ARTICLE

doi:10.1038/nature18964

The Simons Genome Diversity Project: 300 genomes from 142 diverse populations

A list of authors and affiliations appears at the end of the paper.

Here we report the Simons Genome Diversity Project data set: high quality genomes from 300 individuals from 142 diverse populations. These genomes include at least 5.8 million base pairs that are not present in the human reference genome. Our analysis reveals key features of the landscape of human genome variation, including that the rate of accumulation of mutations has accelerated by about 5% in non–Africans compared to Africans since divergence. We show that the ancestors of some pairs of present–day human populations were substantially separated by 100,000 years ago, well before the archaeologically attested onset of behavioural modernity. We also demonstrate that indigenous Australians, New Guineans and Andamanese do not derive substantial ancestry from an early dispersal of modern humans; instead, their modern human ancestry is consistent with coming from the same source as that of other non–Africans.

To obtain a complete picture of human diversity, it is necessary to sequence the genomes of many individuals from diverse locations. To date, the largest whole-genome sequencing survey, the 1000 Genomes Project, analysed 26 populations of European, East Asian, South Asian, American, and sub-Saharan African ancestry[1]. However, this and most other sequencing studies have focused on demographically large populations. Such studies tend to ignore smaller populations that are also important for understanding human diversity. In addition, many of these studies have sequenced genomes to only 4–6-fold coverage. Here, we report the Simons Genome Diversity Project (SGDP): deep genome sequences of 300 individuals from 142 populations chosen to span much of human genetic, linguistic, and cultural variation (Supplementary Data Table 1).

Data set and catalogue of novel variants

We sequenced the samples to an average coverage of 43-fold (range 34–83-fold) at Illumina Ltd; almost all samples (278) were prepared using the same PCR-free library preparation (https://support.illumina.com/content/dam/illumina-marketing/documents/services/FastTrackServices_Methods_Tech_Note.pdf). We aligned reads to the human reference genome hs37d5/hg19 using BWA-MEM (BWA-0.7.12)[2] (Supplementary Information section 1). We genotyped each sample separately using the Genome Analysis Toolkit (GATK)[3], with a modification to eliminate bias towards genotypes matching the reference (Supplementary Information section 1). We developed a filtering procedure that generates a sample-specific mask. At 'filter level 1' which we recommend for most analyses, we retain an average of 2.13 Gb of sequence per sample and identify 34.4 million single nucleotide polymorphisms (SNPs) and 2.1 million insertion/deletion polymorphisms (indels) (Supplementary Information section 2). We have made the GATK-processed data available in a file small enough to download by FTP, along with software to analyse these data (Supplementary Information section 3). The SGDP data set highlights the incompleteness of current catalogues of human variation, with the fraction of heterozygous positions not discovered by the 1000 Genomes Project being 11% in the KhoeSan and 5% in New Guineans and Australians (Extended Data Fig. 1; Supplementary Data Table 1). We used FermiKit[4] to map short reads against each other, store the assemblies in a compressed form that retains all the information required for polymorphism discovery and analysis, and identified SNPs by comparing against the human reference. We find that FermiKit has comparable sensitivity and specificity to GATK for SNP discovery and genotyping, and is more accurate for indels (Supplementary Information section 4). FermiKit also identified 5.8 Mb of contigs that are present in the SGDP but absent in the human reference genome presumably because they are deleted there; these contigs which we have made publicly available can be used as 'decoys' to improve read mapping (Supplementary Information section 5). Finally, we called copy number variants[5] and used lobSTR[6,7] to genotype 1.6 million short tandem repeats (STRs) (Supplementary Information section 6). The high quality of the STR genotypes ($r^2 = 0.92$ to capillary sequencing calls) is evident from their accurate reconstruction of population relationships, even for difficult-to-genotype mononucleotide repeats (Extended Data Fig. 2).

The structure of human genetic diversity

To obtain an overview of population relationships, we carried out ADMIXTURE[8] (Extended Data Fig. 3) and principal component analysis[9] (Extended Data Fig. 4a). We also built neighbour-joining trees based on pairwise divergence per nucleotide (Fig. 1a) and F_{ST} (Extended Data Fig. 4b) whose topologies are consistent with previous findings that the deepest splits among human populations are among Africans. We computed heterozygosity—the proportion of diallelic genotypes per base pair—and recapitulate previous findings that the highest genetic diversity is found in sub-Saharan Africa and that there is a much lower ratio of X-to-autosome diversity in non-Africans than in Africans (Fig. 1b)[10]. A surprise is that African 'pygmy' hunter-gatherers have reduced X-to-autosome diversity ratios relative to all other sub-Saharan Africans. This pattern is just as strong even after we remove the third of chromosome X known to be subject to the strongest natural selection, suggesting that the finding is driven by demographic history rather than by natural selection (Supplementary Information section 7). It has been suggested that the reduced X-to-autosome heterozygosity ratio in non-Africans is due to ongoing male-driven admixture[10,11]. Male non-pygmy admixture into pygmies is well-documented[12,13], so this process could explain these findings.

Comparisons of ancient to present-day human genomes have shown that all non-Africans today possess Neanderthal ancestry[14] with

more in eastern non-Africans[15,16], and that Australo-Melanesians, and to a lesser extent other eastern non-Africans, possess Denisovan ancestry[17–19]. However, these studies only analysed genomes from a handful of populations. We computed statistics informative about Neanderthal and Denisovan ancestry and provide a fine-scale view of these ancestry distributions worldwide (Fig. 1c, d; Supplementary Data Table 1; Supplementary Information section 8). We do not detect any population with a higher proportion of Neanderthal ancestry than is present in East Asians. However, we do find suggestive evidence of an excess of Denisovan ancestry in some South Asians compared to other Eurasians. This signal may not have been detected before because earlier surveys of archaic introgression largely excluded South Asians (Fig. 1d; Supplementary Data Table 1).

The time course of human population separation

We studied demographic history by leveraging the fact that variation across the genome in divergent sites per base pair can be used to reconstruct population size changes and separations. We used the pairwise sequential Markovian coalescent (PSMC)[20] to reconstruct population size changes, and the multiple sequentially Markovian coalescent (MSMC)[21] to study the time course of population separations. We infer that the population ancestral to all present day humans began to develop

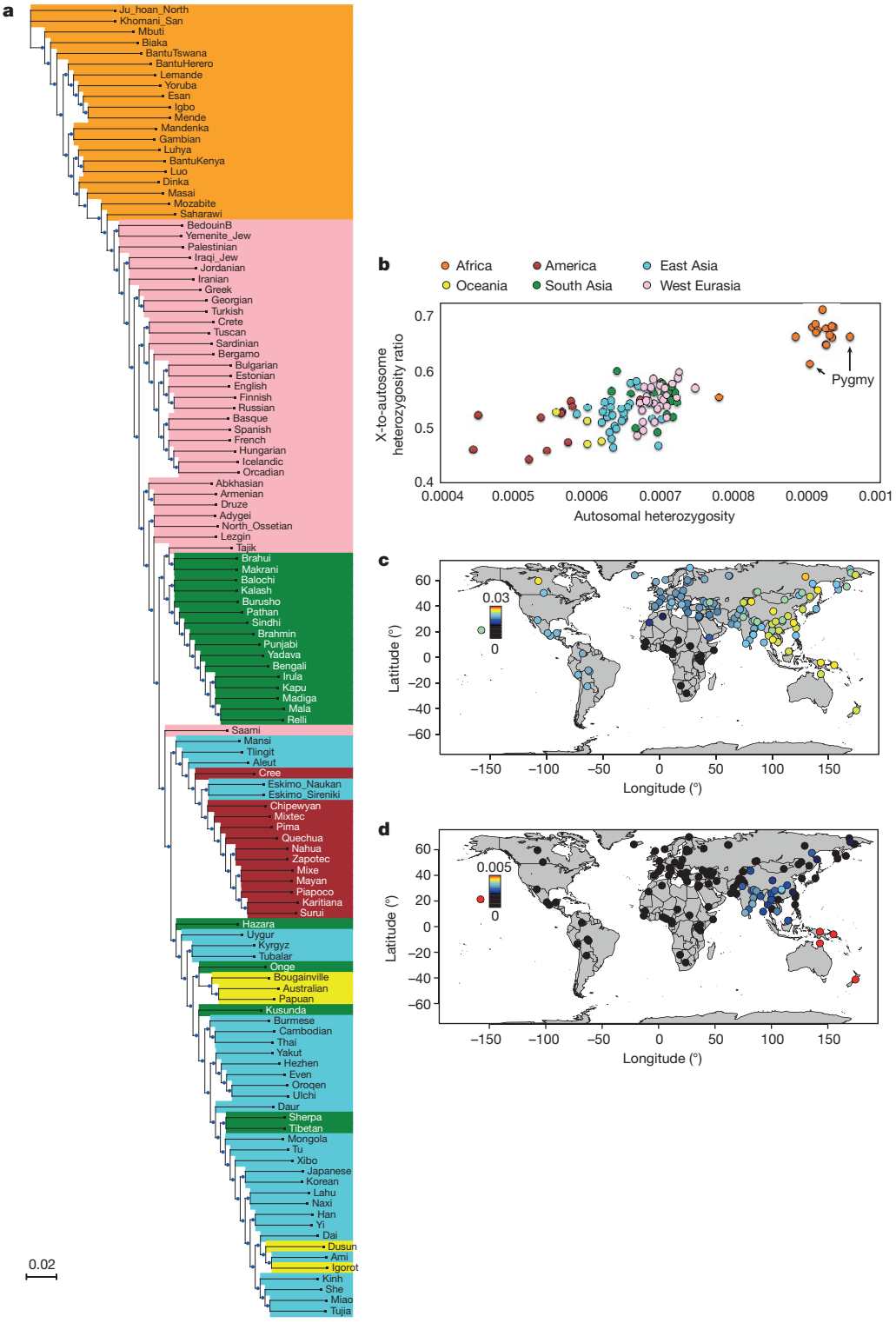


Figure 1 | Genetic variation in the SGDP. a, Neighbour-joining tree of relationships based on pairwise divergence. **b**, Plot of autosomal heterozygosity against the X-to-autosome heterozygosity ratio, showing the reduction in this ratio in non-Africans and pygmies. **c**, Estimate of Neanderthal ancestry with a heat map scale of 0–3%. **d**, Estimate of Denisovan ancestry with a heat map scale of 0–0.5% to bring out subtle differences in mainland Eurasia (Oceanian groups with as much as 5% Denisovan ancestry are saturated in bright red).

