LETTER

doi:10.1038/nature13408

Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA

Emilia Huerta-Sánchez[1,2,3]*, Xin Jin[1,4]*, Asan[1,5,6]*, Zhuoma Bianba[7]*, Benjamin M. Peter[2], Nicolas Vinckenbosch[2], Yu Liang[1,5,6], Xin Yi[1,5,6], Mingze He[1,8], Mehmet Somel[9], Peixiang Ni[1], Bo Wang[1], Xiaohua Ou[1], Huasang[1], Jiangbai Luosang[1], Zha Xi Ping Cuo[10], Kui Li[11], Guoyi Gao[12], Ye Yin[1], Wei Wang[1], Xiuqing Zhang[1,13,14], Xun Xu[1], Huanming Yang[1,15,16], Yingrui Li[1], Jian Wang[1,16], Jun Wang[1,15,17,18,19] & Rasmus Nielsen[1,2,20,21]

As modern humans migrated out of Africa, they encountered many new environmental conditions, including greater temperature extremes, different pathogens and higher altitudes. These diverse environments are likely to have acted as agents of natural selection and to have led to local adaptations. One of the most celebrated examples in humans is the adaptation of Tibetans to the hypoxic environment of the high-altitude Tibetan plateau[1–3]. A hypoxia pathway gene, *EPAS1*, was previously identified as having the most extreme signature of positive selection in Tibetans[4–10], and was shown to be associated with differences in haemoglobin concentration at high altitude. Re-sequencing the region around *EPAS1* in 40 Tibetan and 40 Han individuals, we find that this gene has a highly unusual haplotype structure that can only be convincingly explained by introgression of DNA from Denisovan or Denisovan-related individuals into humans. Scanning a larger set of worldwide populations, we find that the selected haplotype is only found in Denisovans and in Tibetans, and at very low frequency among Han Chinese. Furthermore, the length of the haplotype, and the fact that it is not found in any other populations, makes it unlikely that the haplotype sharing between Tibetans and Denisovans was caused by incomplete ancestral lineage sorting rather than introgression. Our findings illustrate that admixture with other hominin species has provided genetic variation that helped humans to adapt to new environments.

The Tibetan plateau (at greater than 4,000 m) is inhospitable to human settlement because of low atmospheric oxygen pressure (~40% lower than at sea level), cold climate and limited resources (for example, sparse vegetation). Despite these extreme conditions, Tibetans have successfully settled in the plateau, in part due to adaptations that confer lower infant mortality and higher fertility than acclimated women of low-altitude origin. The latter tend to have difficulty bearing children at high altitude, and their offspring typically have low birth weights compared to offspring from women of high-altitude ancestry[1,2]. One well-documented pregnancy-related complication due to high altitude is the higher incidence of preeclampsia[2,11] (hypertension during pregnancy). In addition, the physiological response to low oxygen differs between Tibetans and individuals of low-altitude origin. For most individuals, acclimatization to low oxygen involves an increase in blood haemoglobin levels. However, in Tibetans, the increase in haemoglobin levels is limited[3], presumably because high haemoglobin concentrations are associated with increased blood viscosity and increased risk of cardiac events, thus resulting in a net reduction in fitness[12,13].

Recently, the genetic basis underlying adaptation to high altitude in Tibetans was elucidated[4–10] using exome and single nucleotide polymorphism (SNP) array data. Several genes seem to be involved in the response but most studies identified *EPAS1*, a transcription factor induced under hypoxic conditions, as the gene with the strongest signal of Tibetan specific selection[4–10]. Furthermore, SNP variants in *EPAS1* showed significant associations with haemoglobin levels in the expected direction in several of these studies; individuals carrying the derived allele have lower haemoglobin levels than individuals homozygous for the ancestral allele. Here, we re-sequence the complete *EPAS1* gene in 40 Tibetan and 40 Han individuals at more than 200× coverage to further characterize this impressive example of human adaptation. Remarkably, we find the source of adaptation was likely to be due to the introduction of genetic variants from archaic Denisovan-like individuals (individuals closely related to the Denisovan individual from the Altai Mountains[14]) into the ancestral Tibetan gene pool.

After applying standard next-generation sequencing filters (see Methods), we call a total of 477 SNPs in a region of approximately 129 kilobases (kb) in the combined Han and Tibetan samples (Supplementary Tables 1 and 2). We compute the fixation index (F_{ST}; see Methods) between Han and Tibetans, and confirm that it is highly elevated in the *EPAS1* region as expected under strong local selection (Extended Data Fig. 1). Indeed, by comparison to 26 populations from the Human Genome Diversity Panel[15,16] (Fig. 1) it is clear that the variants in this region are far more differentiated than one would expect given the average genome-wide differentiation between Han and Tibetans (F_{ST} ~0.02, ref. 4). The only other genes with comparably large frequency differences between any closely related populations are the previously identified loci associated with lighter skin pigmentation in Europeans, *SLCA45A2* and *HERC2* (refs 17–20), although in these examples the populations compared (for example, Hazara and French, Brahui and Russians) are more genetically differentiated than Han and Tibetans. In populations as closely related as Han and Tibetans, we find no examples of SNPs with as much differentiation as seen in *EPAS1*, illustrating the uniqueness of its selection signal.

F_{ST} is particularly elevated in a 32.7-kb region containing the 32 most differentiated SNPs (green box in Extended Data Fig. 1 and Supplementary Table 3), which is the best candidate region for the advantageous mutation(s). We therefore focus the subsequent analyses primarily on this region. Phasing the data (see Methods) to identify Han and Tibetan haplotypes in this region (Fig. 2), we find that Tibetans carry a high-frequency haplotype pattern that is strikingly different from both their

[1]BGI-Shenzhen, Shenzhen 518083, China. [2]Department of Integrative Biology, University of California, Berkeley, California 94720 USA. [3]School of Natural Sciences, University of California, Merced, California 95343 USA. [4]School of Bioscience and Bioengineering, South China University of Technology, Guangzhou 510006, China. [5]Binhai Genomics Institute, BGI-Tianjin, Tianjin 300308, China. [6]Tianjin Translational Genomics Center, BGI-Tianjin, Tianjin 300308, China. [7]The People's Hospital of Lhasa, Lhasa 850000, China. [8]Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa 50011, USA. [9]Department of Biological Sciences, Middle East Technical University, 06800 Ankara, Turkey. [10]The Second People's Hospital of Tibet Autonomous Region, Lhasa 850000, China. [11]The People's Hospital of the Tibet Autonomous Region, Lhasa 850000, China. [12]The hospital of XiShuangBanNa Dai Nationalities, Autonomous Jinghong, 666100 Yunnan, China. [13]The Guangdong Enterprise Key Laboratory of Human Disease Genomics, BGI-Shenzhen, 518083 Shenzhen, China. [14]Shenzhen Key Laboratory of Transomics Biotechnologies, BGI-Shenzhen, 518083 Shenzhen, China. [15]Princess Al Jawhara Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah 21589, Saudi Arabia. [16]James D. Watson Institute of Genome Science, 310008 Hangzhou, China. [17]Department of Biology, University of Copenhagen, Ole MaaløesVej 5, 2200 Copenhagen, Denmark. [18]Macau University of Science and Technology, AvenidaWai long, Taipa, Macau 999078, China. [19]Department of Medicine, University of Hong Kong 999077, Hong Kong. [20]Department of Statistics, University of California, Berkeley, California 94720, USA. [21]Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark.
*These authors contributed equally to this work.

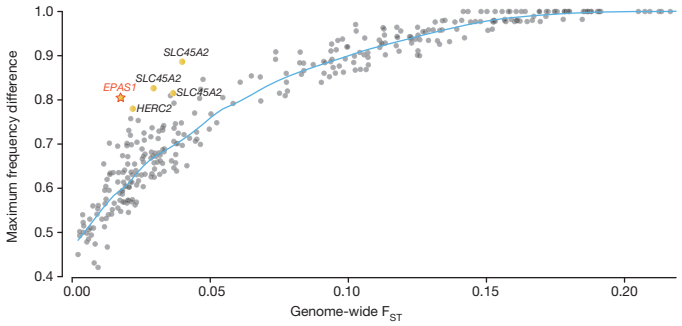


Figure 1 | Genome-wide F_{ST} versus maximal allele frequency difference. The relationship between genome-wide F_{ST} (*x* axis) computed for each pair of the 26 populations and maximal allele frequency difference (*y* axis), first explored in ref. 19. Maximal allele frequency difference is defined as the largest frequency difference observed for any SNP between a population pair. The 26 populations are from the Human Genome Diversity Panel (HGDP). The labels highlight genes that harbour SNPs previously identified as having strong local adaptation. The grey points represent the observed relationship between population differentiation (F_{ST}) and maximal allele frequency difference; the more differentiated populations tend to have mutations with larger frequency differences. The star symbol and the yellow symbols represent outliers; these are populations that are not highly differentiated but where we find some mutations that have higher frequency differences than expected (light blue line).

Figure 2 | Haplotype pattern in a region defined by SNPs that are at high frequency in Tibetans and at low frequency in Han Chinese. Each column is a polymorphic genomic location (95 in total), each row is a phased haplotype (80 Han and 80 Tibetan haplotypes), and the coloured column on the left denotes the population identity of the individuals. Haplotypes of the Denisovan individual are shown in the top two rows (green). The black cells represent the presence of the derived allele and the grey space represents the presence of the ancestral allele (see Methods). The first and last columns correspond to the first and last positions in Supplementary Table 3, respectively. The red and blue arrows indicate the 32 sites in Supplementary Table 3. The blue arrows represent a five-SNP haplotype block defined by the first five SNPs in the 32.7-kb region. Asterisks indicate sites at which Tibetans share a derived allele with the Denisovan individual.

minority haplotypes and the common haplotype observed in Han Chinese For example, the region harbours a highly differentiated 5-SNP haplotype motif (AGGAA) within a 2.5-kb window that is only seen in Tibetan samples and in none of the Han samples (the first five SNPs in Supplementary Table 3, and blue arrows in Fig. 2). The pattern of genetic variation within Tibetans appears even more unusual because none of the variants in the five-SNP motif is present in any of the minority haplotypes of Tibetans. Even when subject to a selective sweep, we would not generally expect a single haplotype to contain so many unique mutations not found on other haplotypes.

We investigate whether a model of selection on either a *de novo* mutation (SDN) or selection on standing variation (SSV) could possibly lead to so many fixed differences between haplotype classes in such a short region within a single population. To do so, we simulate a 32.7-kb region under these models assuming different strengths of selection and conditioning on the current allele frequency in the sample (see Methods). We find that the observed number of fixed differences between the haplotype classes is significantly higher than what is expected by simulations under any of the models explored (Extended Data Fig. 2). Thus the degree of differentiation between haplotypes is significantly larger than expected from mutation, genetic drift and directional selection alone. In other words, it is unlikely ($P < 0.02$ under either a SSV scenario or under a SDN scenario) that the high degree of haplotype differentiation could be caused by a single beneficial mutation landing by chance on a background of rare SNPs, which are then brought to high frequency by selection. The remaining explanations are the presence of strong epistasis between many mutations, or that a divergent population introduced the haplotype into Tibetans by gene flow or through ancestral lineage sorting.

