LETTER

doi:10.1038/nature12961

The genomic landscape of Neanderthal ancestry in present-day humans

Sriram Sankararaman[1,2], Swapan Mallick[1,2], Michael Dannemann[3], Kay Prüfer[3], Janet Kelso[3], Svante Pääbo[3], Nick Patterson[1,2] & David Reich[1,2,4]

Genomic studies have shown that Neanderthals interbred with modern humans, and that non-Africans today are the products of this mixture[1,2]. The antiquity of Neanderthal gene flow into modern humans means that genomic regions that derive from Neanderthals in any one human today are usually less than a hundred kilobases in size. However, Neanderthal haplotypes are also distinctive enough that several studies have been able to detect Neanderthal ancestry at specific loci[1,3–8]. We systematically infer Neanderthal haplotypes in the genomes of 1,004 present-day humans[9]. Regions that harbour a high frequency of Neanderthal alleles are enriched for genes affecting keratin filaments, suggesting that Neanderthal alleles may have helped modern humans to adapt to non-African environments. We identify multiple Neanderthal-derived alleles that confer risk for disease, suggesting that Neanderthal alleles continue to shape human biology. An unexpected finding is that regions with reduced Neanderthal ancestry are enriched in genes, implying selection to remove genetic material derived from Neanderthals. Genes that are more highly expressed in testes than in any other tissue are especially reduced in Neanderthal ancestry, and there is an approximately fivefold reduction of Neanderthal ancestry on the X chromosome, which is known from studies of diverse species to be especially dense in male hybrid sterility genes[10–12]. These results suggest that part of the explanation for genomic regions of reduced Neanderthal ancestry is Neanderthal alleles that caused decreased fertility in males when moved to a modern human genetic background.

To search systematically for Neanderthal haplotypes, we developed a method based on a conditional random field[13] (CRF) that combines information from three features of genetic variation that are informative of Neanderthal ancestry (Supplementary Information section 1 and Extended Data Fig. 1). The first is the allelic pattern at a single nucleotide polymorphism (SNP): if a non-African individual carries a derived allele seen in Neanderthals but absent from the west-African Yoruba from Ibadan, Nigeria (YRI), the allele is likely to originate from Neanderthals. The second is high sequence divergence of the non-African haplotype to all YRI haplotypes but low divergence to Neanderthal. The third is a haplotype length consistent with interbreeding 37–86-thousand years ago[14]. We trained the CRF using simulations[15], and established its robustness to deviations from the assumed demography (Supplementary Information section 2).

We screened for Neanderthal haplotypes in the 1000 Genomes Project Phase 1 (1KG) data[9], using the Altai Neanderthal genome of 52-fold average coverage to determine alleles present in Neanderthals[2], a six-primate consensus to determine ancestral alleles[16], and 176 YRI genomes as a reference panel assumed to harbour no Neanderthal ancestry (Fig. 1a). Table 1 reports the mean and standard deviation across individuals of the fraction of their ancestry confidently inferred to be Neanderthal (probability > 90%). Figure 1b and Extended Data Fig. 2 plot the fraction of European ($n = 758$) and east-Asian ($n = 572$) haplotypes that descend from Neanderthals at each genomic location (Supplementary Information section 3). We created a tiling path of inferred Neanderthal haplotypes that spans 1.1 gigabases (Gb) over 4,437 contigs (Supplementary Information section 4), thus filling in gaps in the Neanderthal sequence over a number of repetitive regions that cannot be reconstructed from short ancient DNA fragments (Extended Data Fig. 3).

Four features of the Neanderthal introgression map suggest that it is producing reasonable results. First, when we infer Neanderthal ancestry using low-coverage data from Croatian Neanderthals[1] we obtain correlated inferences (Spearman rank correlation $\rho = 0.88$ in Europeans; Supplementary Information section 3). Second, in the African Luhya in Webuye, Kenya (LWK), the proportion of the genome inferred to be Neanderthal is 0.08%, an order of magnitude smaller than in non-African populations (Table 1). Third, the proportion of the genome with confidently inferred Neanderthal ancestry has a mean of 1.38% in east-Asian and 1.15% in European populations (Table 1), consistent with previous reports of more Neanderthal ancestry in east-Asian than

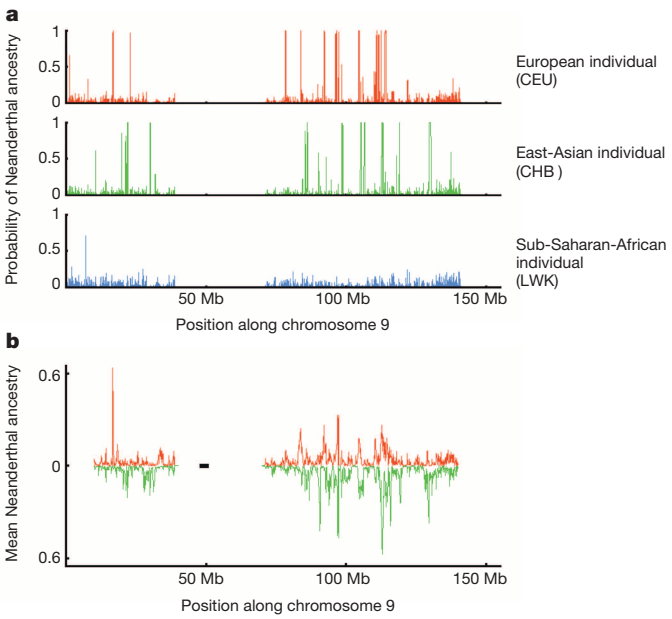


Figure 1 | Maps of Neanderthal ancestry. **a**, Individual maps; the marginal probability of Neanderthal ancestry for one European-American, one east-Asian and one sub-Saharan-African phased genome across chromosome 9. **b**, Population maps; estimate of the proportion of Neanderthal ancestry in European individuals (red) and east-Asian individuals (green), averaged across all individuals from each population in non-overlapping 100-kb windows on chromosome 9. The black bar denotes the coordinates of the centromere. The plot is limited to segments of the chromosome that pass filters (see Supplementary Information section 8). CEU, residents of Utah, US, with northern and western European ancestry (from the Centre d'Etude du Polymorphisme Humain (CEPH) collection); CHB, Han Chinese in Beijing, China; LWK, African Luhya in Webuye, Kenya.

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [3]Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany. [4]Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA.

©2014 Macmillan Publishers Limited. All rights reserved

Table 1 | Genome-wide estimates of Neanderthal ancestry

Region	Population	Number of individuals	Neanderthal ancestry on autosomes (%)	Neanderthal ancestry on the X chromosome (%)
Europe	CEU	85	1.17 ± 0.08	0.21 ± 0.17
	FIN	93	1.20 ± 0.07	0.19 ± 0.14
	GBR	89	1.15 ± 0.08	0.20 ± 0.15
	IBS	14	1.07 ± 0.06	0.23 ± 0.18
	TSI	98	1.11 ± 0.07	0.25 ± 0.20
East Asia	CHB	97	1.40 ± 0.08	0.30 ± 0.21
	CHS	100	1.37 ± 0.08	0.27 ± 0.21
	JPT	89	1.38 ± 0.10	0.26 ± 0.21
America	CLM	60	1.14 ± 0.12	0.22 ± 0.16
	MXL	66	1.22 ± 0.09	0.21 ± 0.15
	PUR	55	1.05 ± 0.12	0.20 ± 0.15
Africa	LWK	97	0.08 ± 0.02	0.04 ± 0.07
	ASW	61	0.34 ± 0.22	0.07 ± 0.11