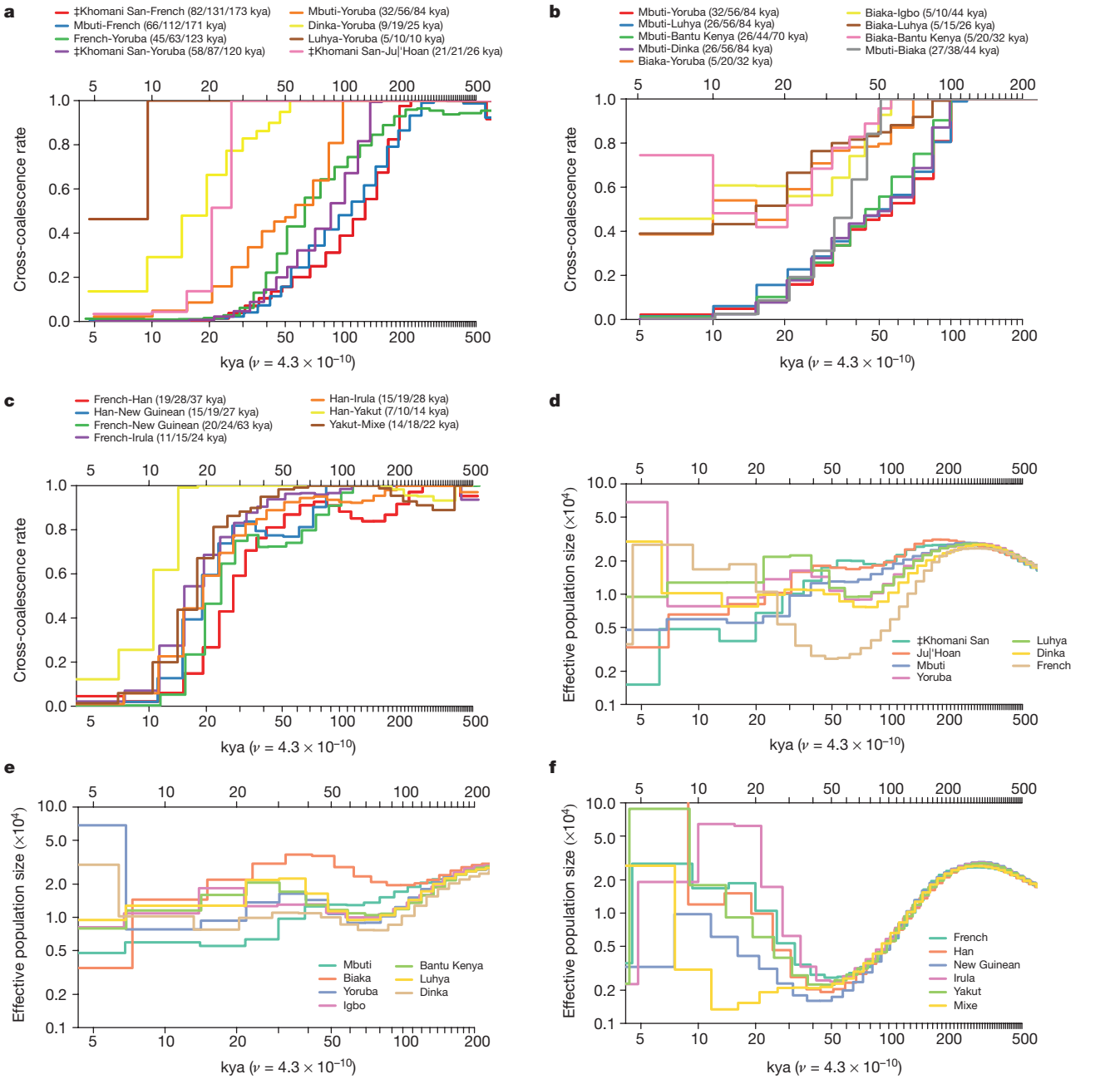


Figure 2 | Cross-coalescence rates and effective population sizes for selected population pairs. a–c, Cross-coalescence rates as a function of time in thousands of years ago (kya) estimated using MSMC, with four haplotypes per pair. In each subfigure legend, we give the point estimate of the date at which 25%, 50% and 75% of lineages in the pair of populations have coalesced into a common ancestral population. We generated these plots using data phased with the 1000 Genomes reference panel (method PS1 described in Supplementary Information section 9), but only show pairs of populations for which the cross-coalescence rates are relatively insensitive to the phasing approach. **a**, Selected African cross-coalescence rates. **b**, Central African rainforest hunter–gatherer cross-coalescence rates. **c**, Ancient non-African cross coalescence rates. **d–f**, Effective population sizes inferred using PSMC, using one diploid genome per population, for the same populations that we used in **a–c**.

substructure at least 200 thousand years ago (kya), which is most apparent when comparing the ancestors of some present-day African hunter–gatherers (southern African KhoeSan and central African Mbuti pygmies) to other populations (Fig. 2a). However, it is also clear that this substructure developed slowly, as all pairs of present-day populations including African hunter–gatherers share a substantial subset of their ancestors as recently as a hundred thousand years ago[22–25]. Quoting the time at which MSMC infers that more than 50% (25–75%) of lineages for a pair of populations are descended from the same ancestral population, we estimate that non-Africans separated substantially from KhoeSan 131 (82–173) kya and almost as anciently from the Mbuti around 112 (66–171) kya. Within Africa (Fig. 2a, b), we infer that the Yoruba separated substantially from the KhoeSan 87 (58–120) kya; from the Mbuti 56 (32–84) kya; and from the Dinka 19 (9–25) kya. We estimate a relatively rapid 21 (21–26) kya separation of northern and southern KhoeSan[23,26] potentially reflecting isolation since the last glacial maximum; and 38 (27–44) kya separation between western (Biaka) and eastern (Mbuti) pygmies, confirming very old substructure between these two central African hunter–gatherer groups[27]. Outside Africa, the most ancient structure dates to around 50 kya (Fig. 2c) during or shortly after the deepest part of the shared non-African bottleneck 40–60 kya, consistent with the archaeological evidence of the dispersal of modern humans into Eurasia during this period. We are not confident about the estimates of the date of

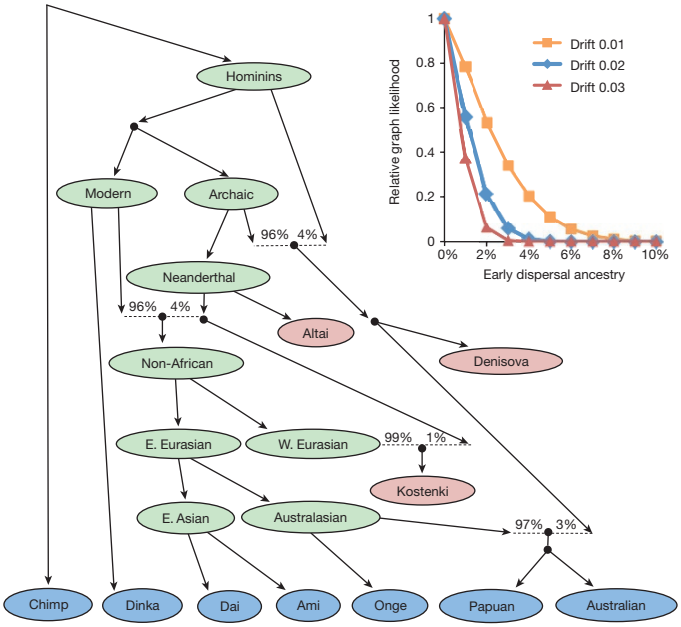


Figure 3 | Present-day populations have negligible ancestry from an early dispersal of modern humans out of Africa. Best-fitting admixture graph model of relationships among Australians, New Guineans, Andamanese and other diverse populations. Present-day populations are shown in blue, ancient samples in red, and select inferred ancestral nodes in green. Dotted lines indicate admixture events, all of which involve archaic humans. All *f*-statistic relationships are accurately fit to within 2.1 standard errors. Inset, results of adding putative early dispersal admixture to the graph model for different assumptions about when the early lineage split off. We specify the split time in terms of the genetic drift above the 'Non-African' node, with 0.01 units of drift representing on the order of ten thousand years. The (approximate) model likelihood is maximized with zero early dispersal ancestry, and no more than a few per cent is consistent with the data.

separation of Australians, New Guineans and Andamanese from other populations because we find that these inferences change depending on the computational method we use for phasing, probably due to these populations not being represented in the 1000 Genomes haploid genome reference panel (Supplementary Information section 9). We caution that the date estimates also do not take into account uncertainty about the true value of the human mutation rate, which could plausibly be 30% higher or lower than the point estimate we use[28].