We search for potential donor populations in two different data sets: the 1000 Genomes Project[21] and whole genome data from ref. 14. We originally defined the *EPAS1* 32.7-kb region boundaries by the level of observed differentiation between the Tibetans and Han only (Supplementary Table 3, Extended Data Fig. 1 and Fig. 2) as described in the previous section. In that region, the most common haplotype in Tibetans is tagged by the distinctive five-SNP motif (AGGAA; the first five SNPs in Fig. 2), not found in any of our 40 Han samples. We first focus on this five-SNP motif and determine whether it is unique to Tibetans or if it is found in other populations.

Intriguingly, when we examine the 1000 Genomes Project data set, we discover that the Tibetan five-SNP motif (AGGAA) is not present in any of these populations, except for a single CHS (Southern Han Chinese) and a single CHB (Beijing Han Chinese) individual. Extended Data Fig. 3 contains the frequencies of all the haplotypes present in the fourteen 1000 Genomes populations[21] at these five SNP positions. Furthermore, when we examine the data set from ref. 14 containing both modern (Papuan, San, Yoruba, Mandeka, Mbuti, French, Sardinian, Han Dai, Dinka, Karitiana, and Utah residents of northern and western European ancestry (CEU)) and archaic (high-coverage Denisovan and low-coverage Croatian Neanderthal) human genomes[14], we discover that the five-SNP motif is completely absent in all of their modern human population samples (Supplementary Table 4). Therefore, apart from one CHS and one CHB individual, none of the other extant human populations sampled to date carry this five-SNP haplotype. Notably, the Denisovan haplotype at these five sites (AGGAA) exactly matches the five-SNP Tibetan motif (Supplementary Table 4 and Extended Data Fig. 3).

We observe the same pattern when focusing on the entire 32.7-kb region and not just the five-SNP motif. Twenty SNPs in this region have unusually high frequency differences of at least 0.65 between Tibetans and all the other populations from the 1000 Genomes Project (Extended Data Fig. 4). However, in Tibetans, 15 out of these 20 SNPs are identical to the Denisovan haplotype generating an overall pattern of high haplotype similarity between the selected Tibetan haplotype and the Denisovan haplotype (Supplementary Tables 5–7). Interestingly, five of these SNPs in the region are private SNPs shared between Tibetans and the Denisovan, but not shared with any other population worldwide, except

for two SNPs at low frequency in Han Chinese (Extended Data Fig. 4 and Supplementary Table 7).

If we consider all SNPs (not just the most differentiated) in the 32.7-kb region annotated in humans, to build a haplotype network[22] using the 40 most common haplotypes, we observe a clear pattern in which the Tibetan haplotype is much closer to the Denisovan haplotype than any modern human haplotype (Fig. 3 and Extended Fig. 5a; see Extended Data Fig. 6a, b for haplotype networks constructed using other criteria). Furthermore, we find that the Tibetan haplotype is slightly more divergent from other modern human populations than the Denisovan haplotype is, a pattern expected under introgression (see Methods and Extended Data Fig. 5b). Raw sequence divergence for all sites and all haplotypes shows a similar pattern (Extended Data Fig. 7). Moreover, the divergence between the common Tibetan haplotype and Han haplotypes is larger than expected for comparisons among modern humans, but well within the distribution expected from human–Denisovan comparisons (Extended Data Fig. 8). Notably, sequence divergence between the Tibetans' most common haplotype and Denisovan is significantly lower ($P = 0.0028$) than expected from human–Denisovan comparisons (Extended Data Fig. 8). We also find that the number of pairwise differences between the common Tibetan haplotype and the Denisovan haplotype ($n = 12$) is compatible with the levels one would expect from mutation accumulation since the introgression event (see Methods for Extended Data Fig. 8). Finally, if we compute D (ref. 14) and S^* (refs 23, 24), two statistics that have been designed to detect archaic introgression into modern humans, we obtain significant values (D-statistic $P < 0.001$, and S^* $P \leq 0.035$) for the 32.7-kb region using multiple null models of no gene flow (see Methods, Supplementary Tables 8–10, and Extended Data Figs 9 and 10a).

Thus, we conclude that the haplotype associated with altitude adaptation in Tibetans is likely to be a product of introgression from Denisovan or Denisovan-related populations. The only other possible explanation is ancestral lineage sorting. However, this explanation is very unlikely as it cannot explain the significant D and S^* values and because it would require a long haplotype to be maintained without recombination since the time of divergence between Denisovans and humans (estimated to be at least 200,000 years (ref. 14)). The chance of maintaining a 32.7-kb fragment in both lineages throughout 200,000 years is conservatively estimated at $P = 0.0012$ assuming a constant recombination of 2.3×10^{-8} per base pair (bp) per generation (see Methods). Furthermore, the haplotype would have to have been independently lost in all African and non-African populations, except for Tibetans and Han Chinese.

We have re-sequenced the *EPAS1* region and found that Tibetans harbour a highly differentiated haplotype that is only found at very low frequency in the Han population among all the 1000 Genomes populations, and is otherwise only observed in a previously sequenced Denisovan individual[14]. As the haplotype is observed in a single individual in both CHS and CHB samples, it suggests that it was introduced into humans before the separation of Han and Tibetan populations, but subject to selection in Tibetans after the Tibetan plateau was colonized. Alternatively, recent admixture from Tibetans to Hans may have introduced the haplotype to nearby Han populations outside Tibet. The CHS and CHB individuals carrying the five-SNP Tibetan–Denisovan haplotype (Extended Data Fig. 3) show no evidence of being recent migrants from Tibet (see Methods and Extended Data Fig. 10b), suggesting that if the haplotype was carried from Tibet to China by migrants, this migration did not occur within the last few generations.

Previous studies examining the genetic contributions of Denisovans to modern humans[14,25] suggest that Melanesians have a much larger Denisovan component than either Han or Mongolians, even though the latter populations are geographically much closer to the Altai mountains[14,25]. Interestingly, the putatively beneficial Denisovan *EPAS1* haplotype is not observed in modern-day Melanesians or in the high-coverage Altai Neanderthal[26] (Supplementary Table 4). Evidence has been found for

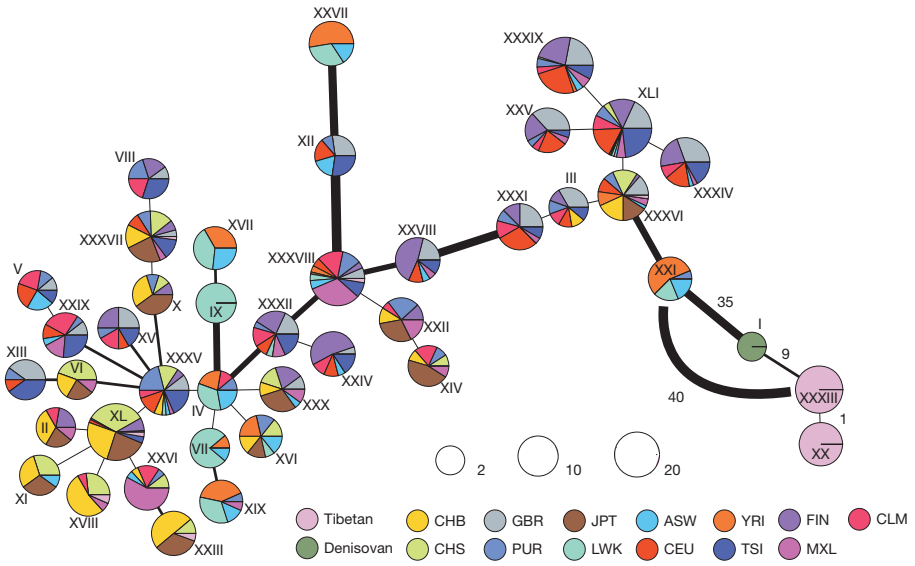


Figure 3 | A haplotype network based on the number of pairwise differences between the 40 most common haplotypes. The haplotypes were defined from all the SNPs present in the combined 1000 Genomes and Tibetan samples: 515 SNPs in total within the 32.7-kb *EPAS1* region. The Denisovan haplotypes were added to the set of the common haplotypes. The R software package pegas[23] was used to generate the figure, using pairwise differences as distances. Each pie chart represents one unique haplotype, labelled with Roman numerals, and the radius of the pie chart is proportional to the $\log_2$(number of chromosomes with that haplotype) plus a minimum size so that it is easier to see the Denisovan haplotype. The sections in the pie provide the breakdown of the haplotype representation amongst populations. The width of the edges is proportional to the number of pairwise differences between the joined haplotypes; the thinnest edge represents a difference of one mutation. The legend shows all the possible haplotypes among these populations. The numbers (1, 9, 35 and 40) next to an edge (the line connecting two haplotypes) in the bottom right are the number of pairwise differences between the corresponding haplotypes. We added an edge afterwards between the Tibetan haplotype XXXIII and its closest non-Denisovan haplotype (XXI) to indicate its divergence from the other modern human groups. Extended Data Fig. 5a contains all the pairwise differences between the haplotypes presented in this figure. ASW, African Americans from the south western United States; CEU, Utah residents with northern and western European ancestry; GBR, British; FIN, Finnish; JPT, Japanese; LWK, Luhya; CHS, southern Han Chinese; CHB, Han Chinese from Beijing; MXL, Mexican; PUR, Puerto Rican; CLM, Colombian; TSI, Toscani; YRI, Yoruban. Where there is only one line within a pie chart, this indicates that only one population contains the haplotype.

Denisovan admixture throughout southeast Asia (as well as in Melanesians) based on a global analysis of SNP array data from 1,600 individuals from a diverse set of populations[27], and this finding has been recently confirmed by ref. 26. Therefore, it appears that sufficient archaic admixture into populations near the Tibetan region occurred to explain the presence of this Denisovan haplotype outside Melanesia. Furthermore, the haplotype may have been maintained in some human populations, including Tibetans and their ancestors, through the action of natural selection.

Recently, a few studies have supported the idea of adaptive introgression from archaic humans to modern humans as having a role in the evolution of immunity-related genes (HLA (ref. 28) and *STAT2* (ref. 29)) and in the evolution of skin pigmentation genes (*BNC2* (refs 23, 30)). Our findings imply that one of the most clear-cut examples of human adaptation is likely to be due to a similar mechanism of gene flow from archaic hominins into modern humans. With our increased understanding that human evolution has involved a substantial amount of gene flow from various archaic species, we are now also starting to understand that adaptation to local environments may have been facilitated by gene flow from other hominins that may already have been adapted to those environments.