For each computationally phased genome in each population, we estimated the probability of Neanderthal ancestry at each SNP and the fraction of autosomal and X chromosome SNPs that are confidently of Neanderthal origin in each individual (marginal probability >90%). The table reports the average and standard deviation of this statistic across individuals within each population. ASW, people with African ancestry in Southwest United States; CEU, Utah residents with northern and western European ancestry (from the Centre d'Etude du Polymorphisme Humain (CEPH) collection); CHB, Han Chinese in Beijing, China; CHS, Han Chinese in South China; CLM, Colombians in Medellin, Colombia; FIN, Finnish in Finland; GBR, British from England and Scotland, UK; IBS, Iberian populations in Spain; JPT, Japanese in Tokyo, Japan; LWK, African Luhya in Webuye, Kenya; MXL, people with Mexican ancestry in Los Angeles, California; PUR, Puerto Ricans in Puerto Rico; TSI, Toscani in Italy.

in European populations [7,17]. Fourth, the standard deviation in Neanderthal ancestry among individuals from within the same population is 0.06–0.10%, in line with theoretical expectation (Supplementary Information section 3), showing that Neanderthal ancestry calculators that estimate differences on the order of a per cent[18] are largely inferring statistical noise.

The Neanderthal introgression map reveals locations where Neanderthal ancestry is inferred to be as high as 62% in east-Asian and 64% in European populations (Fig. 1b and Extended Data Fig. 2). Several of these regions provide evidence of positive selection if we assume a model in which the distribution of Neanderthal ancestry has been governed by neutral drift; however, this assumption is problematic in light of the evidence for widespread negative selection against Neanderthal ancestry reported below (Supplementary Information section 5). To further explore whether Neanderthal alleles may have been affected by positive selection, we examined the 5% of genes with the highest inferred Neanderthal ancestry. We do not detect tissue-specific expression patterns; however genes involved in keratin filament formation and some other biological pathways are significantly enriched in Neanderthal ancestry in European populations, east-Asian populations, or both (Extended Data Table 1 and Supplementary Information section 6). Thus, Neanderthal alleles that affect skin and hair may have helped modern humans to adapt to non-African environments. We also investigated the relevance of Neanderthal alleles to present-day human biology by identifying alleles of Neanderthal origin (Supplementary Information section 7), and overlapping this list with alleles that have been associated with phenotypes of medical relevance[19,20]. We identify alleles of Neanderthal origin that affect lupus, biliary cirrhosis, Crohn's disease, optic-disk size, smoking behaviour, IL-18 levels and type 2 diabetes[20] (Extended Data Table 2).

The most striking feature of the introgression map is large 'deserts' of Neanderthal ancestry: on a 10-megabase (Mb) scale on the autosomes, there are 4 windows in European and 14 in east-Asian populations with Neanderthal ancestry $< 0.1\%$ (Extended Data Fig. 2 and Supplementary Information section 8). Two analyses show that these deserts are not artefacts of reduced power to detect ancestry. First, these regions were detected using a probability threshold for calling a segment as Neanderthal of >25%, which results in a much higher sensitivity to true segments of Neanderthal ancestry than does a threshold of >90% (Supplementary Information section 3). Second, when we estimate Neanderthal ancestry in regions of low recombination rate where Neanderthal haplotypes are longer so that we have more power to detect them, we see a decreased Neanderthal ancestry proportion, opposite to the expectation from increased power ($\rho = 0.221$, $P = 4.4 \times 10^{-4}$ in European individuals; $\rho = 0.226$, $P = 1.9 \times 10^{-4}$ in east-Asian individuals) (Supplementary Information section 8). As we also observe multi-megabase regions of increased Neanderthal ancestry, part of the explanation for the ancestry deserts is likely to be small population sizes shortly after interbreeding (Supplementary Information section 8). However, selection also seems to have contributed to Neanderthal ancestry deserts, as we detect a correlation to functionally important regions (below).

To explore whether selection provides part of the explanation for regions of reduced Neanderthal ancestry, we tested for a correlation of Neanderthal ancestry to a previously published 'B statistic', in which low B implies a high density of functionally important elements[21]. We find that low B is significantly correlated to low Neanderthal ancestry: $\rho = 0.32$ in European populations ($P = 4.9 \times 10^{-87}$) and $\rho = 0.31$ in east-Asian populations ($P = 3.88 \times 10^{-68}$) (Fig. 2 and Supplementary Information section 8). This is not an artefact of reduced power, as there is expected to be reduced genetic variation in regions of low B which should make introgressed Neanderthal haplotypes stand out more clearly (Extended Data Table 3 and Supplementary Information section 2). We also estimated Neanderthal ancestry in quintiles of B statistic using an approach that is not biased by varying mutation rates, recombination rates, or genealogical tree depth[22], and confirmed that the quintile with the highest B has significantly higher Neanderthal

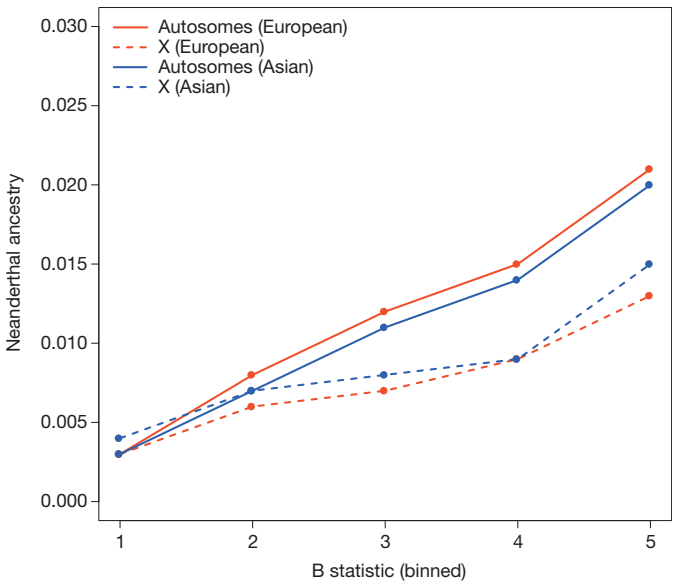


Figure 2 | Functionally important regions are deficient in Neanderthal ancestry. The median of the proportion of Neanderthal ancestry (estimated as the average over the marginal probability of Neanderthal ancestry assigned to each individual allele at a SNP) within quintiles of a B statistic that measures proximity to functionally important regions (1-low, 5-high). We show results on the autosomes and the X chromosome, and in European and east-Asian populations.