Early modern human dispersals contributed little to non-Africans

There is intense debate about whether present-day Australians, New Guineans and Asian 'Negrito' populations are descended from the same source population as mainland Eurasians, or whether they also derive some ancestry from an early, independent dispersal of modern humans into Asia[29–31]. To explore this scenario rigorously, we fit an admixture graph[32]—a phylogenetic tree incorporating mixture events—to the allele frequency correlations among Neanderthals, Denisovans, Upper Paleolithic Europeans, East Asians, New Guineans, Australians, and Andamanese. We obtain a good fit to the data if we include known Neanderthal and Denisovan introgression and model all modern human ancestry in New Guineans, Australians and Andamanese as part of an eastern clade together with mainland East Asians (Supplementary Information section 11; Fig. 3). Furthermore, when we manually introduce a deeply diverging modern human lineage contributing ancestry to Australians, New Guineans, and Andamanese (or when we repeat the analysis in a model without Andamanese), no position or proportion of the deep lineage improves the fit. If this putative source population branched off the main lineage leading to non-Africans more than about 10–20 thousand years before the separation of European and East Asian ancestors, we obtain an upper bound of a few per cent for the possible contribution to Australians and New Guineans (Fig. 3 inset; Supplementary Information section 11). These results are at odds with an inference of substantial early dispersal ancestry in a previous analysis of an Australian genome[31]; however, that study used a less complete model that, notably, did not include the known Denisovan admixture into Australo-Melanesians[17]. The findings for Australians are also unlikely to be due to some unusual feature of the individuals we sequenced, as when we compared three different groups of Australian samples for which there is published genome-wide data, we found them all to be consistent with descending from a common homogeneous population since separation from New Guineans (Supplementary Information section 10). These results are not in conflict with skeletal and archaeological evidence of an early modern human presence outside of Africa[29,33], as early migrations could have occurred but not contributed substantially to present-day populations. The possibility of populations that once flourished but did not contribute substantially to living groups is especially plausible now that ancient DNA from the ~45 kya Ust'-Ishim[28] and the ~40 kya Oase 1 individuals[34] has documented their existence.

Accelerated mutation accumulation in non-Africans

The SGDP data provide an opportunity to compare the rates at which mutations have accumulated across populations. We restricted our analyses to samples for which our genotypes are likely to be most reliable (this included restricting to samples which were all processed in the same way), and we used the highest level of filtering ('level 9') (Supplementary Information section 7). We pooled samples by region to increase power, and for all pairs of regions, computed the expected number of positions where, if we picked a random chromosome from both, region A would mismatch chimpanzee and region B would be identical to chimpanzee (or vice versa). If the rate of accumulation of mutation has been the same since the two populations diverged, these numbers are expected to be equal[35]. However, when we compute the ratio of mutations on one lineage or the other since separation, we find a subtle (average of 0.5%) but significant excess of mutations in non-Africans relative to sub-Saharan Africans ($3.3 < |Z| < 9.4$ standard errors from zero; Extended Data Table 1). Because any difference must reflect events since non-African/African population divergence, which is a less than a tenth of average genetic divergence (Fig. 2a), this implies a greater difference in mutation accumulation rates since population divergence (~5%). We were concerned that these results might be biased by the fact that the human genome reference sequence is more closely related to non-Africans than to Africans, or by higher levels of heterozygosity in Africans, as both of these issues could make detection of divergent sites in Africans more difficult. However, we replicated the findings after remapping to chimpanzee, which is equally distant to all present populations, and after restricting analyses to the X chromosome in males (as males only have a single X chromosome, this procedure avoids bias due to different error rates in detecting heterozygous genotypes in populations with different rates of heterozygosity) (Extended Data Fig. 5). These observations are most likely to be explained by acceleration in the rate of mutation accumulation in non-Africans, since the same signal appears in comparisons to sub-Saharan Africans related in different ways to non-Africans (Extended Data Table 1). It is known that the rate of CCT > CTT mutations differs across human populations. However, this particular mutation class was found to be enriched relative to Africans in Europeans but not in East Asians, and thus cannot explain our signal[36]. One of several possible explanations for these findings is a decrease in the generation interval in non-Africans compared to Africans since separation[37].

No species-wide sweeps in modern humans

Finally, we used the SGDP data set to address the hypothesis that the widespread appearance of modern human behaviour in the

archaeological record after ~50 kya was driven by one or a few changes in neurological genes that swept through the population shortly before this time[38]. We first applied the 3P-CLR method[39] to search for locations in the genome with low allele frequency differentiation between KhoeSan and other modern humans, combined with high differentiation between modern and archaic (Neanderthal and Denisovan) humans, as might be expected from a selective sweep in the ancestors of all modern humans (Supplementary Information section 12) (Extended Data Fig. 6). We found no strong outlier signals, although a caveat is that the scan has limited power and we could not apply it to filtered sections of the genome. We also applied the PSMC method[20] to estimate the average time since the most recent common ancestor (TMRCA) of individuals' two chromosomes in the genomic regions within the largest 3P-CLR peaks (38 peaks corresponding the top 0.1%). In none of the regions did we infer that the great majority of all pairs of modern humans share a common ancestor <100 kya, as would be expected for a sweep just before ~50 kya years ago (Supplementary Data Table 2).

As a second approach to scanning for species-wide selective sweeps, we applied the PSMC to infer TMRCA for SGDP samples across the entire genome. This analysis found no regions where the great majority of pairs of human genomes are inferred to share a common ancestor <100 kya (the largest fraction seen anywhere in the genome is 68%; Extended Data Fig. 7).

Taken together, these results do not rule out the possibility that genetic changes contributed in a meaningful way to changes in human behaviour after 50 kya; for example, changing selection can produce shifts in the frequencies of pre-existing mutations to bring a population to a new and advantageous set-point for a phenotype as occurred in the case of height differences between northern and southern Europeans[40]. For polygenic selection, however, genetics is not a creative force, and instead responds to selection pressures imposed by novel environmental conditions or lifestyles. Thus, our results provide evidence against a model in which one or a few mutations were responsible for the rapid developments in human behaviour in the last 50,000 years. Instead, changes in lifestyles due to cultural innovation or exposure to new environments are likely to have been driving forces behind the rapid transformations in human behaviour in the last 50,000 years[41,42].

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 18 September 2015; accepted 23 June 2016.
Published online 21 September 2016.

1. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
2. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
3. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
4. Li, H. FermiKit: assembly-based variant calling for Illumina resequencing data. Preprint at http://arxiv.org/abs/1504.06574 (2015).
5. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349,** aab3761 (2015).
6. Gymrek, M. & Erlich, Y. Profiling short tandem repeats from short reads. *Methods Mol. Biol.* **1038,** 113–135 (2013).
7. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.* **22,** 1154–1162 (2012).
8. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
9. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
10. Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat. Genet.* **41,** 66–70 (2009).
11. Keinan, A. & Reich, D. Can a sex-biased human demography account for the reduced effective population size of chromosome X in non-Africans? *Mol. Biol. Evol.* **27,** 2312–2321 (2010).
12. Verdu, P. *et al.* Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. *Mol. Biol. Evol.* **30,** 918–937 (2013).
13. Joiris, D. V. The framework of central African hunter-gatherers and neighbouring societies. *African Study Monographs* **Suppl. 28,** 57–79 (2003).
14. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
15. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
16. Wall, J. D. *et al.* Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics* **194,** 199–209 (2013).
17. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468,** 1053–1060 (2010).
18. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).
19. Skoglund, P. & Jakobsson, M. Archaic human ancestry in East Asia. *Proc. Natl Acad. Sci. USA* **108,** 18301–18306 (2011).
20. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).
21. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46,** 919–925 (2014).
22. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43,** 1031–1034 (2011).
23. Schlebusch, C. M. *et al.* Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* **338,** 374–379 (2012).
24. Veeramah, K. R. *et al.* An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Mol. Biol. Evol.* **29,** 617–630 (2012).
25. Labuda, D., Zietkiewicz, E. & Yotova, V. Archaic lineages in the history of modern humans. *Genetics* **156,** 799–808 (2000).
26. Pickrell, J. K. *et al.* The genetic prehistory of southern Africa. *Nat. Commun.* **3,** 1143 (2012).
27. Patin, E. *et al.* Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet.* **5,** e1000448 (2009).
28. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514,** 445–449 (2014).
29. Groucutt, H. S. *et al.* Rethinking the dispersal of Homo sapiens out of Africa. *Evol. Anthropol.* **24,** 149–164 (2015).
30. Reyes-Centeno, H., Hubbe, M., Hanihara, T., Stringer, C. & Harvati, K. Testing modern human out-of-Africa dispersal models and implications for modern human origins. *J. Hum. Evol.* **87,** 95–106 (2015).
31. Rasmussen, M. *et al.* An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334,** 94–98 (2011).
32. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192,** 1065–1093 (2012).
33. Liu, W. *et al.* The earliest unequivocally modern humans in southern China. *Nature* **526,** 696–699 (2015).
34. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524,** 216–219 (2015).
35. Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat. Genet.* **47,** 126–131 (2015).
36. Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl Acad. Sci. USA* **112,** 3439–3444 (2015).
37. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* **15,** 47–70 (2014).
38. Klein, R. G. & Edgar, B. *The dawn of human culture.* (Wiley, 2002).
39. Racimo, F. Testing for ancient selection using cross-population allele frequency differentiation. *Genetics* **202,** 733–750 (2015).
40. Turchin, M. C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat. Genet.* **44,** 1015–1019 (2012).
41. Mcbrearty, S. & Brooks, A. S. The revolution that wasn't: a new interpretation of the origin of modern human behavior. *J. Hum. Evol.* **39,** 453–563 (2000).
42. Renfrew, C. *Prehistory: the Making of the Human Mind.* (Modern Library, 2009).
43. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12,** 246 (2011).
44. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 7 (2015).
45. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