METHODS SUMMARY

DNA samples included in this work were extracted from peripheral blood of 41 unrelated Tibetan individuals living at more than 4,300 m above sea level within the Himalayan Plateau, with informed consent. Tibetan identity was based on self-reported family ancestry. The individuals were from two villages of Dingri (4,300 m altitude) and Naqu (4,600 m altitude). These individuals are a subset of the 50 individuals exome-sequenced analysed in ref. 4. Samples of 40 Han Chinese (CHB) are from the 1000 Genomes Project. A combined strategy of long-range PCR and next-generation sequencing was used to decipher the whole *EPAS1* gene and its ±30-kb flanking region. We designed 38 pairs of long-range PCR primers to amplify the region in 41 Tibetan and 40 Han individuals. PCR products from all individuals were fragmented and indexed, then sequenced to higher than 260-fold depth for each individual with the Illumina Hiseq2000 sequencer. The reads were aligned to the UCSC human reference genome (hg18) using the SOAPaligner. Genotypes of each individual at every genomic location of the *EPAS1* gene and the flanking region were called by SOAPsnp. To make comparisons with the 40 Han easier, we only used 40 Tibetan samples for this study.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 December 2013; accepted 28 April 2014.

Published online 2 July; corrected online 13 August 2014 (see full-text HTML version for details).

1. Moore, L. G., Young, D., McCullough, R. E., Droma, T. & Zamudio, S. Tibetan protection from intrauterine growth restriction (IUGR) and reproductive loss at high altitude. *Am. J. Hum. Biol.* **13,** 635–644 (2001).
2. Niermeyer, S. *et al.* Child health and living at high altitude. *Arch. Dis. Child.* **94,** 806–811 (2009).
3. Wu, T. *et al.* Hemoglobin levels in Quinghai-Tibet: different effects of gender for Tibetans vs. Han. *J. Appl. Physiol.* **98,** 598–604 (2005).
4. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329,** 75–78 (2010).
5. Bigham, A. *et al.* Identifying signature of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet.* **6,** e1001116 (2010).
6. Simonson, T. S. *et al.* Genetic evidence for high-altitude adaptation in Tibet. *Science* **329,** 72–75 (2010).
7. Beall, C. M. *et al.* Natural selection on EPAS1 (HIF2a) associated with low hemoglobin concentration in Tibetan highlanders. *Proc. Natl Acad. Sci. USA* **107,** 11459–11464 (2010).
8. Peng, Y. *et al.* Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Mol. Biol. Evol.* **28,** 1075–1081 (2011).
9. Xu, S. *et al.* A genome-wide search for signals of high-altitude adaptation in Tibetans. *Mol. Biol. Evol.* **28,** 1003–1011 (2011).
10. Wang, B. *et al.* On the origin of Tibetans and their genetic basis in adapting high-altitude environments. *PLoS ONE* **6,** e17002 (2011).
11. Moore, L. G. *et al.* Maternal adaptation to high-altitude pregnancy: an experiment of nature—a review. *Placenta* **25,** S60–S71 (2004).
12. Vargas, E. & Spielvogel, H. Chronic mountain sickness, optimal hemoglobin, and heart disease. *High Alt. Med. Biol.* **7,** 138–149 (2006).
13. Yip, R. Significance of an abnormally low or high hemoglobin concentration during pregnancy: special consideration of iron nutrition1'2'3. *Am. J. Clin. Nutr.* **72,** 272S–279S (2000).
14. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
15. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100–1104 (2008).
16. Rosenberg, N. A. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann. Hum. Genet.* **70,** 841–847 (2006).
17. Soejima, M. & Koda, Y. Population differences of two coding SNPs. in pigmentation-related genes SLC24A5 and SLC45A2. *Int. J. Legal Med.* **121,** 36–39 (2007).
18. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genet.* **39,** 1443–1452 (2007).
19. Coop, G. *et al.* The role of geography in human adaptation. *PLoS Genet.* **5,** e1000500 (2009).
20. Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19,** 826–837 (2009).
21. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
22. Paradis, E. Pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* **26,** 419–420 (2010).
23. Vernot, B. & Akey, J. Resurrecting Surviving neandertal lineages from modern human genomes. *Science* (2014).
24. Plagnol, V. & Wall, J. D. Possible ancestral structure in human populations. *PLoS Genet.* **2,** e105 (2006).
25. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature* **468,** 1053–1060 (2010).
26. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).
27. Skoglund, P. & Jakobsson, M. Archaic human ancestry in East Asia. *Proc. Natl Acad. Sci. USA* **108,** 18301–18306 (2011).
28. Abi-Rached, L. *et al.* The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* **334,** 89–94 (2011).
29. Mendez, F. L., Watkins, J. C. & Hammer, M. F. A haplotype at STAT2 introgressed from Neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am. J. Hum. Genet.* **91,** 265–274 (2012).
30. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research was funded by the State Key Development Program for Basic Research of China, 973 Program (2011CB809203, 2012CB518201, 2011CB809201, 2011CB809202), China National GeneBank-Shenzhen and Shenzhen Key Laboratory of Transomics Biotechnologies (no. CXB201108250096A). This work was also supported by research grants from the US NIH; R01HG003229 to R.N. and R01HG003229-08S2 to E.H.S. We thank F. Jay, M. Liang and F. Casey for useful discussions.

Author Contributions R.N., Ji.W. and Ju.W. supervised the project. X.J., A., Z.B., Y.L., X.Y., M.H., P.N., B.W., X.O., H., J.L., Z.X.P.C., K.L., G.G., Y.Y., W.W., X.Z., X.X., H.Y., Y.L., Ji.W. and Ju.W. collected and generated the data, and performed the preliminary bioinformatic analyses to call SNPs and indels from the raw data. E.H.-S. and N.V. filtered the data and B.M.P. phased the data. E.H.-S. performed the majority of the population genetic analysis with some contributions from B.M.P. and M.S. E.H.-S. and R.N. wrote the manuscript with critical input from all the authors.

Author Information Sequence data have been deposited in the Sequence Read Archive under accession number SRP041218. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Ju.W. (wangj@genomics.cn) or Ji.W. (wangjian@genomics.cn) or R.N. (rasmus_nielsen@berkeley.edu).

METHODS

DNA samples. DNA samples included in this work were extracted from peripheral blood of 41 unrelated Tibetan individuals living at more than 4,300 m above sea level within the Himalayan Plateau. Tibetan identity was based on self-reported family ancestry. The individuals were from two villages of Dingri (20 samples prefixed DR (Dingri); 4300m altitude) and Naqu (21 samples prefixed NQ (Naqu); 4,600 m altitude). All participants gave a self-report of at least three generations living at the sampling site, and provided informed consent for this study. These individuals are a subset of the 50 individuals exome-sequenced from ref. 4. Samples of Han Chinese (CHB) are from the 1000 Genomes Project[21] (40 samples prefixed NA in the 1000 Genomes Project).

A combined strategy of long-range PCR and next-generation sequencing was used to decipher the whole *EPAS1* gene and its ±30-kb flanking region (147 kb in total). We designed 38 pairs of long-range PCR primers to amplify the region in 41 Tibetan and 40 Han individuals. PCR products from all individuals were fragmented and indexed, then sequenced to higher than 260-fold depth for each individual with the Illumina Hiseq2000 sequencer. The reads were aligned to the UCSC human reference genome (hg18) using the SOAPaligner[31]. Genotypes of each individual at every genomic location of the *EPAS1* gene and the flanking region were called by SOAPsnp[32]. To make comparisons easier with the Han samples, we only used 40 Tibetan samples for this study.

Data filtering. The coverage for each individual is roughly 200×. Genotypes of each individual at every site in the *EPAS1* gene and the flanking region were called by SOAPsnp[32], which resulted in 700 SNPs in the combined Tibetan–Han sample. However, we filtered out some sites post genotype calling by performing standard filters that are applied in the analyses of next generation sequencing data. Specifically, of the 700 SNPs called, we removed SNPs that fell into the following categories: SNPs that were not in Hardy–Weinberg equilibrium; SNPs that were located in hard-to-map regions (the SNP is located at a position with mappability = 0, using the Duke Unique 35 track); SNPs where more than half the heterozygote individuals have a statistically unequal distribution of counts for the two alleles (that is, the counts of the two alleles deviate from the expectation of 50% of counts for each assuming a binomial distribution); SNPs with different quality-score distributions for the two alleles; SNPs that fell on or near a deletion or insertion; and SNPs that were tri-allelic. A total of 477 SNPs in the combined sample remained after applying all the filters. Within Tibetans, 354 sites (out of the 477 sites) were SNPs, and within Han Chinese, 364 sites (out of the 477) were SNPs. After filtering, we used Beagle to phase the Tibetan and Han individuals together[33].

Human Genome Diversity Panel data. We downloaded the HGDP SNP data from the University of Chicago website (http://hgdp.uchicago.edu/data/plink_data/) and followed the filtering criteria in ref. 34. We used the formula of ref. 35 to compute F_{ST} between pairs of populations. We used the intersection of SNPs between the 50 Tibetan exomes from ref. 7 and the HGDP SNPs, resulting in 8,362 SNPs. Note, the number 354 quoted in the previous section refers to Tibetan SNPs from the full re-sequencing of the *EPAS1* gene in this study.

We calculated F_{ST} for each pair of populations and also scored the frequencies of the SNP with the largest frequency difference between pairs of populations. Using the genotypes from the 26 populations we have re-created Fig. 2a of ref. 34 using the SNPs overlapping in two data sets: the 50 Tibetan exomes data set and the HGDP[15,16] data set. The re-created figure (Fig. 1) displays the maximum SNP frequency difference against the mean F_{ST} across all SNPs for each pair of the HGDP populations. We added one data point to this figure consisting of the mean Fst between Tibetans and Han (F_{ST} is approximately 0.018) and the SNP with the largest frequency difference between Han Chinese and Tibetans (approximately 0.8), which is a SNP in the *EPAS1* gene.

Tibetan and Han haplotypes at the 32.7-kb highly differentiated region. The 32.7-kb region was identified as the region of highest genetic differentiation between Tibetans and Han Chinese (green box in Extended Data Fig. 1), providing the strongest candidate region for the location of the selective sweep. To examine the haplotypes in this region, we first filtered out SNPs that occurred with a frequency of ≤5% or ≥95% in both the Tibetan and Han samples; that is, SNPs that were very common or very rare in both populations simultaneously. We computed the number of pairwise differences between every pair of haplotypes. We then ordered the haplotypes based on their number of pairwise distances from the most common haplotype in each population separately. Figure 2 is generated using the heatmap.2 function from the gplots package of the statistical computing platform R (ref. 36), with derived and ancestral alleles coloured black and light grey, respectively. We used the chimpanzee sequence to define the ancestral state. However, the chimpanzee allele was missing at one of the 32 most differentiated sites (see arrows in Fig. 2), so in that case we used the orangutan and rhesus macaque alleles to define the ancestral allele.