©2014 Macmillan Publishers Limited. All rights reserved

Table 2 | Enrichment of tissue-specific genes in regions deficient in Neanderthal ancestry

Tissue	European whole genome	European X chromosome	European autosomes	East-Asian whole genome	East-Asian X chromosome	East-Asian autosomes
Adipose	0.93	1	0.81	0.99	1	0.95
Adrenal	0.5	NA	0.5	0.42	NA	0.42
Blood	0.99	0.98	0.99	0.94	0.73	0.94
Brain	1	1	1	1	1	1
Breast	0.98	0.63	0.99	1	0.94	1
Colon	0.64	0.77	0.63	0.94	0.97	0.89
Heart	0.99	0.71	0.99	0.8	0.57	0.81
Kidney	1	0.15	1	1	0.08	1
Liver	0.99	0.99	0.99	1	0.86	1
Lung	0.96	0.64	0.96	0.99	0.87	0.99
Lymph	0.88	0.62	0.9	0.99	0.51	0.99
Ovary	0.84	0.95	0.81	0.62	0.91	0.58
Prostate	1	0.79	1	1	0.73	1
Muscle	0.95	0.7	0.95	0.83	0.1	0.88
Testes	0.0095	0.13	0.016	0.018	0.039	0.055
Thyroid	0.86	0.62	0.88	0.87	0.94	0.86

Tissue-specific genes (defined as those that are significantly more highly expressed in the specified tissue than in any of the 15 other tissues) are compared to all other expressed genes in that tissue. Of the 16 tissues tested, only testes-specific genes are significantly enriched in the regions deficient in Neanderthal ancestry, defined as locations in which all sites across all individuals are assigned a marginal probability of Neanderthal ancestry of <10% (47% of genes in European individuals and 52% of genes in east-Asian individuals fall into this category). NA, no tissue-specific genes for this tissue on the X chromosome. Units are *P* values.

ancestry than the other quintiles ($P = 7 \times 10^{-4}$) (Extended Data Table 4 and Supplementary Information section 9).

The largest deserts of Neanderthal ancestry are on the X chromosome, where the mean Neanderthal ancestry is about a fifth of the autosomes (Table 1). The power of our CRF to detect Neanderthal ancestry is higher on the X chromosome than on the autosomes (Extended Data Table 5 and Supplementary Information section 2), implying that this observation cannot be an artefact of reduced power. At least some of the reduction in Neanderthal ancestry that we observe on the X chromosome must be due to selection, since—just as on the autosomes—we observe that Neanderthal ancestry is positively correlated with B statistic ($\rho = 0.276, P = 3.1 \times 10^{-4}$ for European populations; $\rho = 0.176$, $P = 0.02$ for east-Asian populations) (Fig. 2 and Supplementary Information section 8). Studies in many species have shown that genes responsible for reduced male fertility disproportionally map to the X chromosome (refs 10–12). We reasoned that this 'large X effect'[23] could explain why the X chromosome was more resistant to introgression of Neanderthal ancestry than the autosomes.

If male hybrid sterility is contributing to our observations, a prediction is that the responsible genes will be disproportionally expressed in testes[24]. To test this hypothesis, we analysed gene transcripts from 16 human tissues[25] and defined 'tissue-specific' genes as those with a significantly higher expression level in that tissue than any other. We found that only genes that are specific to testes were enriched in regions of low Neanderthal ancestry (when compared with all other genes expressed in the same tissue). This effect remained significant after permuting gene annotations while preserving the correlation structure between Neanderthal ancestry and gene expression ($P = 0.0095$ in European popululatons; $P = 0.018$ in east-Asian populations) (Table 2 and Supplementary Information section 6). However, hybrid sterility is not the only factor responsible for selection against Neanderthal material, as Neanderthal ancestry is also depleted in conserved pathways such as RNA processing ($P < 0.05$; Extended Data Table 2 and Supplementary Information section 6).

We have shown that interbreeding of Neanderthals and modern humans introduced alleles onto the modern human genetic background that were not tolerated, which probably resulted in part from their contributing to male hybrid sterility. The resulting reduction in Neanderthal ancestry was quantitatively large: in the fifth of the genome with highest B, Neanderthal ancestry is 1.54 ± 0.15 times the genome-wide average (Extended Data Table 4 and Supplementary Information section 9)[22]. If we assume that this subset of the genome was unaffected by selection, this implies that the proportion of Neanderthal ancestry shortly after introgression must have been >3% rather than the approximately 2% seen today. The large effect of negative selection on present-day levels of Neanderthal ancestry may explain why the proportion of Neanderthal ancestry is significantly higher in present-day east-Asian than in European populations (Table 1)[7,17]; there is evidence that east-Asian population sizes have been smaller than European populations for some of the time since their separation[26], which could have resulted in less efficient selection to remove Neanderthal-derived deleterious alleles. The evidence for male hybrid sterility is particularly remarkable when compared with mixed populations of present-day humans in which no convincing signals of selection against alleles inherited from one of the mixing populations have been found despite high power to detect such effects[27]. Thus, although the time of separation between Neanderthals and modern humans was only about five times larger than that between present-day European and west-African populations[2], the biological incompatibility was far greater. A potential explanation is the 'snowball effect', whereby hybrid sterility genes are expected to accumulate in proportion to the square of the substitutions between two taxa because two interacting loci need to change to produce an incompatibility ('Dobzhansky-Muller incompatibilities')[28]. An important direction for future work is to explore whether similar phenomena have affected other interbreeding events between diverged hominin groups.

METHODS SUMMARY

We infer the distribution of Neanderthal ancestry at each allele in a phased test genome using a conditional random field (CRF)[13]. The data we use consist of a panel of African reference genomes, the high-coverage Neanderthal genome, and a genetic map. For a set of CRF parameters, we can compute the probability of Neanderthal ancestry at each allele using the forward–backward algorithm (Supplementary Information section 1). Parameter estimation in the CRF needs training data (haplotypes labelled with true Neanderthal ancestries). We estimated the CRF parameters from data simulated under a demography relating European individuals, west-African individuals and Neanderthals (Supplementary Information section 2). We assessed the robustness of the CRF by measuring its false discovery rate in simulated demographic models that are different from the ones used in the simulation to train the CRF (Supplementary Information section 2). We also assessed the robustness to errors in phasing and in the genetic map (Supplementary Information section 2). For some analyses we estimate Neanderthal ancestry proportion based on all alleles where the marginal probability of Neanderthal ancestry as inferred by the CRF is greater than a specified threshold. For other analyses, we compute it as the average marginal probability of Neanderthal ancestry.

Online Content Any additional Methods, 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 5 September; accepted 18 December 2013.
Published online 29 January 2014.

1. Green, R. E. *et al.* A draft sequence of the Neanderthal genome. *Science* **328,** 710–722 (2010).
2. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).

©2014 Macmillan Publishers Limited. All rights reserved

3. Abi-Rached, L. *et al.* The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* **334,** 89–94 (2011).
4. 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).
5. Mendez, F. L., Watkins, J. C. & Hammer, M. F. Neanderthal origin of genetic variation at the cluster of OAS immunity genes. *Mol. Biol. Evol.* **30,** 798–801 (2013).
6. Yotova, V. *et al.* An X-linked haplotype of Neanderthal origin is present among all non-African populations. *Mol. Biol. Evol.* **28,** 1957–1962 (2011).
7. Wall, J. D. *et al.* Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics* **194,** 199–209 (2013).
8. Lachance, J. *et al.* Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150,** 457–469 (2012).
9. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
10. Tucker, P. K., Sage, R. D., Wilson, A. C. & Eichler, E. M. Abrupt cline for sex chromosomes in a hybrid zone between two species of mice. *Evolution* **46,** 1146–1163 (1992).
11. Good, J. M., Dean, M. D. & Nachman, M. W. A complex genetic basis to X-linked hybrid male sterility between two species of house mice. *Genetics* **179,** 2213–2228 (2008).
12. Presgraves, D. C. Sex chromosomes and speciation in *Drosophila. Trends Genet.* **24,** 336–343 (2008).
13. Lafferty, J., McCallum, A. & Pereira, F. C. N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proc. 18th Int. Conf. Machine Learn.* 282–289 (2001).
14. Sankararaman, S., Patterson, N., Li, H., Paabo, S. & Reich, D. The date of interbreeding between Neanderthals and modern humans. *PLoS Genet.* **8,** e1002947 (2012).
15. Hellenthal, G. & Stephens, M. msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics* **23,** 520–521 (2007).
16. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18,** 1814–1828 (2008).
17. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
18. Durand, E. Y. *Neanderthal Ancestry Estimator* White paper 23-05 http://23andme.https.internapcdn.net/res/pdf/hXitekfSJe1lcly7-Q72XA_23-05_Neanderthal_Ancestry.pdf (23andMe, 2011).
19. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
20. The SIGMA Type 2 Diabetes Consortium.. Sequence variants in *SLC16A11* are a common risk factor for type 2 diabetes in Mexico. *Nature* http://dx.doi.org/10.1038/nature12828 (25 December 2014).
21. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5,** e1000471 (2009).
22. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468,** 1053–1060 (2010).
23. Coyne, J. A. O. H. A. *Speciation and Its Consequences* (eds Otte, D. & Endler, J. A.) 180–207 (Sinauer Associates, 1989).
24. Wu, C.-I. & Davis, A. W. Evolution of postmating reproductive isolation: the composite nature of Haldane's rule and its genetic basis. *Am. Nat.* **142,** 187–212 (1993).
25. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22,** 1775–1789 (2012).
26. Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genet.* **39,** 1251–1255 (2007).
27. Bhatia, G. *et al.* Genome-wide scan of 29,141 African Americans finds no evidence of selection since admixture. Preprint at http://arxiv.org/pdf/1312.2675.pdf (2013).
28. Orr, H. A. & Turelli, M. The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. *Evolution* **55,** 1085–1094 (2001).