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

Acknowledgements We thank the volunteers who donated samples. We thank H. Blanche, N. Boivin, H. Cann (deceased), E. Eichler, H. Greely, M. Petraglia, K. Prüfer, A. Rogers, M. Steinrücken, U. Stenzel and P. Sudmant for comments, critiques, discussions, or advice on assembling samples. We thank S. Fan for uploading 21 genomes to the European Genome-phenome archive. The sequencing was funded by the Simons Foundation (SFARI 280376) and the US National Science Foundation (BCS-1032255). I.M. was supported by a Long Term Fellowship grant LT001095/2014 from the Human Frontier Science program. P.S. was supported by the Wenner-Gren foundation and the Swedish Research Council (VR grant 2014-453). T.W. and M.G. were supported by an NIJ grant 2014-DN-BX-K089. Y.E. was supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund and by NIJ grant 2014-DN-BX-K089. D.L. was supported by the Natural Sciences and Engineering

Research Council of Canada. T.K. was supported by ERC Starting Investigator grant FP7 - 261213. R.S. received support from Russian Foundation for Basic Research (#15-04-02543). S.D. received support from the Russian Foundation for Basic Research (#16-34-00599). R.K., E.K. and S.L. were supported by the Russian Foundation for Basic Research (11-04-00725-a). E.B. was supported by the Russian Foundation for Basic Research (16-06-00303). O.B. was supported by the Russian Scientific Fund (14-04-00827) and by the Russian Foundation for Basic Research (16-04-00890). D.M.B., H.S., E.M., R.V. and M.M. were supported by Institutional Research Funding from the Estonian Research Council IUT24-1 and by the European Regional Development Fund (European Union) through the Centre of Excellence in Genomics to Estonian Biocentre and University of Tartu. D.C. was supported by the Spanish MINECO grant CGL-44351-P. L.B.J. and W.S.W. were supported by NIH grant GM59290. S.A.T. was supported by NIH grants 5DP1ES022577 05, 1R01DK104339-01, and 1R01GM113657-01. C.T.-S. and Y.X. were supported by The Wellcome Trust grant 098051. C.M.B. was supported by NSF grants 0924726 and 1153911. K.T. was supported by CSIR Network Project grant (GENESIS: BSC0121). J.P.S. and Y.S.S. were supported in part by an NIH grant R01-GM094402, and a Packard Fellowship for Science and Engineering. G.R., J.K and S.P. were funded by the Max Planck Society. N.P. and D.R. were supported by NIH grant GM100233 and D.R. is a Howard Hughes Medical Institute investigator.

Author Contributions S.M., Y.E., Y.S.S., S.P., J.K., N.P. and D.R. supervised the study. S.N., N.R., C.G., G.P., F.B., G.D., I.G.R., A.R.J., P.D., D.M.B., C.M.B., C.C., T.H., A.M.-E., O.L.P., E.B., O.B., S.K.-Y., H.S., D.T., L.Y., C.T.-S., Y.X., M.S.A., A.R.-L., C.B., A.D.R., C.J., E.B.S., E.M., J.P., R.V., B.M.H., U.H., R.W.M., A.S., G.S., J.T.S.W., R.K., E.K., S.L., G.A., D.C., M.H., T.K., W.K., C.A.W., D.L., M.B., L.B.J., S.A.T., W.S.W., M.M., S.D., R.S., L.S., K.T. and D.R. assembled samples. S.M., H.L., M.L., I.M., M.G., F.R., J.P.S., M.Z., N.C., A.T., P.S., I.L., S.S., Q.F., G.R., Y.S., N.P. and D.R. performed analyses. S.M., H.L., M.L., I.M., M.G., F.R., M.Z., N.P. and D.R. wrote the manuscript with help from all co-authors.

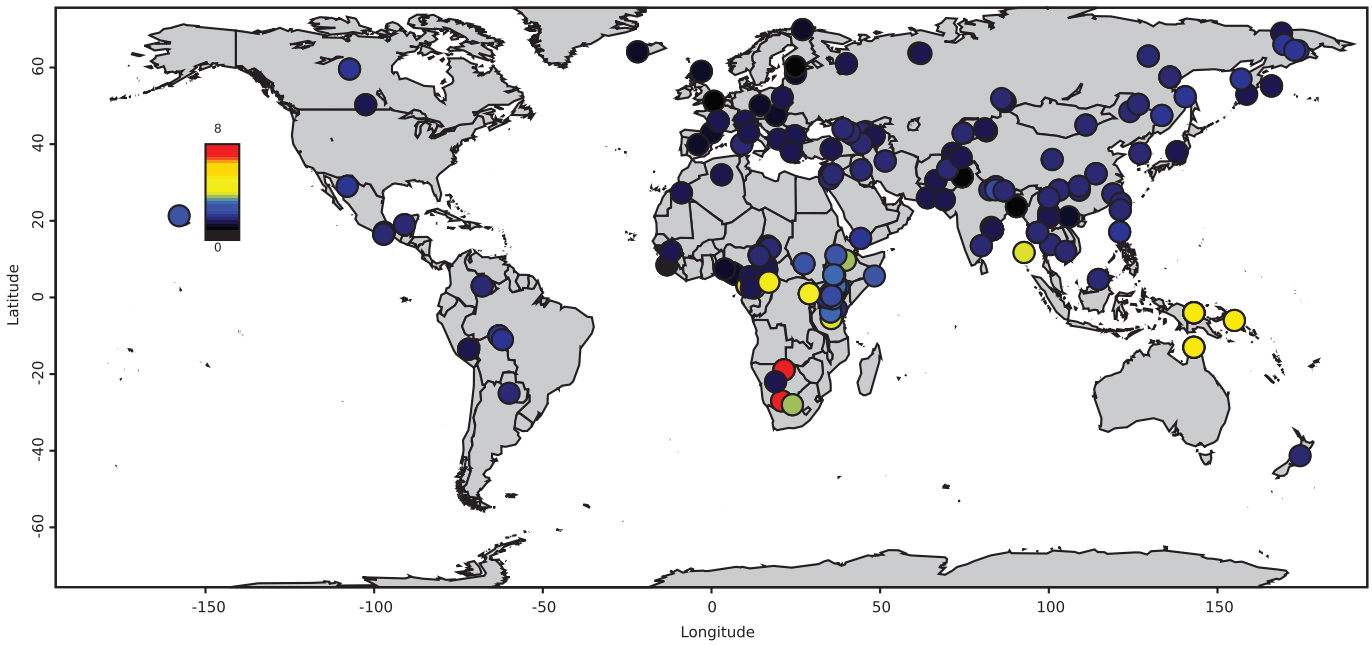
Author Information Raw data for 279 genomes for which the informed consent documentation is consistent with fully public data release are available through the EBI European Nucleotide Archive under accession numbers PRJEB9586 and ERP010710. For the remaining 21 genomes (designated by code 'Y' in the seventh column of Supplementary Data Table 1), data are deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001001959. Data for these 21 genomes can be obtained by submitting to the EGA Data Access Committee a signed letter containing the following text: "(a) I will not distribute the data outside my collaboration; (b) I will not post the data publicly; (c) I will make no attempt to connect the genetic data to personal identifiers for the samples; and (d) I will not use the data for any commercial purposes." Compact versions of the SGDP dataset and software for accessing it are available at (http://genetics.med.harvard.edu/reichlab/Reich_Lab/Datasets.html). The short tandem repeat (STR) genotypes are available through dbVar under accession number nstd128 (http://www.ncbi.nlm.nih.gov/dbvar). Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.M. (shop@genetics.med.harvard.edu) or D.R. (reich@genetics.med.harvard.edu).

Reviewer Information *Nature* thanks P. Bellwood and S. Ramachandran and the other anonymous reviewer(s) for their contribution to the peer review of this work.

