* loci. Panel (b): shown are –log10(p-values) for the traits that are associated with at least one of the 9 loci at $p < 0.05$. For better presentation, p-values are truncated at 5e-8. A list of full names of the traits can be found in Table S5.

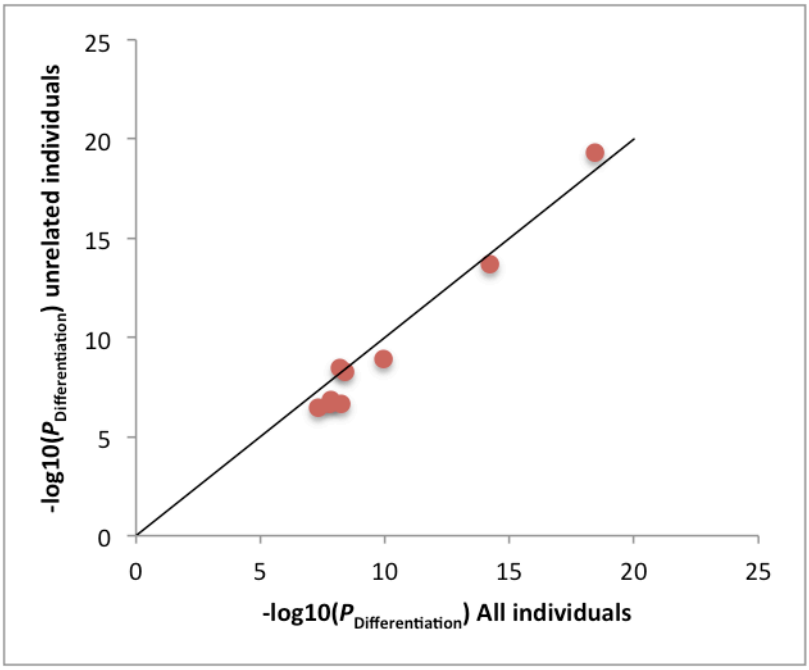


Figure S11 MLMA-LOCO analysis using unrelated individuals (n = 8,538) vs. that using all the individuals (n = 10,295). The set of unrelated individuals were selected at a genetic relatedness (estiamted from all SNPs on HapMap3 with MAF > 0.01) threshold of 0.1 using GCTA-GRM (3). The diagnal line has intercept 0 and slope 1.