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

Acknowledgements We thank A. Briggs, P. Moorjani, M. Przeworski, D. Presgraves and A. Williams for critical comments, and K. Kavanagh for help with Extended Data Fig. 2. We are grateful for support from the Presidential Innovation Fund of the Max Planck Society, NSF HOMINID grant 1032255 and NIH grant GM100233. S.S. was supported by a post-doctoral fellowship from the Initiative for the Science of the Human Past at Harvard University. D.R. is a Howard Hughes Medical Institute Investigator.

Author Contributions S.S., N.P., S.P. and D.R. conceived of the study. S.S., S.M. M.D., K.P., J.K. and D.R. performed analyses. J.K., S.P., N.P. and D.R. supervised the study. S.S. and D.R. wrote the manuscript with help from all co-authors.

Author Information The tiling path of confidently inferred Neanderthal haplotypes, as well as the Neanderthal introgression map, can be found at http://genetics.med.harvard.edu/reichlab/Reich_Lab/Datasets.html. 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 S.S. (sankararaman@genetics.med.harvard.edu) or D.R. (reich@genetics.med.harvard.edu).

©2014 **Macmillan Publishers Limited. All rights reserved**

METHODS

Conditional random field for inferring Neanderthal local ancestry. For a haploid genome in a test population that carries Neanderthal ancestry (for example, Europeans), given the allelic states of a sequence of SNPs along this haplotype, we wish to infer the ancestral state of the allele at each SNP; specifically, whether it has entered modern humans through Neanderthal gene flow. In addition to the test haplotype, the data analysed consist of a panel of haplotypes from the sub-Saharan-African Yoruba (YRI) who we assume harbour no Neanderthal ancestry[1]. To determine the allelic state of the Neanderthals, we used a high-coverage Neanderthal genome[2]. We determined the ancestral and derived allele at each SNP using a six-primate consensus sequence[16]. To estimate the genetic distance between adjacent SNPs, we used the Oxford combined linkage-disequilibrium map[29]. We specified the distribution of the unobserved Neanderthal ancestry states at each SNP given the observed genetic data as a conditional random field (CRF)[13]. We specified CRF feature functions that relate the observed data and the unobserved ancestral state at each SNP ('emission functions') as well as feature functions that relate the unobserved ancestral states at adjacent SNPs ('transition functions'). Thus, the model is a linear-chain CRF. The feature functions and their associated parameters fully specify the distribution of the unobserved ancestral states given the observed data. Given the parameters and the observed data, we were able to infer the marginal probability of Neanderthal ancestry at each SNP of the haploid genome. We computed the marginal probabilities efficiently using the forward–backward algorithm[13,30]. Supplementary Information section 1 presents the mathematical details.

Feature functions. The emission functions couple the unobserved ancestral state at a SNP to the observed features. We used two classes of emission functions.

The first class of emission functions captures information from the joint patterns observed at a single SNP across European and African individuals, and Neanderthals. These features are indicator functions that assume the value '1' when a specific pattern is observed at a SNP and '0' otherwise. We used feature functions that pick out two classes of allelic patterns. One of these features is 1 if at a given SNP, the test haplotype carries the derived allele, all the YRI haplotypes carry the ancestral allele, and either of the two Neanderthal alleles is derived. SNPs with this joint configuration have an increased likelihood of Neanderthal ancestry. In the CRF, an increased likelihood associated with this feature is reflected in the fact that the parameter is positive with a magnitude determined by the informativeness of the feature. The second feature is 1 if at a given SNP the test haplotype carries a derived allele that is polymorphic in the panel of African individuals but absent in the Neanderthal. SNPs with this joint configuration have a decreased likelihood of Neanderthal ancestry.

The second class of emission functions uses multiple SNPs to capture the signal of Neanderthal ancestry. Specifically, we compared the divergence of the test haplotype to the Neanderthal sequence to the minimum divergence of the test haplotype to all African haplotypes over non-overlapping 100-kilobase (kb) windows (the size scale we expect for Neanderthal haplotypes today based on the time of Neanderthal gene flow into modern humans[14]). In a region of the genome in which the test haplotype carries Neanderthal ancestry, we expect the test haplotype to be closer to the Neanderthal sequence than to most modern human sequences (albeit with a large variance), and we expect the pattern to be reversed outside these regions. We chose the derived allele at heterozygous sites so that this distance was effectively the minimum distance of the potentially introgressed test haplotype to one of the two Neanderthal haplotypes.

The transition feature function modulates the correlation of the ancestral states at adjacent SNPs. We defined this feature function as an approximation, at small genetic distance, to the log of the transition probabilities of a standard Markov process of admixture between two populations. This approximation makes parameter estimation in the CRF efficient.

Parameter estimation. To estimate the parameters of the CRF, we needed haplotypes labelled with Neanderthal ancestries; that is, training data. As we do not in fact know the true Neanderthal state in any individual, we estimated the CRF parameters on data simulated under a demographic model. We estimated parameters by maximizing the L2-regularized conditional log likelihood[30] using a limited-memory version of LBFGS (limited memory Broyden–Fletcher–Goldfarb–Shanno)[31]. We fixed the value of the parameter associated with the L2 penalty at 10 although a broad range of values seem to work in practice. We assumed a simple demographic model relating African and European individuals and Neanderthals with Neanderthal–modern human admixture occurring 1,900 generations ago[14] and a fraction of Neanderthal ancestry of 3%[1]. The model parameters were broadly constrained by the observed allele frequency differentiation between the west-African YRI and European-American CEU populations, and by the observed excess sharing of alleles between European and African populations relative to Neanderthals. The simulations incorporated hotspots of recombination[11] as well as the reduced power to detect low-frequency alleles from low-coverage sequencing data[12].