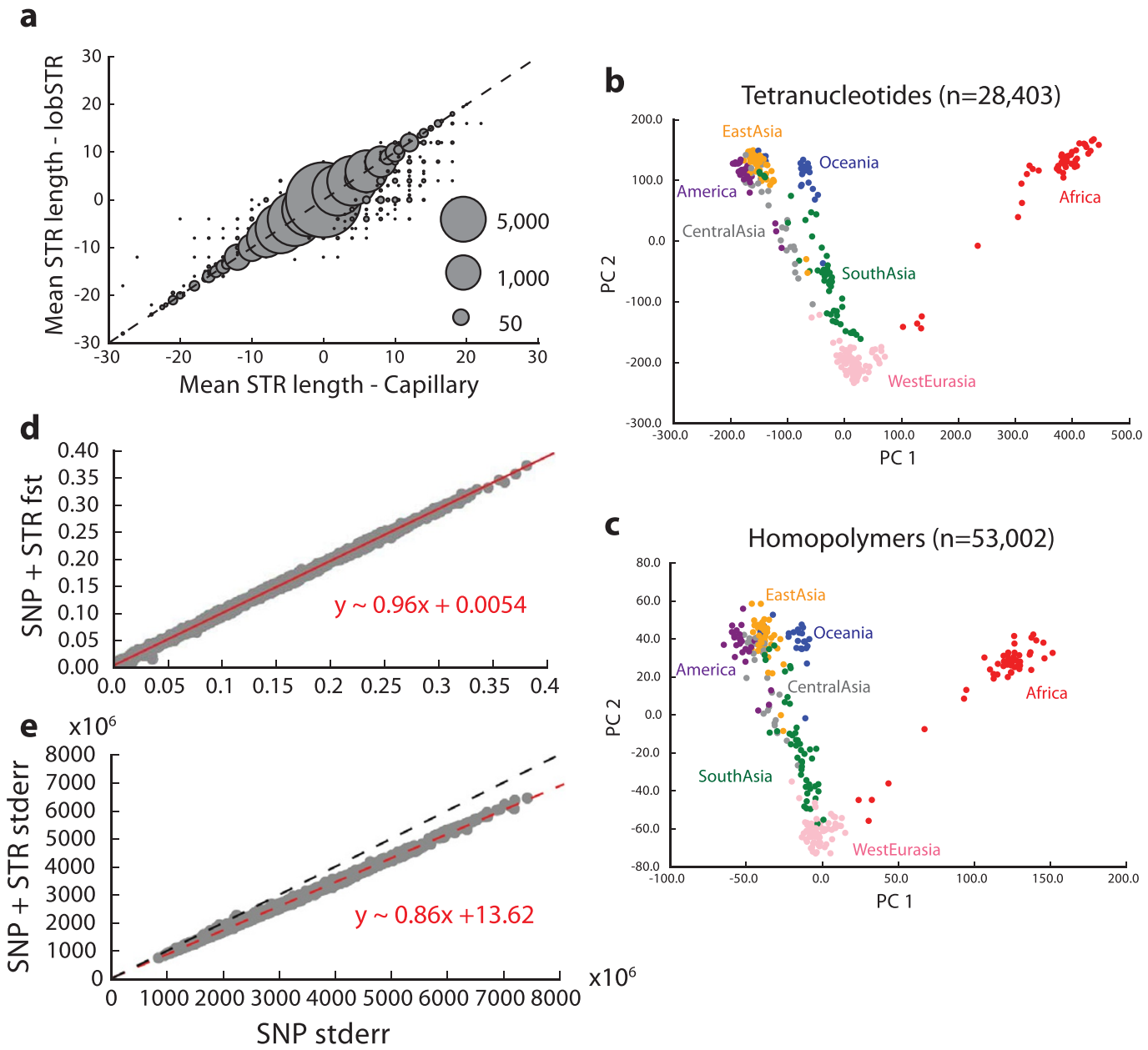
Swapan Mallick[1,2,3], Heng Li[2]*, Mark Lipson[1]*, Iain Mathieson[1]*, Melissa Gymrek[2,4,5,6], Fernando Racimo[7], Mengyao Zhao[1,2,3], Niru Chennagiri[1,2,3], Susanne Nordenfelt[1,2,3], Arti Tandon[1,2], Pontus Skoglund[1,2], Iosif Lazaridis[1,2], Sriram Sankararaman[1,2]†, Qiaomei Fu[1,2,8], Nadin Rohland[1,2], Gabriel Renaud[9], Yaniv Erlich[6,10,11], Thomas Willems[6,12], Carla Gallo[13], Jeffrey P. Spence[14], Yun S. Song[15,16,17], Giovanni Poletti[13], Francois Balloux[18], George van Driem[19], Peter de Knijff[20], Irene Gallego Romero[21,22], Aashish R. Jha[23], Doron M. Behar[24], Claudio M. Bravi[25], Cristian Capelli[26], Tor Hervig[27], Andres Moreno-Estrada[28], Olga L. Posukh[29,30], Elena Balanovska[31], Oleg Balanovsky[31,32,33], Sena Karachanak-Yankova[34], Hovhannes Sahakyan[24,35], Draga Toncheva[34], Levon Yepiskoposyan[35], Chris Tyler-Smith[36], Yali Xue[36], M. Syafiq Abdullah[37], Andres Ruiz-Linares[38], Cynthia M. Beall[39], Anna Di Rienzo[23], Choongwon Jeong[23], Elena B. Starikovskaya[40], Ene Metspalu[24,41], Jüri Parik[24], Richard Villems[24,41,42], Brenna M. Henn[43], Ugur Hodoglugil[44], Robert Mahley[45], Antti Sajantila[46], George Stamatoyannopoulos[47], Joseph T. S. Wee[48], Rita Khusainova[49,50], Elza Khusnutdinova[49,50], Sergey Litvinov[24,49,50], George Ayodo[51], David Comas[52], Michael F. Hammer[53], Toomas Kivisild[24,54], William Klitz[6], Cheryl A. Winkler[55], Damian Labuda[56], Michael Bamshad[57], Lynn B. Jorde[58], Sarah A. Tishkoff[59], W. Scott Watkins[60], Mait Metspalu[24], Stanislav Dryomov[40,61], Rem Sukernik[40,62], Lalji Singh[63]†, Kumarasamy Thangaraj[63], Svante Pääbo[9], Janet Kelso[9], Nick Patterson[2] & David Reich[1,2,3]

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [3]Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA. [5]Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts 02139, USA. [6]New York Genome Center, New York, New York 10013, USA. [7]Department of Integrative Biology, University of California, Berkeley, California 94720-3140, USA. [8]Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, IVPP, CAS, Beijing 100044, China. [9]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany. [10]Department of Computer Science, Columbia University, New York, New York 10027, USA. [11]Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032, USA. [12]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [13]Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima 15102, Perú. [14]Computational Biology Graduate Group, University of California, Berkeley, California 94720, USA. [15]Computer Science Division, University of California, Berkeley, California 94720, USA. [16]Department of Statistics, University of California, Berkeley, California 94720, USA. [17]Department of Mathematics and Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [18]Genetics Institute, University College London, Gower Street, London WC1E 6BT, UK. [19]Institute of Linguistics, University of Bern, Bern CH-3012, Switzerland. [20]Department of Human and Clinical Genetics, Postzone S5-P, Leiden University Medical Center, 2333 ZA Leiden, Netherlands. [21]School of Biological Sciences, Nanyang Technological University, 637551 Singapore. [22]Lee Kong Chian School of Medicine, Nanyang Technological University, 636921 Singapore. [23]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. [24]Estonian Biocentre, Evolutionary Biology group, Tartu 51010, Estonia. [25]Laboratorio de Genética Molecular Poblacional, Instituto Multidisciplinario de Biología Celular (IMBICE), CCT-CONICET La Plata/CIC Buenos Aires/Universidad Nacional de La Plata, La Plata B1906APO, Argentina. [26]Department of Zoology, University of Oxford, Oxford OX1 3PS, UK. [27]Department of Clinical Science, University of Bergen, Bergen 5021, Norway. [28]National Laboratory of Genomics for Biodiversity (LANGEBIO), CINVESTAV, Irapuato, Guanajuato 36821, Mexico. [29]Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences, Novosibirsk 630090, Russia. [30]Novosibirsk State University, Novosibirsk 630090, Russia. [31]Research Centre for Medical Genetics, Moscow 115478, Russia. [32]Vavilov Institute for General Genetics, Moscow 119991, Russia. [33]Moscow Institute for Physics and Technology, Dolgoprudniy 141700, Russia. [34]Department of Medical Genetics, National Human Genome Center, Medical University Sofia, Sofia 1431, Bulgaria. [35]Laboratory of Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences of Armenia, Yerevan 0014, Armenia. [36]The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. [37]RIPAS Hospital, Bandar Seri Begawan, Brunei. [38]Department of Genetics, Evolution and Environment, University College London WC1E 6BT, UK. [39]Department of Anthropology, Case Western Reserve University, Cleveland, Ohio 44106-7125, USA. [40]Laboratory of Human Molecular Genetics, Institute of Molecular and Cellular Biology, Siberian Branch of Russian Academy of Sciences, Novosibirsk 630090, Russia. [41]Department of Evolutionary Biology, University of Tartu, Tartu 51010, Estonia. [42]Estonian Academy of Sciences, Tallinn 10130, Estonia. [43]Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York 11794, USA. [44]NextBio, Illumina, Santa Clara, California 95050, USA. [45]Gladstone Institutes, San Francisco, California 94158, USA. [46]Department of Forensic Medicine, University of Helsinki, Helsinki 00014, Finland. [47]Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, Washington 98195, USA. [48]National Cancer Centre Singapore, 169610 Singapore. [49]Institute of Biochemistry and Genetics, Ufa Research Centre, Russian Academy of Sciences, Ufa 450054, Russia. [50]Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa 450074, Russia. [51]Jaramogi Oginga Odinga University of Science and Technology, Bondo 40601, Kenya. [52]Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona 08003, Spain. [53]ARL Division of Biotechnology, University of Arizona, Tucson, Arizona 85721, USA. [54]Division of Biological Anthropology, University of Cambridge, Fitzwilliam Street, Cambridge CB2 1QH, UK. [55]Basic Research Laboratory, Center for Cancer Research, NCI, Leidos Biomedical Research, Inc., Frederick National Laboratory, Frederick, Maryland 21702, USA. [56]CHU Sainte-Justine, Pediatrics Departement, Université de Montréal, Québec H3T 1C5, Canada. [57]Department of Pediatrics, University of Washington, Seattle, Washington 98119, USA. [58]Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA. [59]Departments of Genetics and Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [60]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, Utah 84112, USA. [61]Department of Paleolithic Archaeology, Institute of Archaeology and Ethnography, Siberian Branch of Russian Academy of Sciences, Novosibirsk 630090, Russia. [62]Altai State University, Barnaul 656000, Russia. [63]CSIR-Centre for Cellular and Molecular Biology, Hyderabad 500 007, India. †Present addresses: Department of Computer Science, University of California at Los Angeles, California 90095, USA and Department of Human Genetics Science, University of California at Los Angeles, California 90095, USA (S.S.); Genome Foundation, Hyderabad 500076, India (L.S).