Simulations, selection on a *de novo* mutation and on standing variation. We used msms[37] to simulate data for a population of constant size with mutation, recombination and positive selection affecting a single site. We simulated under conditions of a current allele frequency of 69 out of 80 in the population, the observed value in the real *EPAS1* data, so that the beneficial mutation had frequency at the present time. We estimated a Tibetan effective population size of $N = 7{,}000$ (see Supplementary Information; section on estimating the effective population size). In addition, we assumed three different selection parameters ($2Ns = 200, 500, 1{,}000$, where s is the selection coefficient of the beneficial mutation) and a recombination rate per bp per generation of 2.3×10^{-8} (this is the average recombination rate in the *EPAS1* gene using the fine-scale estimates from the map of ref. 38). This recombination estimate is very similar to the estimate from the African American map[39] for the *EPAS1* gene which is 2.4×10^{-8}. We set the mutation rate to 2.0×10^{-8} per bp per generation because this is what we estimated using the patterns of genetic diversity in the *EPAS1* gene in Tibetans under an approximate bayesian computation (ABC) framework (see Supplementary Information for details; section on estimating the mutation rate). This mutation-rate estimate is similar to the phylogenetic estimates reviewed in ref. 40. We note that the human–chimpanzee differences in other intronic regions in the genome of the same size has a mean (417) and median (410) close to that found for the *EPAS1* 32.7-kb region (see Supplementary Information; section on the distribution of human–chimpanzee differences in 32.7-kb regions for details), suggesting that this region does not have an unusual mutation rate. In the simulations, we examined a region of 32.7 kb around the selected site and grouped the haplotypes into two groups: those that carried the beneficial allele and those that did not. If k is the number of chromosomes carrying the beneficial mutation, we counted how many mutations within the 32.7-kb region had frequency bigger or equal to (k/80 – 0.05) in the group that harboured the beneficial allele (that is, fixed or almost fixed in that group), and simultaneously had frequency 0 in the other group.

To simulate data for a sweep from standing variation, we used mbs[41] and the same parameters as in the previous set of simulations, but assuming an initial allele frequency of the advantageous allele of 1% when selection starts. To be able to compare the number of almost fixed sites from these simulations to the observed data, we needed to make a call in the Han–Tibetan data set of what could plausibly be the selected site. The most straightforward choice is the site that has the highest Han–Tibetan differentiation; see the circled SNP in Extended Data Fig. 1 (this site also has the largest frequency difference (~0.85) between Tibetans and any of the 1000 Genomes populations). Tibetan individuals with the derived mutation at this site were defined as carrying the selected haplotype, and the remaining individuals were defined as not carrying the selected haplotype. Then we performed the same counting of 'almost fixed' sites between these two groups as was done for the simulations. The simulated distribution of almost fixed differences and the real data are shown in Extended Data Fig. 2 (histograms of almost fixed differences).

For the SDN model under a selection parameter of $2Ns = 200, 500, 1{,}000$, the P values are 0.004, 0.006 and 0.006, respectively. Under SSV with a selection parameter of $2Ns = 200, 500, 1{,}000$, the P values are 0.002, 0.012 and 0.015, respectively. We note that increasing the initial frequency of the selected allele (to 5%) also leads to a smaller number of fixed differences than what we observe in the real data, thereby making the simulated SSV scenario similarly unlikely (P values are 0.007, 0.01 and 0.006 for $2Ns = 200, 500, 1{,}000$ respectively). We also note that simulating data with a smaller mutation rate will not result in an increase in the number of fixed differences.

Five-SNP motif. We identified the contiguous five-SNP haplotype motif that is most common in the 40 Tibetan samples, but entirely absent in the 40 Han individuals (see the five-SNP haplotype defined by the first five blue arrows in Fig. 2). The five SNPs comprising this motif (positioned at 46421420, 46422184, 46422521, 46423274 and 46423846), span a 2.5-kb region (46423846–46421420, ~2.5 kb) containing no other SNPs (even when including low- and high-frequency SNPs). The genomic positions of this five-SNP motif were then examined in the phased 1000 Genomes[21] data set to compute the frequency of this five-SNP haplotype in the populations sequenced in the 1000 Genomes project (see Extended Data Fig. 3, and below). We will refer here to this five-SNP motif as the 'core' Tibetan haplotype.

Haplotype frequencies at the five-SNP motif in 1000 Genomes data. For all samples or populations in the 1000 genomes project[21], we interrogated the five sites in the 'core' Tibetan haplotype identified in *EPAS1*, and counted the frequencies of each of the unique haplotypes within each population group of the 1000 genomes. The bar-plot in Extended Data Fig. 3 is a summary of these frequencies within each population, coloured by the unique haplotype sequences present.

Haplotype network. We constructed a haplotype network including Tibetans, Denisovans and the 1000 Genomes samples (ASW, African American from south western United States; CLM, Colombian; CEU, Utah Residents with Northern and Western European ancestry; CHB, Han Chinese from Beijing; CHS, Southern Han Chinese; FIN, Finnish; GBR, British; JPT, Japanese; LWK, Luhya; MXL, Mexican; PUR, Puerto Rican; TSI, Toscani; YRI, Yoruban) for the 32.7-kb *EPAS1* region in the combined 1000 Genomes samples. To limit the number of haplotypes to display, we identified the 40 most common haplotypes. There are a total of 515 SNPs in the 32.7-kb *EPAS1* region that pass all quality filters. We used the R (ref. 36) software package

pegas (ref. 22) to build a tree that connects haplotypes based on pairwise differences (Fig. 3). The Denisovan individual is homozygous at all the 515 sites. Note that for clarity (to reduce the number of colours needed for the plot) and because the small sample of Iberians (18) only contain haplotypes observed in other European samples, Fig. 3 does not include the Iberians (IBS). We find that the Denisovan haplotype is closest to the Tibetan haplotypes. Extended Data Fig. 5a contains all the pairwise differences between the 41 (40 from modern humans and 1 from Denisovan) haplotypes in Fig. 3.

The observed haplotype structure is compatible with the introgression hypothesis. As expected under the introgression hypothesis, the Tibetan haplotype is more distant to the non-Tibetan haplotypes than the Denisovan haplotype because, after the admixture event, the introgressed haplotype accumulated extra mutations. In contrast, the Denisovan haplotype, being the product of DNA extracted from an ancient specimen, did not have time to accumulate a similar number of mutations. This effect is illustrated in Extended Data Fig. 5b. For example, the divergence between the introgressed haplotype (that is, the Tibetan haplotype) and the Yoruban haplotype would be larger than between the observed Denisovan haplotype and the Yoruban haplotype (see Extended Data Fig. 5b and Supplementary Information; section on Extended Data Fig. 5b).

Haplotype network. Figure 3 plots the network of the 40 most common haplotypes. Alternatively, we also used the 20 sites such that the frequency difference between Tibetans and each of the 14 populations from the 1000 Genomes project[21] is at least 0.65 (see Extended Data Fig. 4) to construct a haplotype network (Extended Data Fig. 6a). The resulting region spanned by these SNPs is the same 32.7-kb region as identified previously by considering sites that are the most differentiated between Tibetans and Han (Supplementary Table 3). For more details about Extended Data Fig. 6a, b, see Supplementary Information; section on haplotype networks constructed using other criteria).

Denisovan–human number of pairwise differences at the *EPAS1* 32.7-kb region. We computed the number of pairwise differences as described in Supplementary Information (section on the number of pairwise differences). We removed all the Denisovan sites that had a genotype quality smaller than 40 and a mapping quality smaller than 30, using the same thresholds as in the Denisovan paper[14]. This filtering resulted in the removal of 782 sites (out of 32,746). We also removed another 27 sites within the 32.7-kb region that did not pass our quality filters in Tibetans (see the data filtering section). The total number of SNPs in the combined Tibetan, 1000 Genomes and the Denisovan samples is 520. For the 32.7-kb region in *EPAS1*, we computed the number of pairwise differences between the Denisovan haplotypes and each of Tibetan haplotypes (red histograms, Extended Data Fig. 7). We also computed the number of pairwise differences between the Denisovan haplotypes and each of the haplotypes in the 1000 Genomes Project's populations (see the blue histograms in Extended Data Fig. 7). Notice that for this comparison, we compared every site that passes the quality filters even if the site is not polymorphic in modern humans. This is in contrast to Fig. 3 where we only considered the sites that are polymorphic in modern humans. Furthermore, if a site is not polymorphic in our sample, we assumed that all of our samples carry the human reference allele. We plot two histograms in each panel of Extended Data Fig. 7: the distribution of Tibetan–Denisovan comparison (red histogram) and the distribution of pairwise differences between the Denisovan haplotype and each population (blue histogram) from the 1000 Genomes Project (Extended Data Fig. 7).

Denisovan–modern human divergence and modern human–modern human divergence. To compute the genomic distribution of modern human–Denisovan pairwise differences we examined all windows of intronic sequence of size 32.7 kb (using a table from Ensembl with the exon boundaries for all genes) from chromosomes 1 to 6. Within each 32.7-kb region, we removed all the Denisovan sites that had a genotype quality smaller than 40 and a mapping quality smaller than 30. We computed divergence by computing all the pairwise differences between a human haplotype and the Denisovan haplotypes (see Supplementary Information; section on the number of pairwise differences) and dividing by the effective sequence length (that is, all the sites in the 32.7-kb region that passed all the filters; a mapping quality higher or equal to 30 and a genotype quality higher or equal to 40). We only kept the 32.7-kb regions where at least 20,000 sites passed these quality criteria. The modern humans used in these comparisons were the first 80 CEU chromosomes, the first 80 CHS chromosomes and the first 80 CHB chromosomes from the 1000 Genomes data. If a site was not polymorphic in modern humans, we assumed that they carried the reference allele.

We also computed modern human–modern human divergence at the same intronic regions. In this case, we compare modern human populations (CHB versus CHS, CHB versus CEU, CHS versus CEU) by comparing all 80 haplotypes in one group to all 80 haplotypes in the other group for a total of $3 \times 80 \times 80$ comparisons. The distributions of modern human–Denisovan and modern human–modern human pairwise differences are both plotted in Extended Data Fig. 8. We also display the distribution of Tibetan–Han pairwise differences in the 32.7-kb region of the *EPAS1* locus (80 Tibetan and 80 Han for a total of 6,400 comparisons). Finally, we include the pairwise differences between the Denisovan and the Tibetans computed as in Extended Data Fig. 7, standardized by the number of sites that passed all quality filters. This number (12/31,937) leads to a sequence divergence of 0.000375 for the most common Tibetan haplotypes, and this is indeed significantly lower ($P = 0.0028$) than what is expected under the distribution of human–Denisovan divergence (see Extended Data Fig. 8). Supplementary Table 11 contains the details regarding the 12 differences between the Tibetan and the Denisovan haplotypes.