Validation of the conditional random field. We assessed the accuracy of the CRF to predict Neanderthal ancestry using simulated data. Given the marginal probabilities estimated by the CRF, we estimated the precision (fraction of predictions that are truly Neanderthal) and the recall (fraction of true Neanderthal alleles that are predicted) as we varied the threshold on the marginal probability for an allele to be declared Neanderthal. We also evaluated the accuracy when the haplotype phase needed to be inferred and when the genetic map had errors. As the CRF parameter estimation assumes a specific demographic model, we were concerned about the possibility that the inferences might be sensitive to the model assumed. We therefore varied each demographic parameter in turn and applied the CRF to data simulated under these perturbed models, fixing the parameters of the CRF to the estimates obtained under the original model. For each of these perturbed models, we evaluated the false discovery rate (defined as one minus the precision) when we restricted to sites at which the CRF assigns a marginal probability of at least 0.90. Supplementary Information section 2 presents the details.

Preparation of the 1000 Genomes data. We applied the CRF to the computationally phased haplotypes in each of the 13 populations in the 1000 Genomes Project[9] (1KG), excluding the west-African Yoruba (YRI). The CRF requires reference genomes from African individuals and Neanderthals. For the African population, we used 176 haplotypes from 88 YRI individuals. For the Neanderthal genome, for most analyses we used the genotypes called from the recently generated high-coverage Neanderthal sequence[2]. We restricted our analysis to sites passing the filters described in ref. 2 and for which the genotype quality score was $\geq$30. These filters discarded sites that are identified as repeats by the Tandem Repeat Finder annotation for hg19 from the University of California Santa Cruz (UCSC) genome browser (available at http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/simpleRepeat.txt.gz), that have Phred-scaled mapping quality scores of $<$30, or that map to regions in which the alignment is ambiguous or which fall within the upper or lower 2.5th percentile of the sample-specific coverage distribution (applied within the regions of unique mappability binned according to the GC content of the reference genome). For the mappability filter, we used the liberal filter that requires that at least 50% of all 35-mers that overlap a position do not map to any other position in the genome allowing up to one mismatch. We further restricted our analysis to sites that are biallelic across the Neanderthal and the 1000 Genomes Project samples. After filtering, we were able to analyse 26,493,206 SNPs on the autosomes and 817,447 SNPs on the X chromosome. For each haplotype analysed, we also restricted to the set of polymorphic sites in the population containing the haplotype. We obtained genetic distances from the Oxford combined linkage-disequilibrium map lifted over to hg19 coordinates[29]. For the X chromosome, we obtained an appropriate sex-averaged map by scaling the X chromosome linkage-disequilibrium-based map by two-thirds.

Statistics for measuring Neanderthal ancestry. We computed several statistics to summarize the Neanderthal ancestry inferred by the CRF. We estimated the proportion of an individual diploid genome that is confidently inferred to be Neanderthal as the fraction of sites for which the marginal probability is $\geq$90%. To assess variation in the proportion of Neanderthal ancestry along the genome, we computed the fraction of alleles across individuals with marginal probability greater than a specified threshold. This statistic is likely to be affected by variation in power along the genome. Hence, we also consider an estimate of the ancestry proportion obtained by averaging the marginal probability across individuals. Depending on the analyses, these statistics are estimated at a single SNP or in non-overlapping windows of a specified size.

To assess whether the predictions made by the CRF are sensible, we inferred Neanderthal ancestry using the low-coverage genome from the Vindija Neanderthals[1]. For this analysis, we restricted to sites at which there is at least one read with mapping quality score between 60 and 90 and base quality of at least 40. As a second validation analysis, we applied the CRF to the sub-Saharan-African Luhya (LWK), using the parameters estimated with non-African individuals. We empirically assessed the accuracy of the CRF on the 1000 Genomes Project data by assuming that LWK has no Neanderthal ancestry, that the false discovery rate in each non-African population is equal to the false discovery rate in LWK, and using the genome-wide proportion of Neanderthal ancestry estimated in ref. 2. (Supplementary Information section 3). We computed the theoretical standard deviation in the proportion of Neanderthal ancestry[32] assuming a pulse model of admixture with 2% Neanderthal ancestry followed by 2,000 generations of random mating, and 2.03 gigabases as the number of bases of the high-coverage Neanderthal genome that pass filters.

Tiling path of Neanderthal haplotypes. We identified Neanderthal haplotypes as runs of consecutive alleles along a test haplotype assigned a marginal probability of $>$90%. We filtered haplotypes smaller than 0.02 centiMorgans (cM). At each SNP that is covered by at least one such haplotype, we estimated the allelic state as the

©2014 Macmillan Publishers Limited. All rights reserved

consensus allele across the spanning haplotypes. See Supplementary Information section 4.

Functional analysis of introgressed alleles. We defined two subsets of consensus coding sequence (CCDS) genes[33]. We define a gene with 'low Neanderthal ancestry' as one in which all alleles across all individuals have a marginal probability ≤10%. We also require that the genes included in this analysis include at least ≥100 SNPs within a 100-kb window centred at its midpoint (this excluded genes at which the power to infer Neandertal ancestry is reduced due to a reduced number of informative sites). We define a gene with 'high Neanderthal ancestry' as one that is in the top 5% of CCDS genes ranked by the average marginal probability across individual haplotypes. See Supplementary Information section 6 for details.

Functional enrichment analysis. We tested for enrichment of Gene Ontology (GO)[34] categories in genes with low or high Neanderthal ancestry, using the hypergeometric test implemented in the FUNC package[35]. We report multiple-testing corrected *P* values estimated from 1,000 permutations for the GO enrichment analysis (family-wise error rate, FWER). Given the observed correlation between Neanderthal ancestry and B statistic[21], a concern is that the functional categories may not be randomly distributed with respect to B-statistic. To control for this, we assigned a B statistic to each gene (estimated as an average of the B statistic over the length of the gene) as well as a uniform random number. This resulted in 17,249 autosomal genes. Genes were binned into 20-equal sized bins based on the gene-specific B statistic. Within each bin, genes were sorted by Neanderthal ancestry and then by the random number. Genes ranked within the top 5% within each bin were used for the analysis. See Supplementary Information section 6 for details.

Identification of alleles born in Neanderthals and cross-correlation with association-study data. To infer whether an allele segregating in a present-day human population was introduced by Neanderthal gene flow, we defined a probable Neanderthal allele as one with marginal probability of ≥90% and a non-Neanderthal allele as having a marginal probability of ≤10%. A SNP at which all of the confident non-Neanderthal alleles as well as all alleles in YRI are ancestral and all of the confident Neanderthal alleles are derived is inferred to be of Neanderthal origin. This procedure allows for some false negatives in the prediction of the CRF. This procedure yields 97,365 Neanderthal-derived SNPs when applied to the predictions in European and east-Asian individuals. We downloaded the variants listed in the NHGRI (National Human Genome Research Institute) GWAS catalogue[19], retaining entries for which the reported association is a SNP with an assigned rs number, and for which the nominal *P* value is less than 5×10^{-8}. This resulted in 5,022 associations, which we then intersected with the Neanderthal-derived list. See Supplementary Information section 7 for details.

Identification of genomic regions deficient in Neanderthal ancestry. We measured the fraction of alleles across individuals and SNPs that are assigned a marginal probability greater than a chosen threshold measured within non-overlapping 10-megabase (Mb) windows. We chose a threshold of 25% as this threshold was found to lead to high recall in our empirical assessment (Supplementary Information section 3). We reported windows for which this statistic is <0.1%.