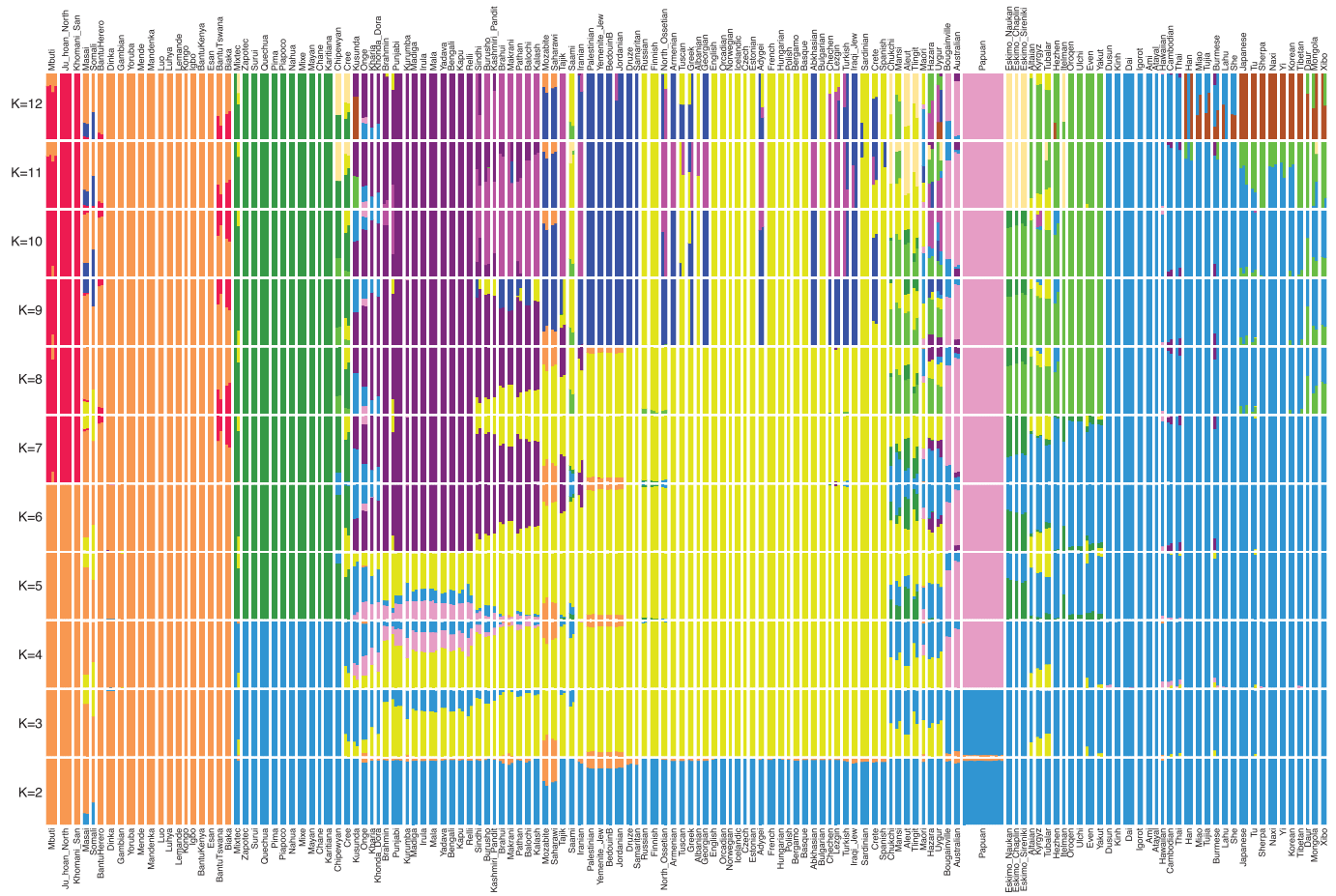
*These authors contributed equally to this work.



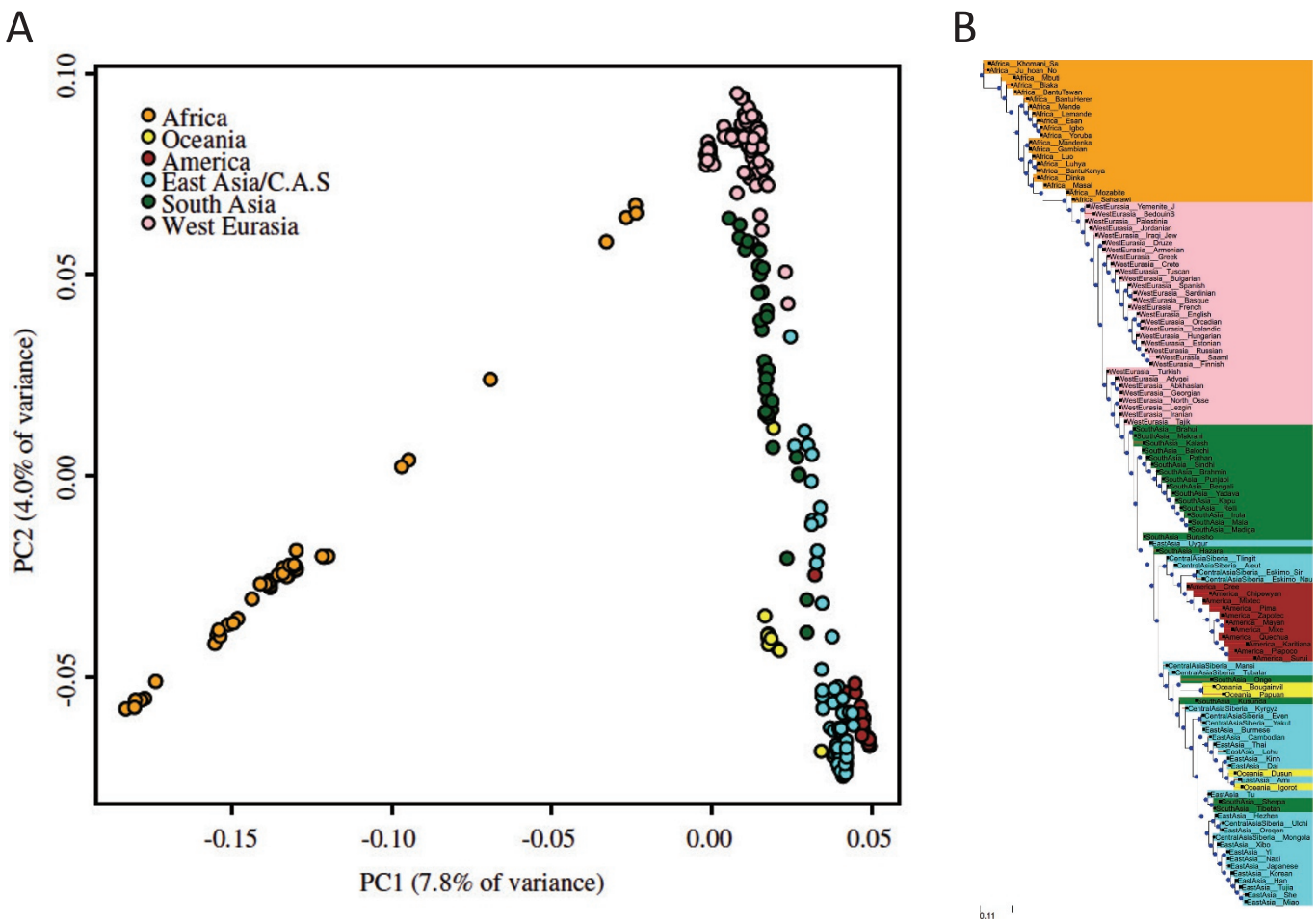
Extended Data Figure 1 | Heat map of fraction of heterozygous sites missed in the 1000 Genomes project. For each sample, we examine all heterozygous sites passing filter level 1, and compute the fraction included as known polymorphisms in the 1000 Genomes project.