To address further the issue of whether 12 differences are expected between the Denisovans and Tibetans under the introgression hypothesis, we computed the number of mutations theoretically expected for an introgressed region of this size, given published estimates of the age of the sample, and the coalescence time within Denisovans. We assumed that mutations occur as a Poisson process and used the estimates of split times from ref. 26 between the called introgressed Denisovan haplotypes and the Denisovan haplotypes (see page 114 of the Supplementary Information of ref. 26). Using these estimates, the number of expected mutations between the Denisovan haplotype and our introgressed haplotype (the Tibetans' most common haplotype) is simply: $(2 \times \text{tMRCA} - \text{age}) \times \text{L} \times \text{mu} = 11.25$, where *tMRCA* is the time to the most recent common ancestor estimated at 394,000 years (394 kyr), $\text{mu} = 0.5 \times 10^{-9}$ per bp per year, L = 32.7 kb, and age is the age of the Denisovan sample, which we conservatively set to 100,000 years. Clearly, the observed value of 12 mutations is remarkably close to the expected number (11.25). In fact, we would need to observe 17 or more mutations to be able to reject the introgression hypothesis at the 5% significance level. If we use our estimate of the mutation rate in the *EPAS1* gene, $\text{mu} = 1.0 \times 10^{-9}$ per bp per year (2.0×10^{-8} per bp per generation), then the expected number of differences is 22.5. Therefore we conclude the number of differences we observe are compatible with the previous estimates of introgressed Denisovan versus sampled Denisovan sequence divergence.

Probability of 32.7-kb haplotype block from shared ancestral lineage. We calculate the probability of a haplotype, of length at least 32.7 kb, shared by modern Tibetans and the archaic Denisovan due to incomplete ancestral lineage sorting. Let r be the recombination rate per generation per bp. Let t be the length of the human and Denisovan branches since divergence. The expected length of a shared ancestral sequence is $1/(r \times t)$. Let this expected length $= L$. Assuming an exponential distribution of admixture tracts, the probability of seeing a shared fragment of length $\geq m$ is $\exp(-m/L)$. However, conditional on observing the Denisovan nucleotide at position j, the expected length is the sum of two exponential random variables with expected lengths L, therefore it follows a Gamma distribution with shape parameter 2, and rate parameter lambda $= 1/L$. Inserting numbers for human branch length after divergence at a conservative lower estimate of 200 kyr, and the Denisovan branch of 100 kyr (divergence minus the estimated age of the Denisovan sample, which can be as old as 100 kyr (refs 14, 26)), and assuming a generation time of 25 years, we get $L = 1/(2.3 \times 10^{-8} \times (300 \times 10^3/25)) = 3623.18$ bp, and the probability of a length of at least $m = 32{,}700$ bp is $1 - \text{GammaCDF}(32700, \text{shape} = 2, \text{rate} = 1/L) = 0.0012$. GammaCDF is the Gamma distribution function and the numbers within the parenthesis are its arguments. Here the recombination of 2.3×10^{-8} is the average recombination rate in *EPAS1* calculated from the estimates in ref. 14. We should mention, both this divergence estimate for the Denisovan–human split and the age of the Denisovan sample are highly conservative[14,25,26], so the actual probability may be considerably less. Also, the haplotype would have to have been independently lost in all African and non-African populations, except for Tibetans and Han Chinese.

Null distributions of *D* statistics under models of no gene flow. As another approach to assess the probability of an ancestral lineage having given rise to the 32.7-kb haplotype that we observe in Tibetans in the absence of gene flow, we compared *D* statistics between human populations under simulations[42] of several demographic models described in ref. 43. *D* statistics were calculated according to equation (2) in ref. 44. The two modern human populations used in computing *D* statistics are Tibetans and either CHB, CEU or YRI (see Supplementary Information for details; section on *D* statistics under Models of no gene flow). All simulations results result in $P < 0.001$ for all comparisons (see Extended Data Fig. 9, Supplementary Tables 8–10 and Supplementary Information).

Genome-wide value of *D* statistics. *D* statistics have been used to assess genome-wide levels of archaic introgression in previous studies[14,25]. To assess whether Tibetans carry more Denisovan admixture than other populations (CEU or CHB), we used the SNP genotype data from ref. 45 and computed *D* statistics as in ref. 44: *D*(chimpanzee, Denisovan, Tibetan and CHB) and *D*(chimpanzee, Denisovan,Tibetan and CEU). At the genome-wide level, using the *D* statistic, we found no evidence that there is more Denisovan admixture in Tibetans than in the Han ($D = 0.000504688$). We also did not find evidence that there is more Denisovan admixture in Tibetans than in the Europeans ($D = 0.001898642$).

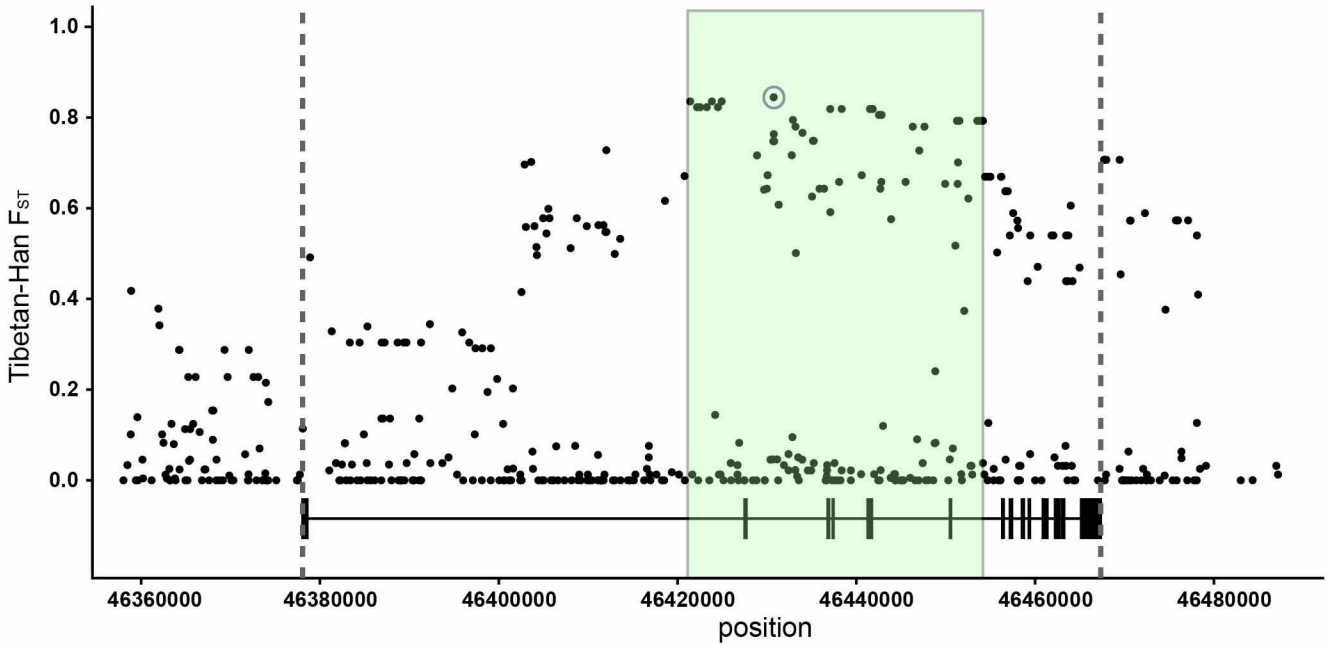
Empirical distributions of *D* statistics for 32.7-kb intronic regions. The *EPAS1* 32.7-kb region was chosen due to its positive selection signal, and not based on a

genome-wide analysis of Denisovan introgression. Therefore, we only performed one test when testing for introgression and did not have to correct *P* values for multiple testing. We do not have Tibetan whole genome sequence data, but as shown in the previous section, genotype array data suggests that the level of Denisovan introgression between Han and Tibetans is similar. Moreover, Tibetans and Han are closely related populations. Therefore, using Han data as a proxy, we can determine whether the observed *D* values at the *EPAS1* region (*D*(TIB, YRI, DEN, chimpanzee) = −0.8818433) is an outlier compared to the distribution of *D* values at other 32.7-kb intronic regions. Using the empirical distribution of *D* values across chromosomes 1 to 22, substituting the 80 Han chromosomes for our 80 Tibetan chromosomes and computing *D*(HAN, YRI, DEN, chimpanzee) for each 32.7-kb intronic region, we obtain a $P < 0.008$. However, as the variance in *D* depends on the number of informative sites, this is probably an overestimate of the true *P* value. In fact, there are no other regions with as many informative sites and as extreme a *D* value as that observed for *EPAS1*. This region is clearly a strong outlier.

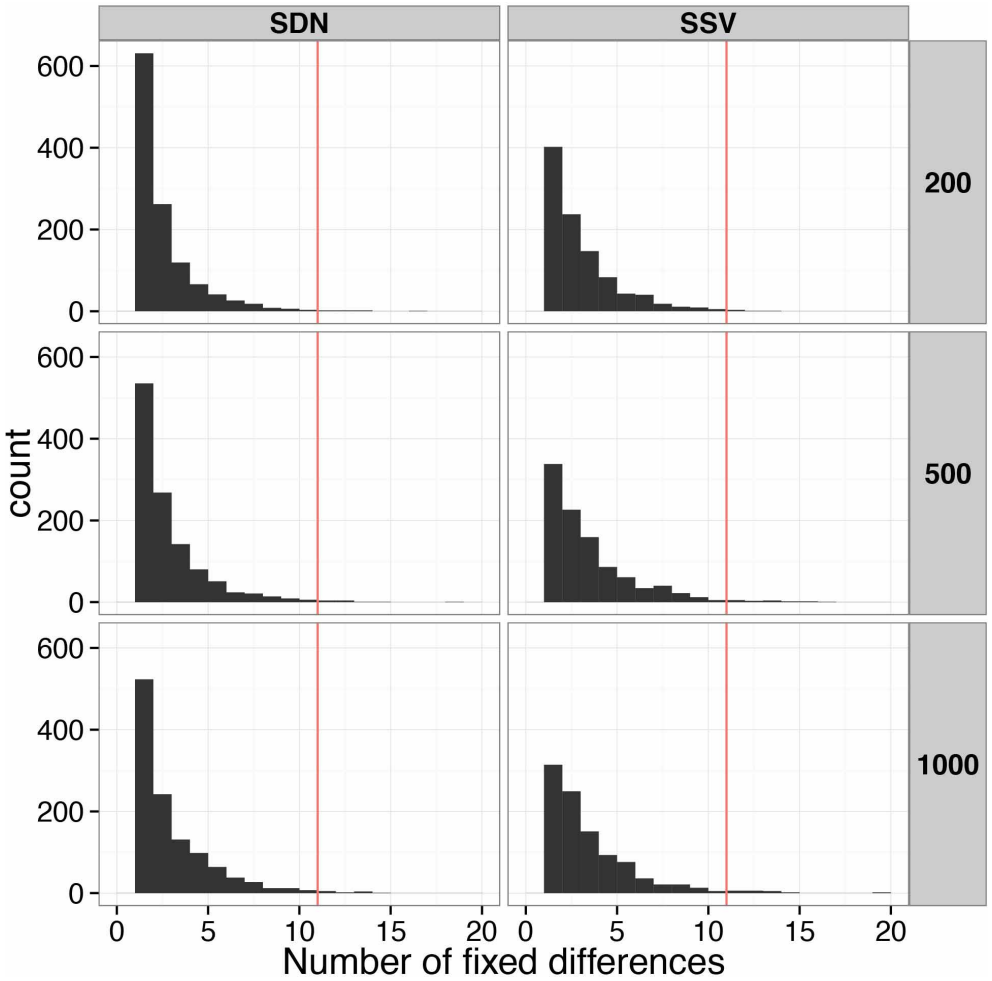
Null distributions of *S statistics under models of no gene flow.** As a final approach for eliminating the hypothesis of ancestral lineage sorting, we follow the methods of ref. 23 to compute *S** (originally derived by ref. 24). *S** was designed to identify regions of archaic introgression. As in the previous section, we used all the four models of ref. 43 that do not include gene flow and simulated data to compute the null distributions of *S**. Distributions are generated from 1000 simulations, and within each simulation we have representation of the 80 Tibetan chromosomes, and 20 Yoruban chromosomes as the outgroup. For each simulated data set we follow ref. 23 and compute *S** on a per-chromosome basis, after sampling at random 20 chromosomes from the Tibetan group and removing SNPs that are observed in the Yoruban chromosomes. The maximum *S** is then recorded. The above process is carried out for 10 random samplings of 20 Tibetan chromosomes and the maximum of the 10 is the final recorded *S**. The exact same procedure is applied to the simulated data and the real data of 80 Tibetan chromosomes. Extended Data Fig. 10a shows that under all four models, *S** is significantly different from the null distribution with all the empirical p-values lying below 0.035. The grey vertical line is the *S** value computed for the real data. The *P* values are 0.035, 0.028, 0.019 and 0.017, respectively, for each model (top to bottom).