To understand the causes for variation in Neanderthal ancestry, we tested for correlation to a B statistic. Each SNP was annotated with the B statistic lifted over to hg19 coordinates. We assessed correlation between B and estimates of Neanderthal ancestry proportion at a nucleotide level as well as at different size scales, and separately on the autosomes and the X-chromosome (Supplementary Information section 8). We also used wavelet decomposition[36] to analyse the correlation of the inferred Neanderthal ancestry between European and east-Asian individuals at multiple size scales (Supplementary Information section 10). Figure 2 reports the relation between the mean marginal probability of Neanderthal ancestry across individuals and quintiles of B statistic at each SNP. To assess significance, we estimated Spearman's correlation ρ and standard errors on ρ using a block jackknife[37] with 10-Mb blocks (Supplementary Information section 8).

To understand the contribution of demography to variation in Neanderthal ancestry along the genome, we measured the coefficient of variation, at a 10-Mb scale, of the proportion of ancestry estimated as defined above. We then applied the CRF to data simulated under diverse demographic models and compared the coefficient of variation to the observed value (Supplementary Information section 8).

Assessment of the power to infer Neanderthal ancestry. To assess the power of the CRF to infer Neanderthal ancestry, we simulated data under diverse demographic models[38]. In one simulation series, we varied effective population size to approximate the effect of background selection and measured recall at a precision of 90%. In a second series, we assessed power on the X-chromosome versus the autosomes by matching the effective population size, recombination rate and mutation rate to estimated values for the X chromosome. See Supplementary Information section 2.

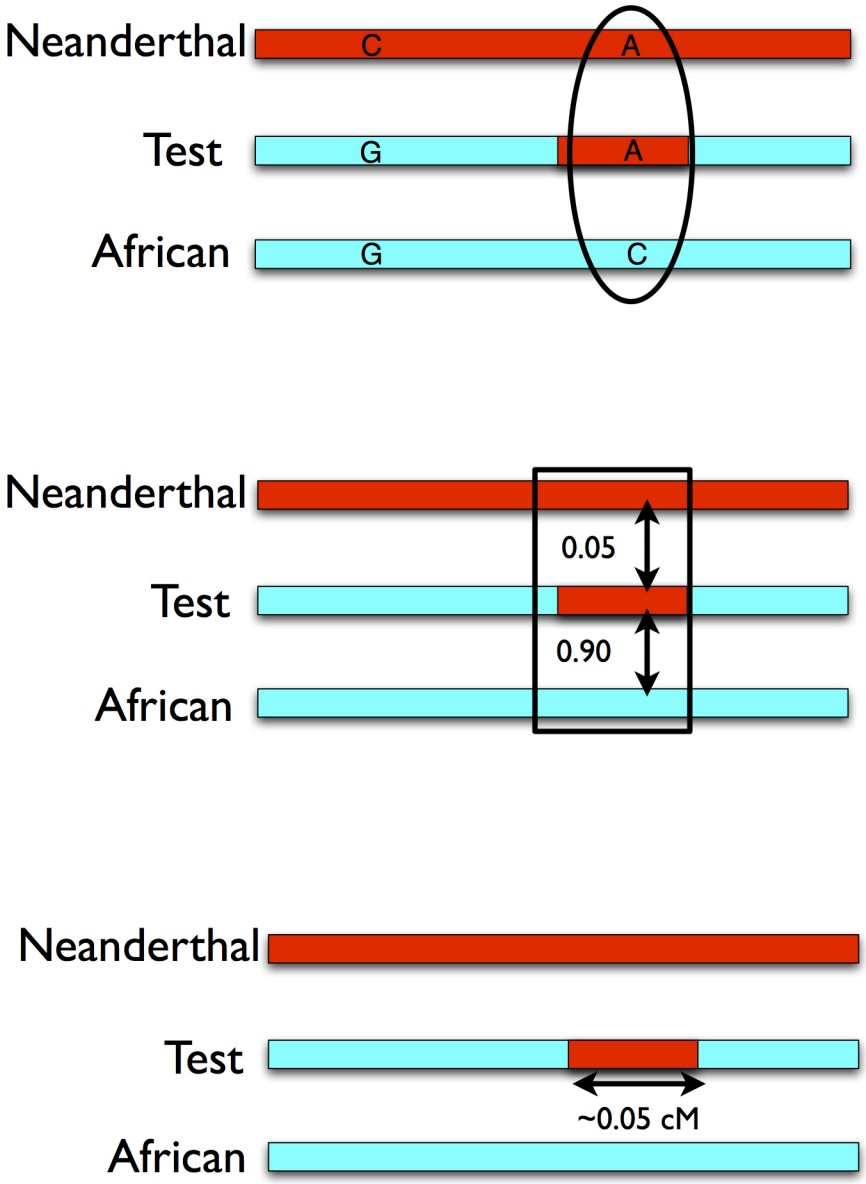
Unbiased estimate of the proportion of Neanderthal ancestry. To estimate the proportion of Neanderthal ancestry in an unbiased way, we divided the genome into quintiles of B, and estimated the proportion of Neanderthal ancestry using a statistic first published in ref. 22. This statistic measures how much closer Denisova is to a non-African individual than to an African individual, divided by the same quantity replacing the non-African individual with Neanderthal. We report the estimated proportion of Neanderthal ancestry in each quintile divided by the genome-wide mean and obtain standard errors using a block jackknife with 100 blocks.

We analysed data from 27 deeply sequenced genomes: 25 present-day humans and the high-coverage Neanderthal and Denisova[17] genomes. For each, we required that sites passed the more stringent set of the two filters described in ref. 2, had a genotype quality of ≥45, and had an ancestral allele that could be determined based on comparison to chimpanzee and at least one of gorilla or orangutan. We computed a *Z*-score for the difference in the ancestry across the bin of highest B statistic versus the rest and used a Bonferroni correction for ten hypotheses (5 hypotheses based on which set of bins we merge and a two-sided test in each). In our main analysis, we analysed both transitions and transversions and pooled genomes for all non-African-individual samples. To establish robustness, we also analysed subsets of the data, repeating the analysis restricting to transversion SNPs, and then further restricting to transversion SNPs only in European and only in eastern-non-African individuals. See Supplementary Information section 9 for details.

Tissue-specific expression. We defined tissue-specific expression levels using the Illumina BodyMap 2.0 RNA-seq data[25], which contains expression data from 16 human tissues. We identified genes that are expressed in a tissue-specific manner using the DESeq package[39] and used a *P*-value cutoff of 0.05. We tested enrichment of tissue-specific genes in regions of high or low Neanderthal ancestry. A concern when testing for enrichment is that clustering of similarly expressed genes coupled with the large size of regions of low Neanderthal ancestry might lead to spurious signals of enrichment. Hence, we devised a permutation test that randomly rotates the annotations of genes (treating each chromosome as a circle) while maintaining the correlation within genes and within Neanderthal ancestry as well as between Neanderthal ancestry and genes. We tested enrichment on the whole genome, on the autosomes alone, and on the X chromosome alone. We generated 1,000 random rotations for each test except for the X chromosome for which we generated all possible rotations. We computed the fraction of permutations for which the *P* value of Fischer's exact test is at least as low as the observed *P* value (see Supplementary Information section 6).