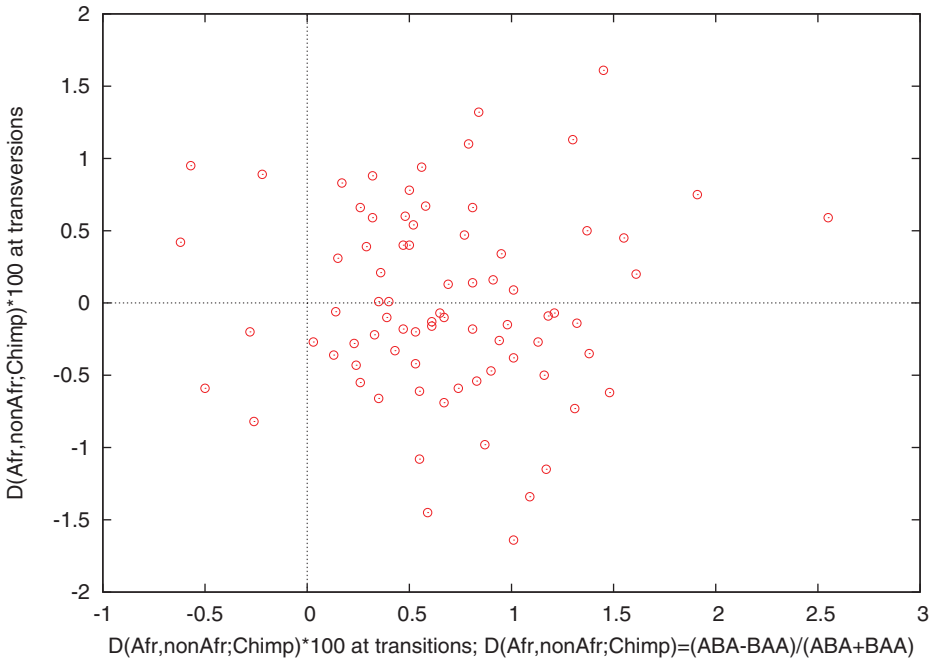
Extended Data Figure 2 | Worldwide variation in human short tandem repeats. a, Mean STR length is reported as the average of the length difference (in base pairs) from the GRCh37 reference for each genotype. Bubble area scales with the number of calls compared at each point. **b**, **c**, The first two principal components after performing principal component analysis on tetranucleotide and homopolymer genotypes, respectively. Colours represent the region of origin of each sample. **d**, Pairwise F_{ST} values between populations computed using only SNPs versus using combined SNP + STR loci. **e**, Block jackknife standard errors for the SNP versus SNP + STR F_{ST} analysis. The red dashed lines give the best-fit line, described by the formula in red. The black dashed line denotes the diagonal.



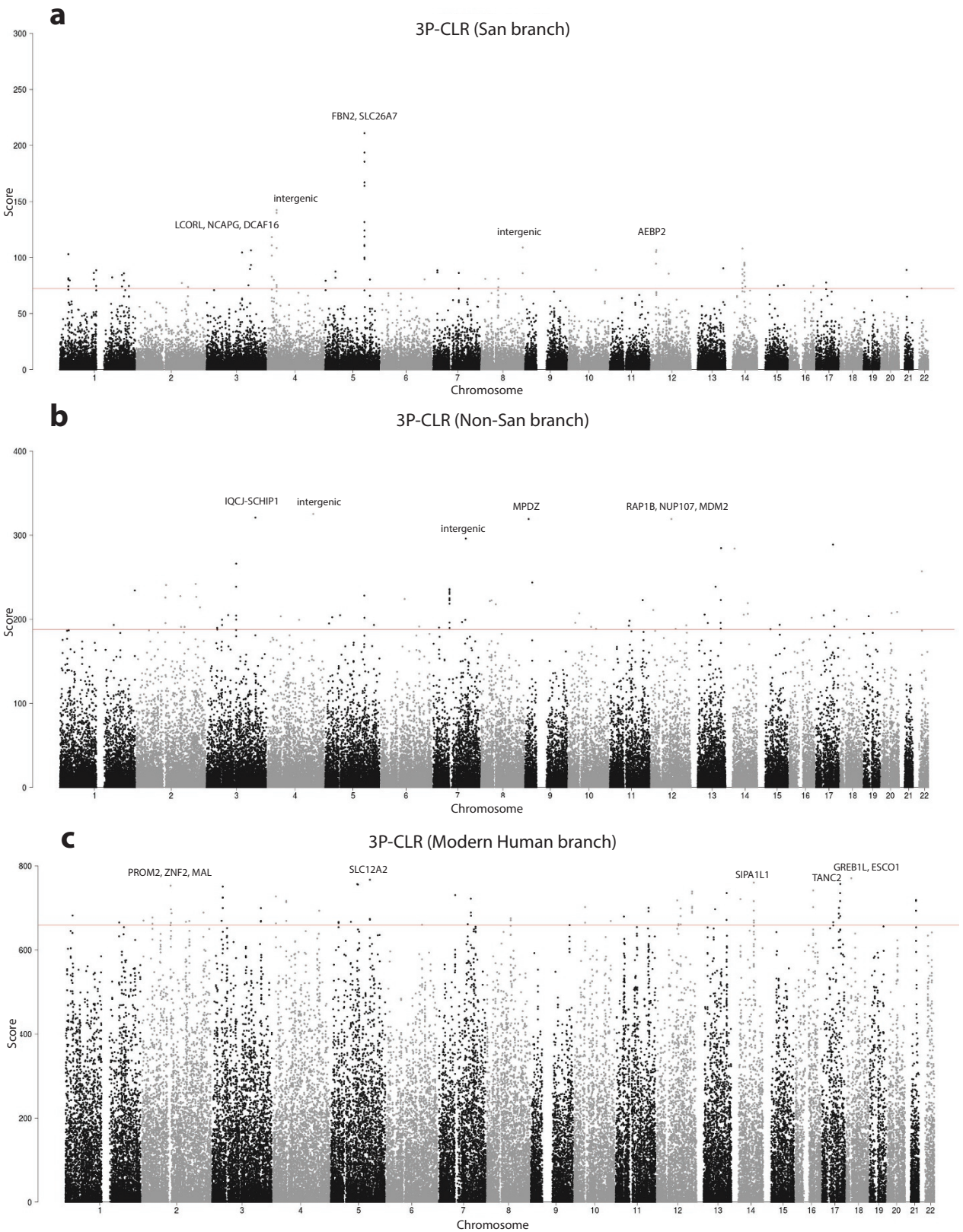
Extended Data Figure 3 | ADMIXTURE analysis. We carried out unsupervised ADMIXTURE 1.23[8,43] analysis over the 300 SGDP individuals in 20 replicates with randomly chosen initial seeds, varying the number of ancestral populations between $K = 2$ and $K = 12$ and using default fivefold cross-validation (–cv flag). We used genotypes of at least filter level 1, and restricted analysis to sites where at least two individuals carried the variant allele (as singleton variants are non-informative for population clustering). After further filtering of sites with at least 99% completeness and performing linkage-disequilibrium-based pruning in PLINK 1.9[44,45] with parameters (–indep-pairwise 1000 100 0.2), a total of 482,515 single nucleotide polymorphisms remained. This figure shows the highest likelihood replicate for each value of K. We found that log likelihood monotonically increases with K, while the value $K = 5$ minimizes cross-validation error (not shown). The solution at $K = 5$ corresponds to major continental groups (Sub-Saharan Africans, Oceanians, East Asians, Native Americans, and West Eurasians), but we show the full range of K here as they illustrate finer-scale population structure that may be useful to users of the data.



Extended Data Figure 4 | Principal component analysis and neighbour joining tree. a, Principal component analysis. **b**, Neighbour-joining tree based on F_{ST} values for all populations with at least two samples.

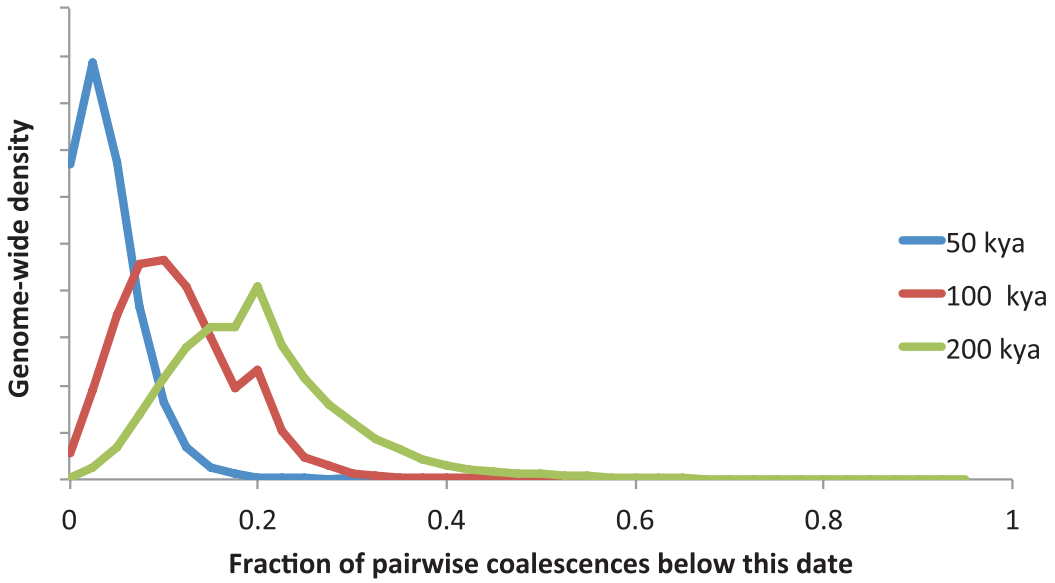


Extended Data Figure 5 | Fewer accumulated mutations in Africans than in non-Africans confirmed by mapping to chimpanzee. We compute a statistic *D* (Population A, Population B, Chimp), measuring the difference in the rate of matching to chimpanzee in Population A compared to Population B. The evidence of mismatching to chimpanzee is seen when we restrict to the male X chromosome to eliminate possible effects due to differences in heterozygosity across populations, and map to the chimpanzee genome which is phylogenetically symmetrically related to all present-day humans. We find that in 78 randomly chosen Population A = African and Population B = non-African pairs of males, transversion substitutions show no consistent skew from zero, but transition substitutions do.