Principal component analyses using Chinese and Tibetan samples. As one single CHB individual carries a haplotype that is very similar to the Denisovan haplotype in *EPAS1* (Extended Data Fig. 6a, b), we wanted to assess whether this similarity might be due to recent gene-flow from Tibetans to CHB (Chinese samples were from the 1000 Genomes Project; Tibetan samples were from ref. 45). If that were true, then we would expect to observe similarities at other loci. Therefore we compute the first and second principal components using all of chromosome 2. For simplicity, we only used chromosome 2 because it contains the *EPAS1* gene and has a sufficiently high number of SNPs to carry out the PCA analysis. We do not have genome-wide genotype calls for the 40 Tibetan samples considered in this study. Therefore, as a proxy, we used the Tibetan genotype data from ref. 45 and compared their Tibetan samples to the CHB and CHS individuals from 1000 Genomes. Extended Data Fig. 10b shows that all the CHB and the CHS individuals cluster together and principal component 1 clearly separates Tibetans from CHB and CHS individuals. Furthermore, the CHB individual with the Denisovan *EPAS1* haplotype (Extended Data Figs 6a, b) clearly clusters with other CHB and CHS individuals and do not show any closer genetic affinity with Tibetans. This suggests that the CHB individual with a Denisovan-like haplotype in *EPAS1* is not a descendant of a recent immigrant from Tibet.

31. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24,** 713–714 (2008).
32. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19,** 1124–1132 (2009).
33. Browning, B. L. & Browning, S. R. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* **88,** 173–182 (2011).
34. Coop, G. *et al.* The role of geography in human adaptation. *PLoS Genet.* **5,** e1000500 (2009).
35. Reynolds, J., Weir, B. S. & Cockerham, C. C. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105,** 767–779 (1983).
36. R Development Core Team R: A language and environment for statistical computing http://www.R-project.org/ (R Foundation for Statistical Computing, 2011).
37. Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure, and selection at a single locus. *Bioinformatics* **26,** 2064–2065 (2010).
38. Myers, S. *et al.* A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310,** 321–324 (2005).
39. Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476,** 170–175 (2011).
40. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature Rev. Genet.* **13,** 745–753 (2012).
41. Teshima, K. M. & Innan, H. mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC Bioinformatics* **10,** 166 (2009).
42. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338 (2002).
43. Sankararaman, S. *et al.* The date of interbreeding between Neandertals and modern humans. *PLoS Genet.* **8,** e1002947 (2012).
44. Durand, E. Y. *et al.* Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28,** 2239–2252 (2011).
45. Simonson, T. S. *et al.* Genetic evidence for high-altitude adaptation in Tibet. *Science* **329,** 72–75 (2010).

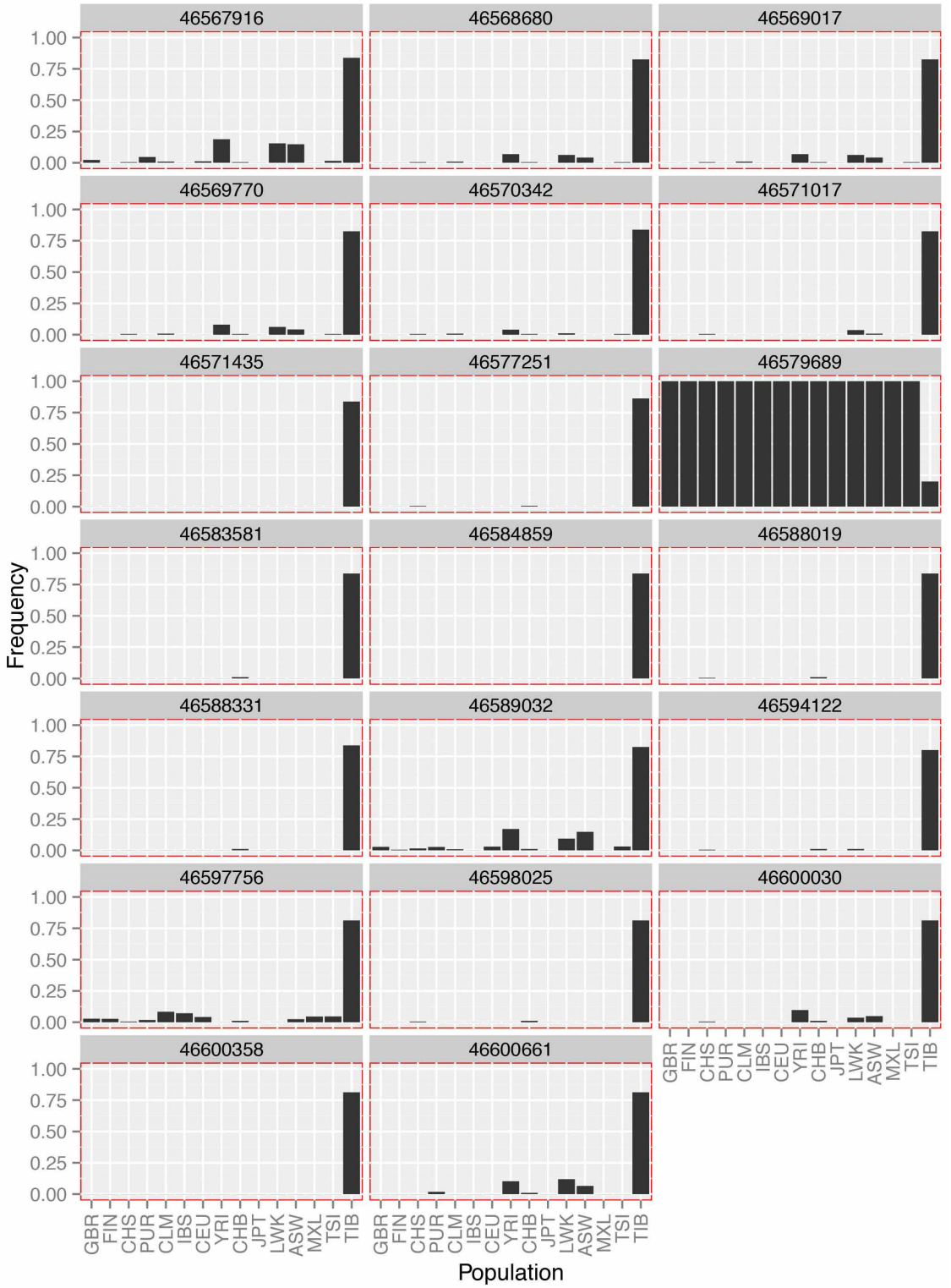


Extended Data Figure 1 | F_{ST} calculated for each SNP between Tibetan and Han populations. Each dot represents the F_{ST} value for each SNP in *EPAS1*. The *x* axis is the physical position in the gene. Positions are based on the hg18 build of the human genome. The green box defines a 32.7-kb region where we observe the largest genetic differentiation between Han Chinese and Tibetans. The first and last positions of this 32.7-kb region correspond to the first and last position of the SNPs listed in Supplementary Table 3. For comparison, in ref. 4 the genome-wide F_{ST} between Han and Tibetans is 0.02. The site with the largest frequency difference (and therefore largest F_{ST}) is circled.