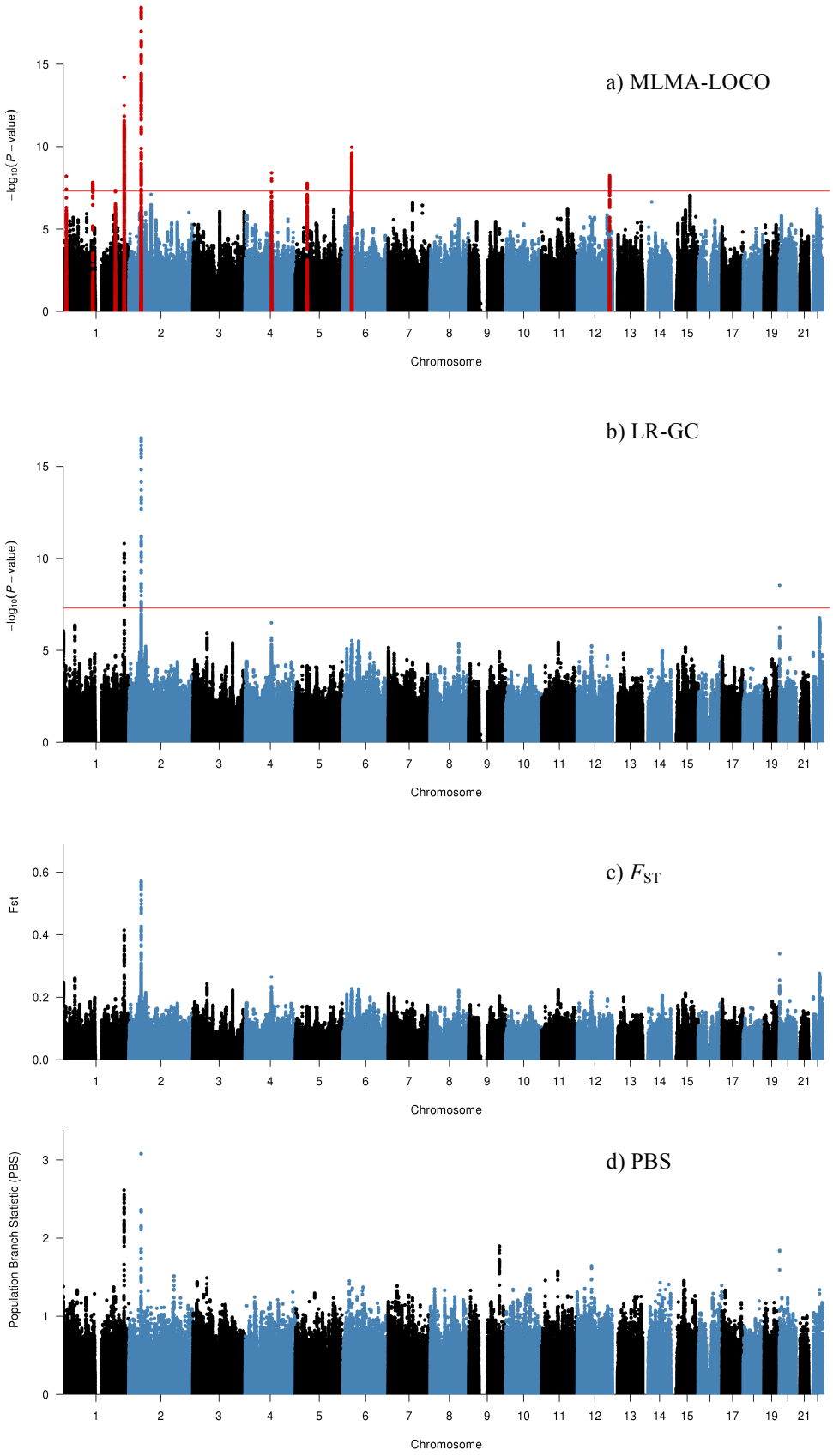


Figure S12 Comprision of the methods for detecting genetic signals of natural selection. In the analysis of linear regression followed by genomic control (LR-GC), the test-statistics from linear regression were divided by the mean test-statistic of all SNPs. F_{ST} values were computed based on Weir and Cockerham's method (6) implemented in GCTA (3). The PBS analysis (7) was performed

using the European subjects from GERA as the reference for PBS score calculation. Note that the F_{ST} and PBS methods are not statistical-testing based approaches and therefore do not control for genome-wide type-I error rate. It seems that there are three SNPs (on chromsomes 9, 20 and 22) that show strong signals in the linear regression and/or PBS analyses but do not reach genome-wide signficance level in the MLMA-LOCO analysis (see the table below). This is because these SNPs are located in regions with strong locus-specific population stratifiation (**Figure S13**).

SNP	CHR	BP	P
rs75237004	9	116,724,010	1.00E-05
rs3746804	20	744,415	1.00E-05
rs5995881	22	41,015,883	1.00E-03