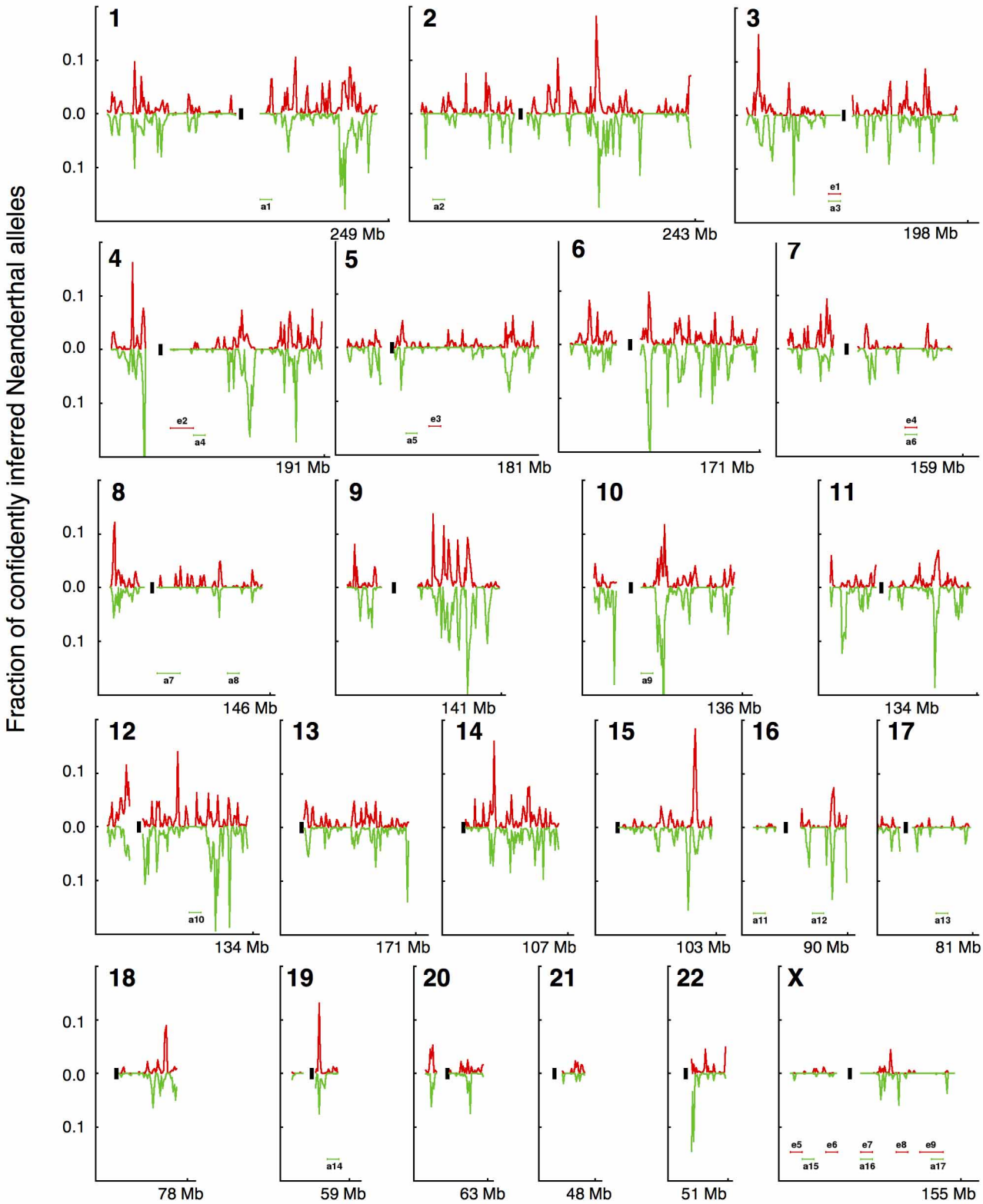
29. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310,** 321–324 (2005).
30. Sutton, C. & McCallum, A. in *Introduction to Statistical Relational Learning* (eds Getoor, L. & Taskar, B.) Ch. 4, 93–128 (MIT Press, 2007).
31. Byrd, R. H., Nocedal, J. & Schnabel, R. B. Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming* **63,** 129–156 (1994).
32. Gravel, S. Population genetics models of local ancestry. *Genetics* **191,** 607–619 (2012).
33. Pruitt, K. D. *et al.* The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19,** 1316–1323 (2009).
34. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
35. Prüfer, K. *et al.* FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* **8,** 41 (2007).
36. Percival, D. B. & Walden, A. T. *Wavelet Methods for Time Series Analysis*. (Cambridge Univ. Press, 2005).
37. Kunsch, H. R. The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17,** 1217–1241 (1989).
38. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338 (2002).
39. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

©2014 Macmillan Publishers Limited. All rights reserved



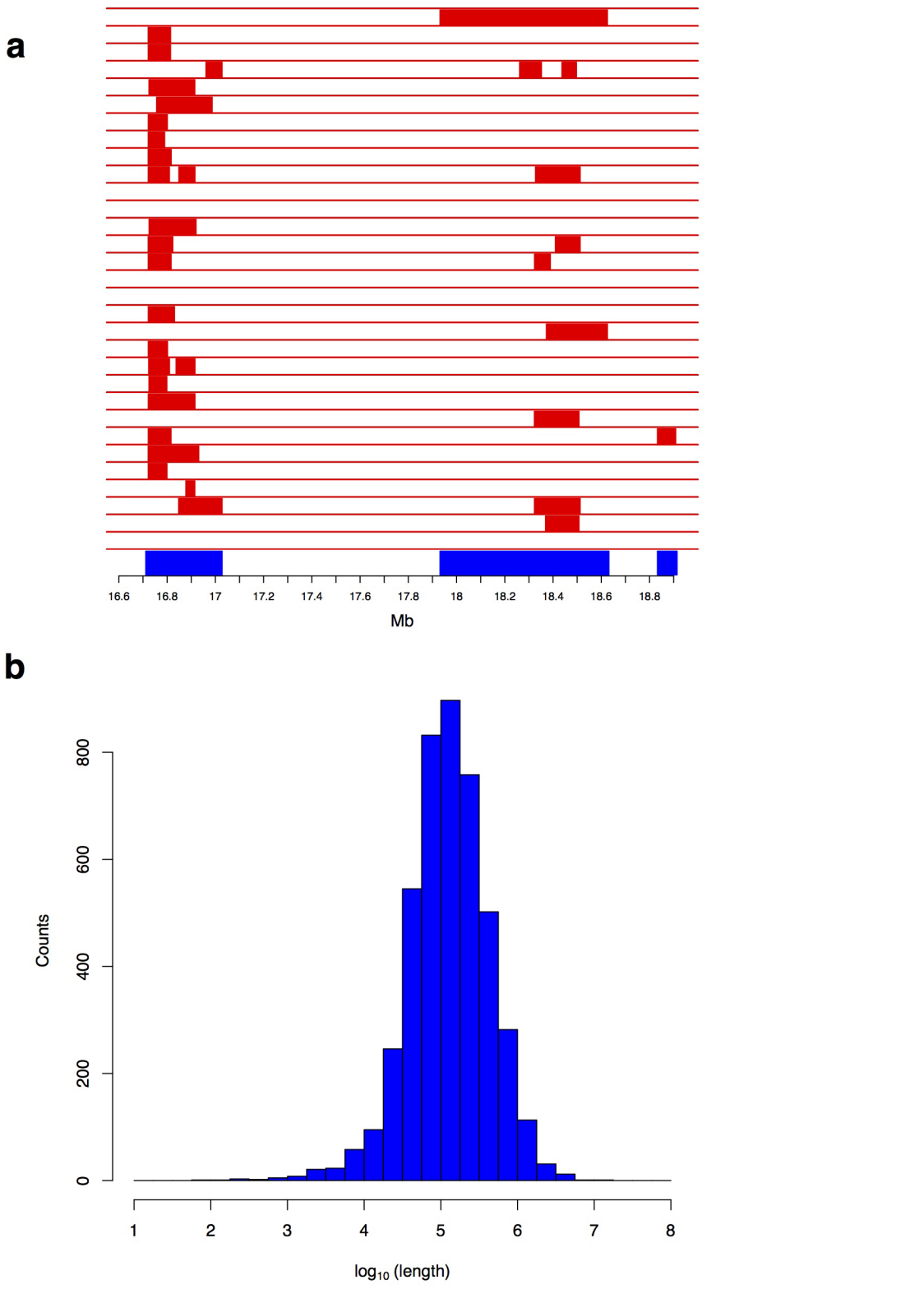
Extended Data Figure 1 | Three features used in the Conditional Random Field for predicting Neanderthal ancestry. Top (feature 1), patterns of variation at a single SNP. Sites at which a panel of sub-Saharan-African individuals carry the ancestral allele and in which the sequenced Neanderthal and the test haplotype carry the derived allele are likely to be derived from Neanderthal gene flow. Middle (feature 2), haplotype divergence patterns. Genomic segments in which the divergence of the test haplotype to the sequenced Neanderthal is low, whereas the divergence to a panel of sub-Saharan-African individuals is high, are likely to be introgressed. Bottom (feature 3), we searched for segments that have a length consistent with what is expected from Neanderthal-to-modern-human gene flow approximately 2,000 generations ago, corresponding to a size of about 0.05 cM = (100 cM per Morgan)/(2,000 generations).