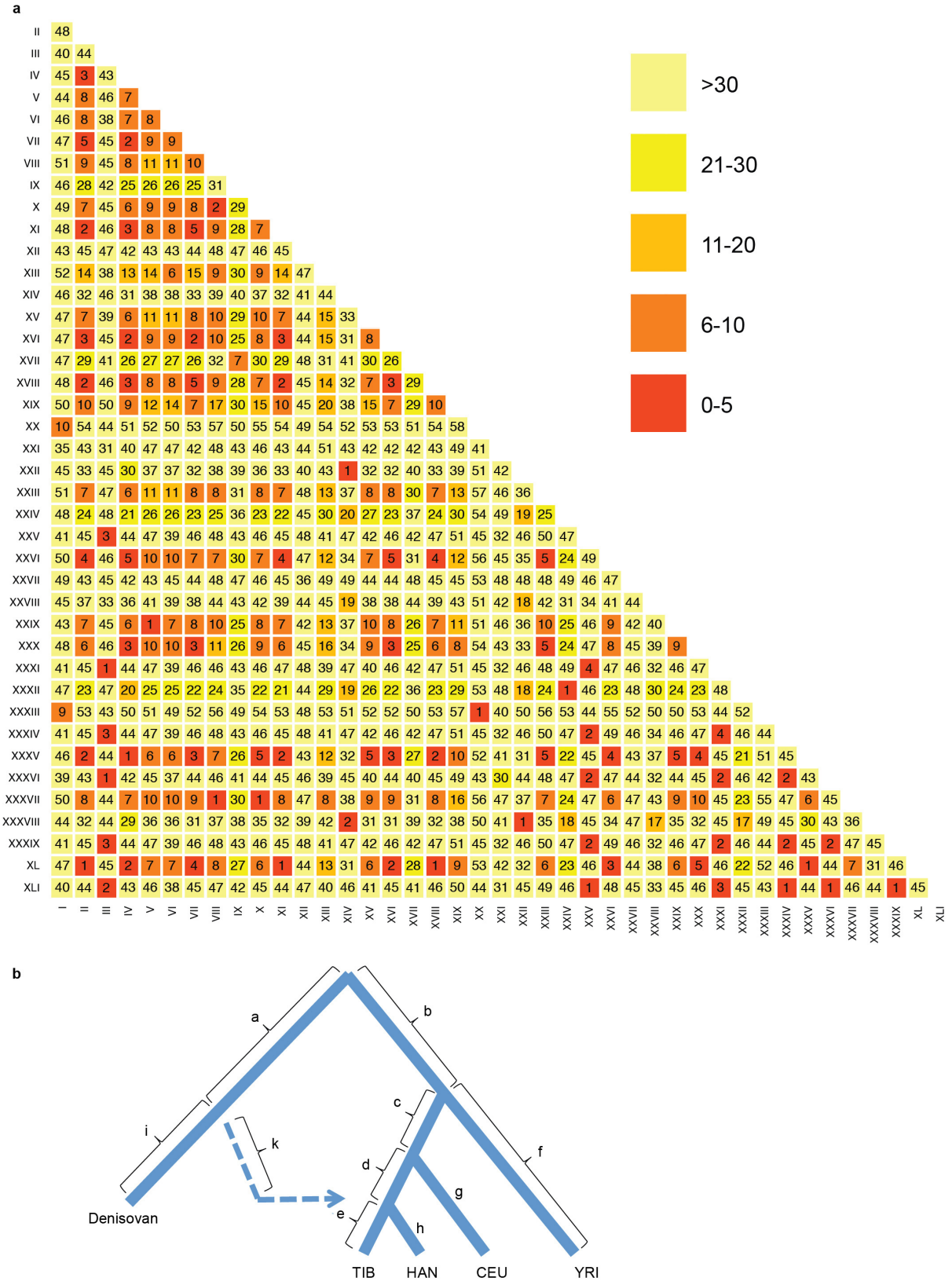


Extended Data Figure 2 | Distribution of fixed differences. The left panel is the distribution of fixed differences between two haplotype groups under a scenario of selection on a *de novo* mutation (see Methods), and the right panel is the distribution under a scenario of selection on standing variation (see Methods) for a region of size ~32.7 kb. The initial frequency of the selected allele in the SSV model is 1%. Each row of panels corresponds to different selection strengths ($2Ns$) from 200 to 1,000. The red lines mark the number of fixed differences observed between the two haplotype classes in the real data for the given window size.

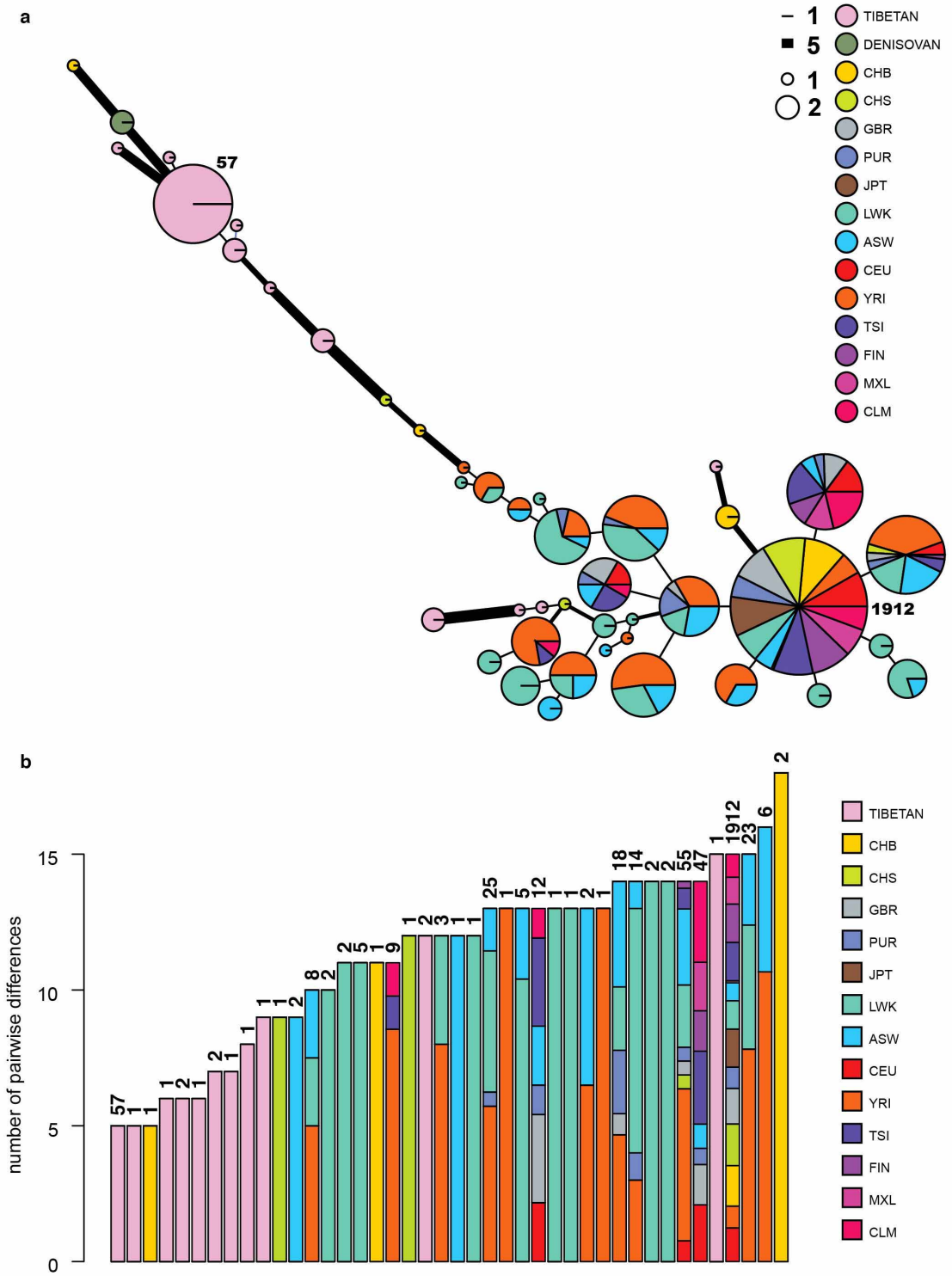
Extended Data Figure 3 | Haplotype frequencies for Tibetans, our Han samples and the populations from the 1000 genomes project for the five-SNP motif in the *EPAS1* region. The *y* axis is the haplotype frequency. The legend shows all the possible haplotypes for the region considered among these populations: ASW, African American from the south western United States; CEU, Utah Residents with Northern and Western European ancestry; CHB, Han Chinese from Beijing; CHS, Southern Han Chinese; CLM, Colombian; FIN, Finnish; GBR, British; HAN, Han Chinese from Beijing; IBS, Iberian; JPT, Japanese; MXL, Mexican; PUR, Puerto Rican; LWK, Luhya; TSI, Toscani; TIB, Tibetan; YRI, Yoruban (see Methods).



Extended Data Figure 4 | Derived allele frequency of the SNPs with the largest frequency difference between Tibetans and the 1000 Genomes Project populations. At these SNPs, the frequency difference between Tibetans and the 1000 Genomes project populations is 0.65 or larger. Positions 46571435, 46579689, 46584859 and 46600358 were not called as SNPs in the 1000 Genomes data, so we assume these positions were fixed for the human reference allele. Note that even though position 46577251, 46588331, 46594122 and 46598025 appear to have a frequency of 0.0 for the populations in the 1000 Genomes data, the derived allele in these SNPs are observed at very low frequency in at least one population (for example, CHB).

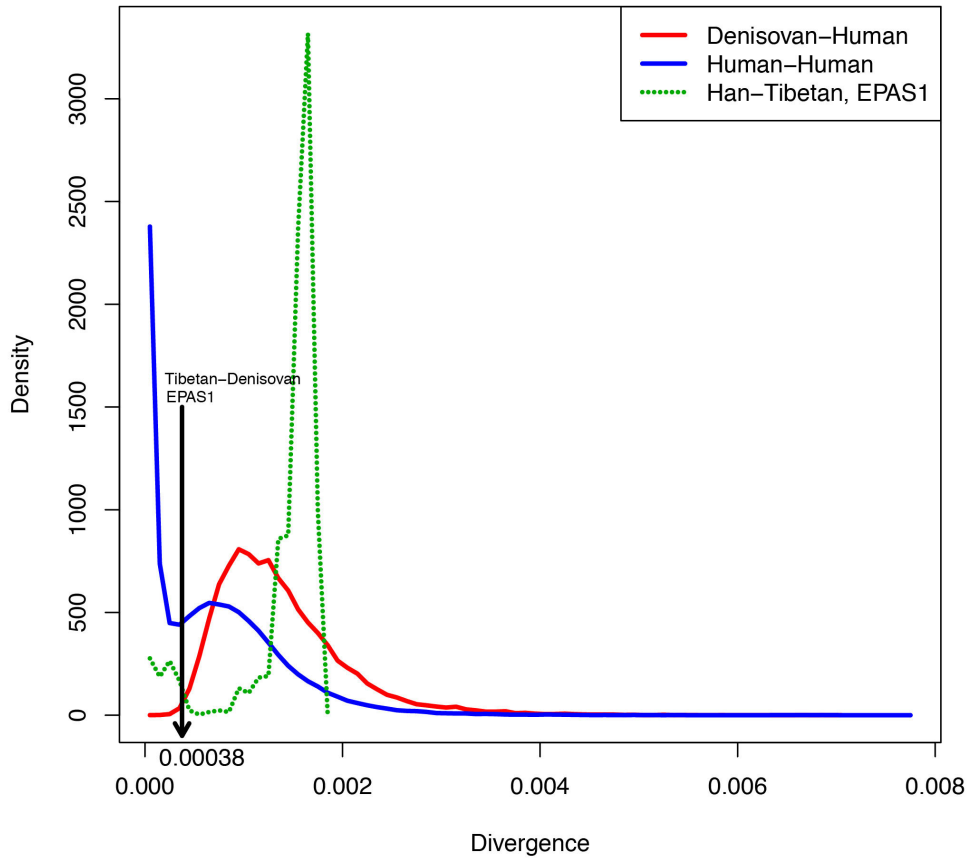


Extended Data Figure 5 | Differences between haplotypes. **a**, The full matrix of pairwise differences between all the unique haplotypes in Fig. 3, for the 40 most common haplotypes identified in the 1000 Genomes and the Tibetan samples in the 32.7-kb region of *EPAS1*. The Denisovan haplotype (of frequency two) was added afterwards for comparison. The unique haplotypes are labelled with Roman numerals (here and in Fig. 3), and the Denisovan haplotype is the first column, haplotype I. Refer to Fig. 3 in the main text and the supplementary material for the representation of populations for each haplotype. **b**, Illustration of the genealogical structure in a model with gene flow from Denisovans to Tibet. Letters a–k are the labels for the branch lengths and are adjacent to their corresponding branches. The divergence between modern human haplotypes and the introgressed haplotype in Tibetans would be larger than the haplotypes in other modern human populations and the Denisovan haplotype (see Methods and Supplementary Information). TIB, CEU and YRI denote Tibetan, European and Yoruban populations. Note that the lengths i and k are unknown as we do not know when these populations went extinct.