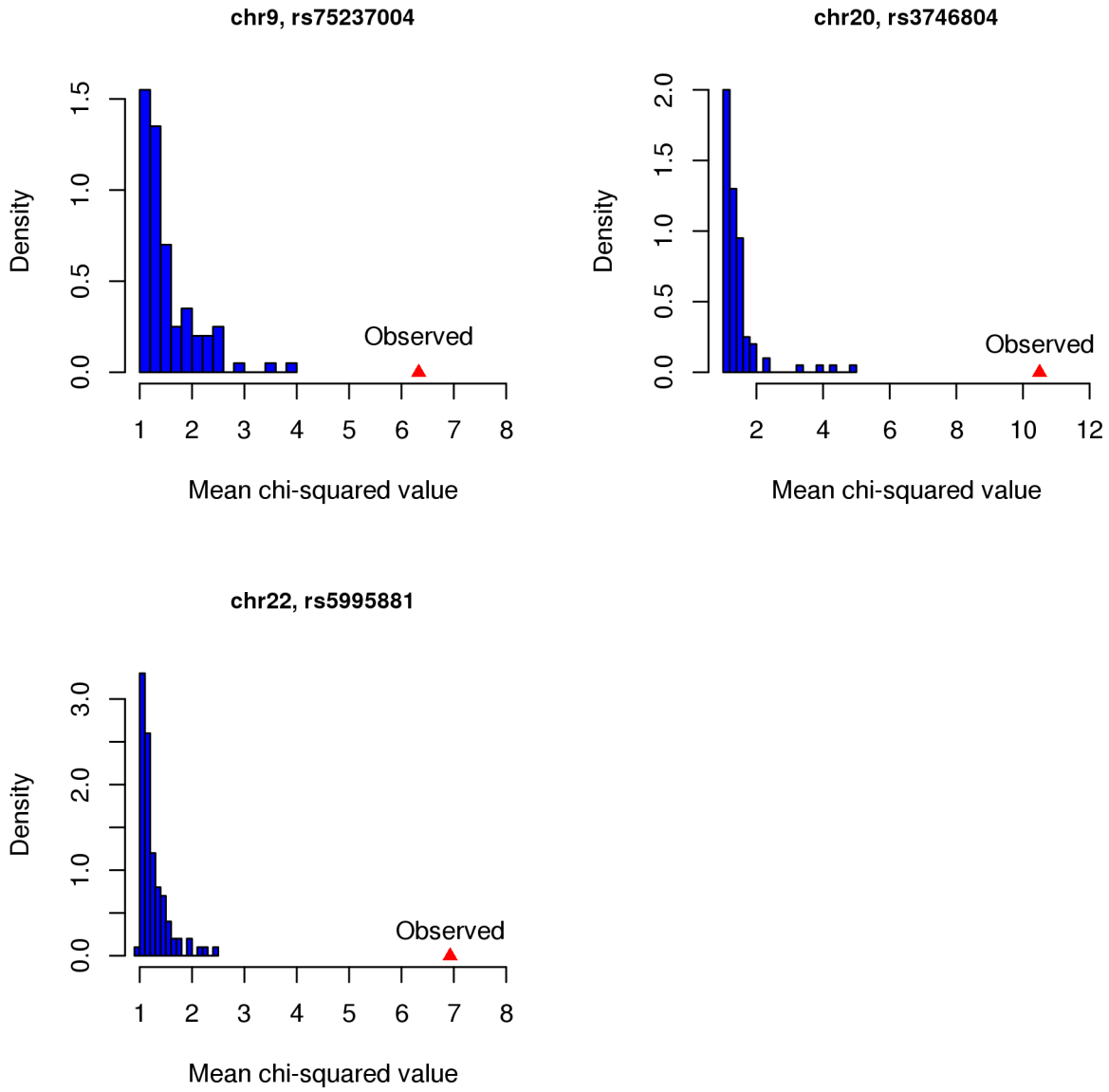


Figure S13 Locus-specific population differentiation. If there is population structure in the sample, SNPs on different chromosomes will be more correlated than expected by chance. Such inter-chromosome correlation is not evenly distributed across the genome but locus-specific, i.e. SNPs in specific genomic regions will tend to be more correlated with SNPs on different chromosomes than average, for example, due to the introgression of DNA from other populations. This is the reason why the MLMA-LOCO analysis, which fits all the SNPs simultaneously as random effects in the model, has the advantage of correcting for locus-specific population structure. In the analysis above, there were three SNPs (rs75237004 on chr9, rs3746804 on chr20, and rs5995881 on chr22) that showed strong signals in the linear regression and/or PBS analyses but were not genome-wide significant in the MLMA-LOCO analysis (**Figure S12**). We hypothesized that this is because these SNPs are located in regions with larger locus-specific population stratification than what we would expect by chance. To test this hypothesis, we performed regression analyses of each of these three

SNPs on all SNPs on the other chromosomes in a set of unrelated individuals of the combined sample (n = 8,538, **Figure S11**). For each of the three SNPs, we randomly sampled 100 MAF-matched SNP (MAF difference < 0.01). For each of the randomly sampled SNPs, we repeated the regression analyses above and calcuated a mean chi-squared value. Shown on each panel of the figure is the distribution of the mean chi-squared values of the 100 randomly sampled SNPs with the observed value labeled by a red triangle on the x-axis. The results are consisent with our hypothesis that these three SNPs are located in regions with strong locus-specific population stratification.