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 2 | Map of Neanderthal ancestry in 1000 Genomes European and east-Asian populations. For each chromosome, we plot the fraction of alleles confidently inferred to be of Neanderthal origin (probability >90%) in non-overlapping 1-Mb windows in Europeans (red) and in east Asians (green). Black bars denote the coordinates of the centromeres. We plot traces in non-overlapping 10-Mb windows that pass filters. We label 10-Mb-scale windows that are deficient in Neanderthal ancestry (e1–e9 (e, European), a1–a17 (a, Asian)) (see Supplementary Information section 8 for details).

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 3 | Tiling path from confidently inferred Neanderthal haplotypes. **a**, Example tiling path at the *BNC2* locus on chromosome 9 in European individuals. Red, confidently inferred Neanderthal haplotypes in a subset of these individuals; blue, resulting tiling path. We identified Neanderthal haplotypes by scanning for runs of consecutive SNPs along a haplotype with a marginal probability >90% and requiring the haplotypes to be at least 0.02 cM long. **b**, Distribution of contig lengths obtained by constructing a tiling path across confidently inferred Neanderthal haplotypes. On merging Neanderthal haplotypes in each of the 1000 Genomes European and east-Asian populations, we reconstructed 4,437 Neanderthal contigs with median length 129 kb.

©2014 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | Gene categories enriched or depleted in Neanderthal ancestry

Biological pathway (GO categorization)	Neandertal ancestry	Europe FWER	East Asian FWER
nucleic acid binding (molecular_function, GO:0003676)	Depleted	0.018	0.032
RNA processing (biological_process, GO:0006396)	Depleted	0.004	0.049
ribonucleoprotein complex (cellular_component, GO:0030529)	Depleted	<0.001	0.027
organelle part (cellular_component, GO:0044422)	Depleted	<0.001	0.037
intracellular organelle part (cellular_component, GO:0044446)	Depleted	<0.001	0.025
mRNA metabolic process (biological_process, GO:0016071)	Depleted	<0.001	0.014
nuclear lumen (cellular_component, GO:0031981)	Depleted	0.039	0.017
nuclear part (cellular_component, GO:0044428)	Depleted	0.005	0.022
keratin filament (cellular_component, GO:0045095)	Enriched	<0.001	<0.001

Enrichment of Gene Ontology categories in genes with depleted or elevated Neanderthal ancestry was assessed using the hypergeometric test implemented in the FUNC package. We report family-wise error rate (FWER) *P* values associated with each GO category (*P* values corrected for the testing of multiple categories).

©2014 Macmillan Publishers Limited. All rights reserved

Extended Data Table 2 | Neanderthal-derived alleles that have been associated with phenotypes in genome-wide association studies

rs id	Coordinates	Derived allele	Derived allele frequency (%)		Phenotype
			Europeans	East Asians	
rs12531711	7:128,617,466	G	10.03	0.17	Systemic lupus erythematosus, Primary biliary cirrhosis
rs3025343	9:136,478,355	A	8.44	0.00	Smoking behavior
rs7076156	10:64,415,184	A	26.52	8.74	Crohn's disease
rs12571093	10:70,019,371	A	16.35	14.86	Optic disc size
rs1834481	11:112,023,827	G	21.50	0.35	Interleukin-18 levels
rs11175593	12:40,601,940	T	1.98	3.32	Crohn's disease
rs75493593	17:6,945,087	T	1.85	12.06	
rs75418188	17:6,945,483	T	1.85	11.54	Type-2 Diabetes
rs117767867	17:6,946,330	T	1.85	11.54	

We identified alleles that are likely to have been introduced by Neanderthal gene flow (Supplementary Information section 7) and intersected these alleles with SNPs that have been shown to be associated with phenotypes (from the NHGRI GWAS catalogue[19] as well from a recent GWAS for type 2 diabetes[20]).

©2014 Macmillan Publishers Limited. All rights reserved

Extended Data Table 3 | Recall of the CRF as a function of the effective population size

Effective population size	Recall
2500	0.552 ± 0.009
5000	0.506 ± 0.009
7500	0.430 ± 0.006
10000	0.384 ± 0.006

Recall is computed at a precision of 90%. Standard errors of the recall are estimated by a block jackknife with 100 blocks.

©2014 Macmillan Publishers Limited. All rights reserved

Extended Data Table 4 | Unbiased estimate of the proportion of Neanderthal ancestry as a function of the B statistic

	11 Non-Africans (transitions + transversions)		**11 Non-Africans** (transversions only)		**4 Europeans** (transversions only)		**7 Eastern** (transversions only)	
	Est.	**Err.**	**Est.**	**Err.**	**Est.**	**Err.**	**Est.**	**Err.**
Quintile 1: B=0-0.63	0.641	0.304	0.672	0.316	0.472	0.397	0.778	0.317
Quintile 2: B=0.63-0.80	0.825	0.209	0.779	0.234	0.849	0.290	0.750	0.236
Quintile 3: B=0.80-0.88	0.578	0.248	0.745	0.298	0.987	0.349	0.647	0.297
Quintile 4: B=0.88-0.94	0.684	0.184	0.676	0.208	0.446	0.256	0.771	0.221
Quintile 5: B=0.94-1.00	1.537	0.152	1.445	0.164	1.502	0.185	1.419	0.177
B≥0.94 vs. B ≤ 0.94	Z=3.82		Z=3.02		Z=3.12		Z=2.58	
Correct for 10 hypotheses	P=0.00066		P=0.013		P=0.0090		P=0.049	

Estimates of the proportion of Neanderthal ancestry in quintiles of B statistics divided by the genome-wide proportion. We find a significant excess of Neanderthal ancestry in the quintile with the highest B statistic relative to the remaining four quintiles (significant after correcting for ten hypotheses).

©2014 Macmillan Publishers Limited. All rights reserved

Extended Data Table 5 | Recall of the CRF on the X chromosome versus the autosomes

	Recall
Autosomes	0.384 ± 0.006
X	0.495 ± 0.009

Simulations are carried out using parameters tailored to be appropriate to each of these compartments of the genome. Recall is computed at a precision of 90%. Standard errors of the recall are estimated by a block jackknife with 100 blocks.

©2014 Macmillan Publishers Limited. All rights reserved