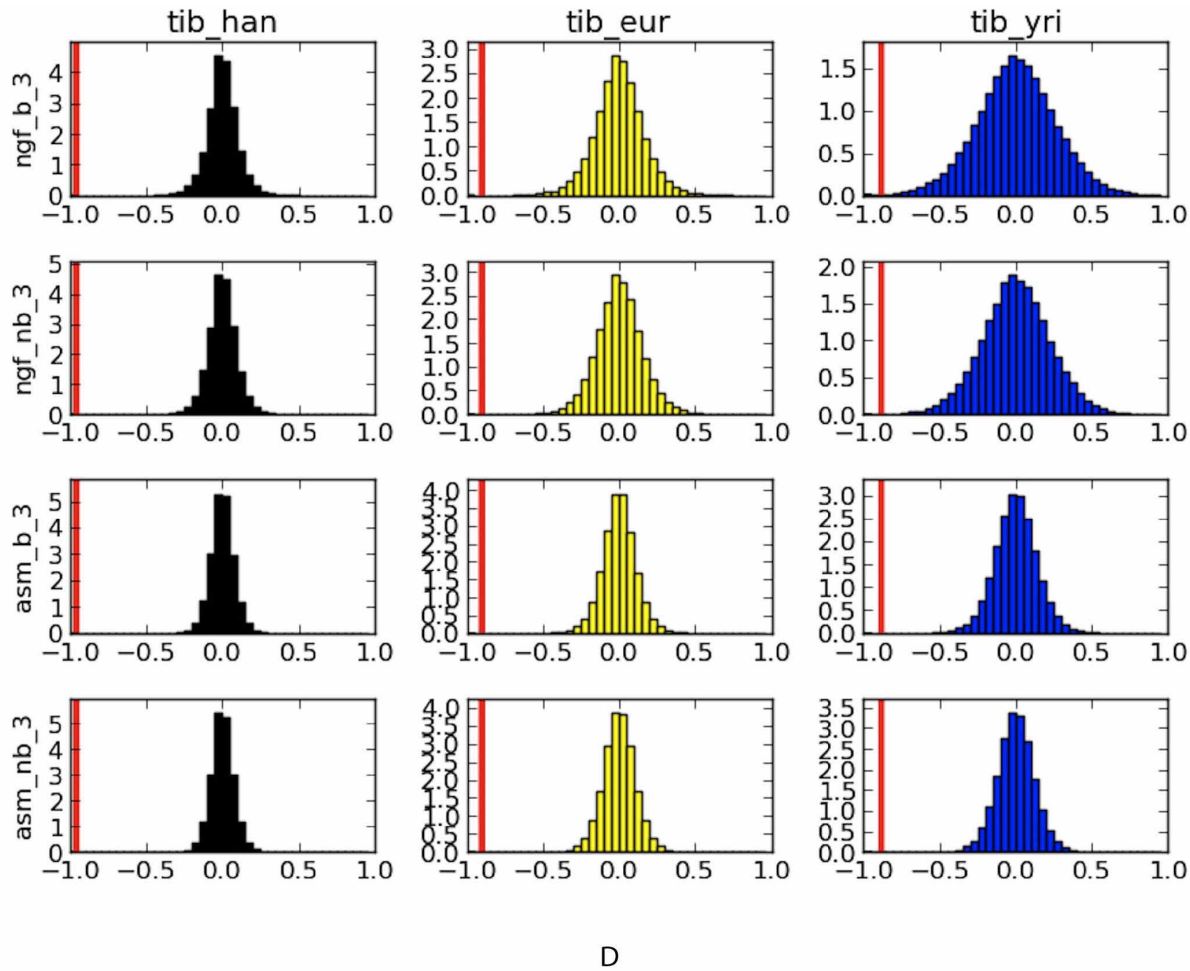


Extended Data Figure 6 | Other haplotype networks. **a**, A haplotype network based on the number of pairwise differences between 43 unique haplotypes defined from the 20 most differentiated SNPs between Tibetans and the 14 populations from the 1000 Genomes Project. The R software package pegas (ref. 22) was used to generate the figure. The haplotype distances are from pairwise differences. Each pie chart represents one unique haplotype and the size of the pie chart is proportional to $\log_2$(number of chromosomes with that haplotype). The sections in the pie provide the breakdown of the haplotypes amongst populations. The width of the edges is proportional to the number of pairwise differences between the joined haplotypes; the thinnest edge width represents a difference of one mutation. The number 57 next to a Tibetan haplotype is the number of Tibetan chromosomes with that haplotype. Similarly, the number 1,912 is the number of chromosomes (across several populations) with that haplotype. **b**, The number of pairwise differences between the Denisovan haplotype and the 43 unique haplotypes defined from the 20 most differentiated SNPs between Tibetans and the 14 populations from the 1000 Genomes Project (same haplotypes as in **a**). Each bar is a unique haplotype, and they are sorted in increasing order of pairwise differences. The colours within each bar represent the proportion of chromosomes with that haplotype broken down by populations. The numbers on top of each bar represent the total number of chromosomes within the 1000 Genomes data set and Tibetans that have the haplotype. Note this is the same data set used to create the haplotype network in panel **a**. Supplementary Tables 5 and 6 contain the 43 haplotypes and the frequencies within each of the populations.

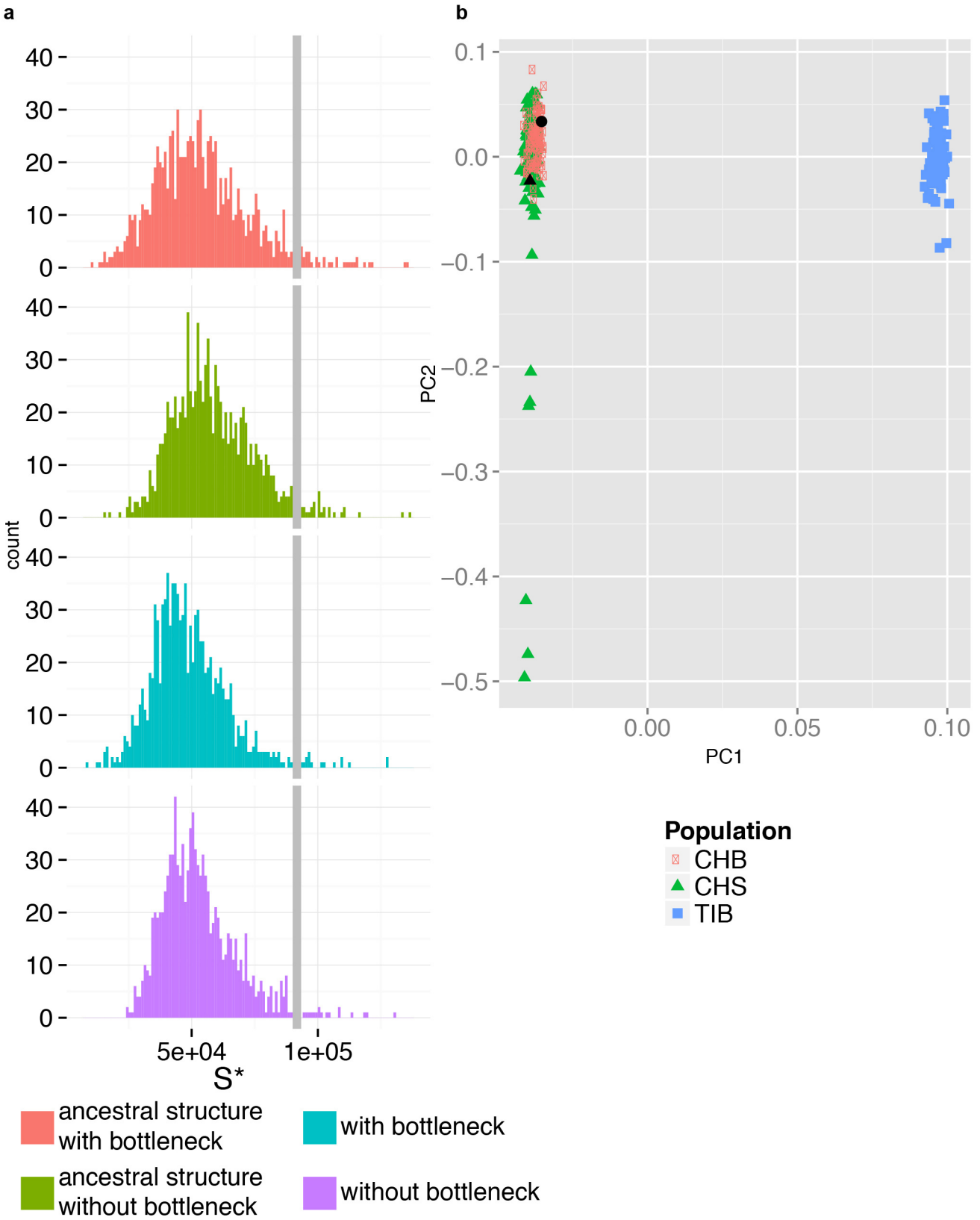
Extended Data Figure 7 | Number of pairwise differences. Red bars are the histograms of the number of pairwise differences between Denisovan and Tibetans. Blue bars are the histograms of the number of pairwise differences between Denisovan and GBR, CHS, FIN, PUR, CLM, IBS, CEU, YRI, CHB, JPT, LWK, ASW, MXL or TSI. All comparisons are within the 32.7-kb region of high differentiation (green box in Extended Data Fig. 1).



Extended Data Figure 8 | Divergence distributions. Modern human–Denisovan divergence (see Methods) for intronic regions of size 32.7 kb is plotted in red. Modern human–modern human divergence for the same intronic regions is plotted in blue. At the *EPAS1* 32.7-kb region, in green, is plotted the Tibetan–Han divergence. The black arrow points to the number of nucleotide differences between the Denisovan and the most common Tibetan haplotype (0.0038). This value is significantly lower than what we observe between modern human–Denisovan (red curve, $P = 0.0028$).



Extended Data Figure 9 | Null distributions of *D* for an assumed Tibet–Han divergence of 3,000 years. Each histogram corresponds to the *D* values obtained under null models without gene flow, and the red vertical bar corresponds to the *D* values observed in the real data. The observed *D* values are significant ($P < 0.001$) even when we assume Tibet–Han divergence of 5,000 or 10,000 years (see Methods and Supplementary Tables 8–10) (model abbreviations are given in the Supplementary Information; section on *D* statistics under models of no gene flow).



Extended Data Figure 10 | *S statistics and PCA plot.** **a**, A measure of introgression, *S**, from ref. 23. Distributions are for 1,000 simulations under the four demographic models described in the Supplementary Information; section on *D* statistics under models of no gene flow. *S** for the Tibetan individuals is shown as a vertical grey line. For all models, the empirical *P* values are 0.035, 0.028, 0.019 and 0.017, respectively, for each model (top to bottom). **b**, Plots the first and second principal components using all the CHS (100 individuals) and the CHB (97 individuals) from the 1000 Genomes and the 77 Tibetan individuals from ref. 45 (see Methods). The black circle and the black triangle represent the single CHB and the CHS individuals carrying the five-SNP Tibetan–Denisovan-haplotype (Extended Data Fig. 3). All SNPs in the intersection between the 1000 Genomes populations and the 77 Tibetan individuals from chromosome 2 were used for this analysis.