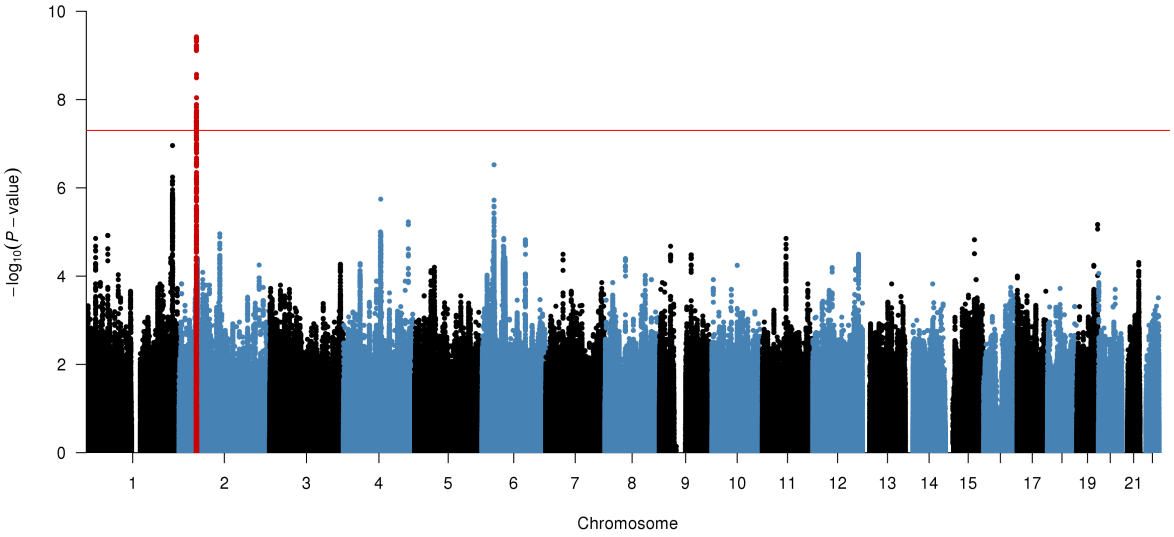


Figure S14 MLMA-LOCO analysis of 150 Tibetan vs. 150 Han subjects. These subjects were randomly sampled from a full set of 3,008 Tibetan and 373 Han subjects collected from Seda and Litang of the Tibetan Plateau. This figure demonstrates why the *EPAS1* locus can be detected in previous studies of small sample size. Note that the increase of power with the increase of sample size in the analysis to detect signal of natural selection is much slower than that in SNP-trait association analysis. This is because the selection analysis is contrasting an alternative model where there is a difference in allele frequency between populations and the differentiation is due to selection, against a null model where there is also a difference in allele frequency and the differentiation is due to genetic drift (rather than no differentiation). Take the F_{ST} analysis for an example. Unlike SNP-trait association analysis where the chi-squared test-statistic is expected to be 1 under the null hypothesis irrespective of sample size, the F_{ST}-based selection analysis has an expected chi-squared test-statistic of $1 + nF_{ST}$ under the null. This explains why we need to adjust the test-statistics by the Genomic Control approach in linear regression analysis (**Figure S12b**).

Text S1 Estimation of divergence time between Han and Chinese

Methods for estimating divergence time between populations utilize the relationship between F_{ST} and T, where F_{ST} is Wright's fixation index and T is the divergence time in generations (8). Under a drift model, $F_{ST} \approx T / 2N_e$ with N_e being the effective population size. We therefore can estimate that T from F_{ST} as $T \approx 2N_eF_{ST}$, where F_{ST} is the mean F_{ST} value calculated from all the genotyped SNPs and N_e can be estimated from the LD between SNPs using the method described in McEvoy et al. (9) (see below). We performed the analysis using 3,008 Tibetan and 373 Han subjects from the TP and an additional set of 1,726 Han subjects from the Eye Hospital of Wenzhou Medical University (WZ). To avoid any potential bias introduced by imputation, we only used genotyped SNPs in common between the two data sets. To remove SNPs with allele frequency difference between cohorts due to batch effects (TP and WZ samples were processed in the same lab but on different batches), we performed a "control-control" analysis to test for the difference in allele frequency between TP-Han and WZ-Han, and removed SNPs with p < 1e-6. We further removed SNPs with MAF < 0.01 and SNPs that were not on the genetic map estimated from 1000G-Han. There were 225,879 SNPs retained for analysis. We estimated the genetic relatedness between subjects using GCTA-GRM and remove one of each pair of subject with estimated relatedness < 0.05 in Tibetans and Han, respectively. There were 1,998 unrelated Tibetans and 2,059 unrelated Han remaining for analysis. We calculated the LD correlations (referring to the same alleles in Tibetans and Han) between all possible pairwise SNPs in 0.0005M to 0.001M distance in each population with the genetic distance data from 1000G-Han (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130507_omni_recombination_rates/) . We then estimated N_e based on the equation $N_e = [1 / r^2_{mean} - 2] / 4c_{mean}$ where r^2_{mean} is the mean LD r^2 between all pairwise SNPs within the regions defined above and c_{mean} is the mean genetic distance between these pairwise SNPs (9).

Text S2 LD score regression analysis

In GWAS especially for those of large sample size, we often observe inflation in test-statistics, as indicated by the excess of $\lambda = \frac{Observed\ median(\chi^2)}{0.455}$ over 1.0 where 0.455 is the expected median(χ^2) under the null hypothesis and λ is defined as the genomic inflation factor. This was previously interpreted as an indication of population structure in the sample not properly accounted for in the association analysis(10). It has been argued that the inflation could be explained by polygenic inheritance, i.e. a large number of genetic variants of small effect influencing the phenotype, and the "Genomic Control" (GC) approach (dividing all the test-statistics by the genomic inflation factor) loses power(11). There is a recent method that is able to distinguish population structure from polygenic inheritance(12). This method is implemented in the LDSC software tool (https://github.com/bulik/ldsc). We performed the LD score regression analysis using the LD scores from 1000G-EAS (provided in the LDSC website) and the summary statistics from the MLMA-LOCO analysis of the combined data set. If the genomic inflation is due to polygenic signals, the regression intercept is expected to be 1 and the regression slope is a function of trait heritability (precisely the proportion of heritability captured by all the SNPs).

Acknowledgements

GERA data: Data came from a grant, the Resource for Genetic Epidemiology Research in Adult Health and Aging (RC2 AG033067; Schaefer and Risch, PIs) awarded to the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics. The RPGEH was supported by grants from the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, the Ellison Medical Foundation, Kaiser Permanente Northern California, and the Kaiser Permanente National and Northern California Community Benefit Programs.

References

1. 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(2):100-106.
2. Yang J*, et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42(7):565-569.
3. Yang J, Lee SH, Goddard ME, & Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88(1):76-82.
4. Zaitlen N*, et al.* (2013) Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet* 9(5):e1003520.
5. Pruim RJ*, et al.* (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26(18):2336-2337.
6. Weir BS & Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38(6):1358-1370.
7. Yi X*, et al.* (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329(5987):75-78.
8. Nei M (1987) *Molecular evolutionary genetics* (Columbia University Press, New York).
9. 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(6):821-829.
10. Devlin B & Roeder K (1999) Genomic control for association studies. *Biometrics* 55(4):997-1004.
11. Yang J*, et al.* (2011) Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* 19(7):807-812.

12. Bulik-Sullivan BK, *et al.* (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47(3):291-295.