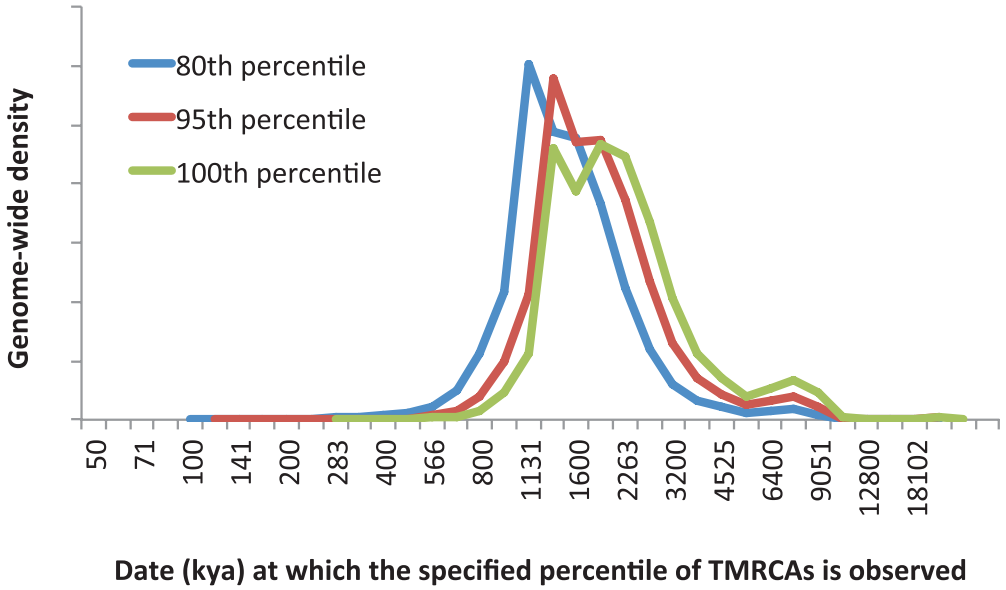


Extended Data Figure 6 | 3P-CLR scan for positive selection. The red line denotes the 99.9% quantile cut-off. The genes in the top five regions are labelled. **a**, Scan for selection on the San terminal branch. **b**, Scan for selection on the non-San terminal branch. **c**, Scan for selection on the ancestral modern human branch.

A



B



C

Date (in kya) at which a specified fraction X of loci have fraction Y of TMRCAs below date

X=Fraction of loci below threshold	Y=80% of TMRCAs less than this date	Y=95% of TMRCAs less than this date	Y=100% of TMRCAs less than this date
0.01%	120 kya	300 kya	430 kya
0.1%	320 kya	500 kya	620 kya
1%	580 kya	810 kya	980 kya

Extended Data Figure 7 | Scan for genomic locations where the great majority of present-day humans share a recent common ancestor. We carried out PSMC analysis on 40 pairs of haploid genomes chosen to sample some of the most deeply divergent present-day human lineages. We recorded the time since the most recent common ancestor (TMRCA) at each position, and rescaled to obtain an estimate of absolute time (Supplementary Information section 12). **a**, Distribution across the genome of the fraction of TMRCAs below specified date cut-offs. For the 100 kya cut-off, the maximum fraction observed anywhere in the genome is 68%. **b**, Distribution across the genome of the date *T* at which specified fractions of sample pairs are inferred to have a TMRCA less than *T*. **c**, Percentile points of the cumulative distribution function of B.

Extended Data Table 1 | Fewer accumulated mutations in Africans than in non-Africans

		All autosomes		All X chromosome		Lowest B quintile		Highest B quintile	
Population A	Population B	D×100	Z	D×100	Z	D×100	Z	D×100	Z
Khoesan	Oceania	-0.35	-8.2	-0.70	-2.7	-0.68	-6.4	-0.14	-1.7
Africa	America	-0.33	-9.4	-0.73	-2.8	-0.65	-7.3	-0.18	-2.6
Khoesan	WestEurasia	-0.30	-7.5	-0.68	-3.1	-0.63	-6.3	-0.17	-2.1
Africa	Oceania	-0.29	-8.5	-0.66	-3.2	-0.55	-6.6	-0.07	-1.0
Africa	WestEurasia	-0.25	-8.5	-0.66	-3.1	-0.49	-6.4	-0.11	-1.8
Khoesan	SouthAsia	-0.24	-6.0	-0.56	-2.7	-0.61	-6.3	-0.11	-1.4
Africa	EastAsia	-0.20	-6.6	-0.65	-2.5	-0.42	-5.2	-0.10	-1.5
Africa	CentralAsiaSiberia	-0.20	-6.2	-0.55	-2.2	-0.48	-6.3	-0.05	-0.7
Pygmy	WestEurasia	-0.19	-4.8	-0.46	-1.4	-0.43	-4.6	-0.04	-0.5
Africa	SouthAsia	-0.18	-6.4	-0.50	-2.0	-0.46	-6.3	-0.03	-0.5
CentralAsiaSiberia	Oceania	-0.13	-3.9	-0.15	-0.6	-0.09	-1.1	-0.03	-0.4
Pygmy	SouthAsia	-0.13	-3.3	-0.38	-1.1	-0.38	-4.2	0.02	0.2
EastAsia	Oceania	-0.13	-4.1	0.00	0.0	-0.17	-2.1	0.04	0.6
Khoesan	Pygmy	-0.10	-2.6	-0.14	-0.4	-0.16	-1.6	-0.12	-1.5
SouthAsia	WestEurasia	-0.08	-4.3	-0.20	-1.2	-0.05	-1.0	-0.10	-2.7
CentralAsiaSiberia	WestEurasia	-0.06	-2.2	-0.16	-0.8	-0.01	-0.2	-0.09	-1.6
EastAsia	WestEurasia	-0.06	-2.1	-0.00	-0.0	-0.08	-1.0	-0.02	-0.3
CentralAsiaSiberia	EastAsia	-0.00	-0.2	-0.18	-1.1	0.07	1.2	-0.08	-1.8
Africa	Pygmy	-0.00	-0.1	-0.06	-0.2	0.03	0.4	-0.06	-0.8
EastAsia	SouthAsia	0.02	0.7	0.22	1.7	-0.04	-0.7	0.08	1.7
CentralAsiaSiberia	SouthAsia	0.02	0.7	0.05	0.3	0.02	0.4	-0.00	-0.0
America	Oceania	0.03	0.9	0.11	0.4	0.10	1.1	0.13	1.7
Oceania	WestEurasia	0.08	2.3	-0.03	-0.1	0.10	1.1	-0.04	-0.6
Africa	Khoesan	0.10	2.9	0.17	0.7	0.23	2.6	0.07	1.0
America	WestEurasia	0.11	3.6	0.11	0.4	0.19	2.2	0.08	1.3
CentralAsiaSiberia	Pygmy	0.14	3.4	0.32	0.9	0.43	4.5	-0.04	-0.4
Oceania	SouthAsia	0.14	4.8	0.22	0.9	0.13	1.7	0.04	0.7
EastAsia	Pygmy	0.15	3.6	0.49	1.4	0.37	3.9	0.04	0.5
America	EastAsia	0.18	5.9	0.09	0.3	0.28	3.6	0.11	1.8
America	CentralAsiaSiberia	0.18	6.2	0.34	1.7	0.23	2.9	0.18	3.1
America	SouthAsia	0.18	6.4	0.34	1.5	0.22	3.0	0.18	3.1
Oceania	Pygmy	0.24	5.4	0.46	1.3	0.45	4.6	0.02	0.2
CentralAsiaSiberia	Khoesan	0.25	6.0	0.57	2.9	0.64	6.3	0.09	1.1
EastAsia	Khoesan	0.25	6.2	0.68	3.2	0.59	5.9	0.14	1.7
America	Pygmy	0.26	5.9	0.58	1.6	0.58	5.7	0.09	1.0
America	Khoesan	0.37	8.7	0.76	3.3	0.77	7.3	0.22	2.5

We compute a statistic *D* (Population A, Population B, Chimp), measuring the difference in the rate of matching to chimpanzee in Population A compared to Population B. For all the autosomes, we observe highly significant signals ($3.3 < |Z| < 9.4$) of excess mismatching to chimpanzee in non-Africans compared to Africans, using a standard error from a block jack-knife. We highlight $|D| > 0.002$ in blue, and $|Z| > 3$ in yellow. The deviations from zero are greatest in subsets of the genome where the time since two populations split comprises a relatively larger fraction of the total genetic divergence time between the populations; this is the direction expected from a mutation accumulation change since divergence. Compared to all the autosomes as a baseline, a least squares fit indicate that the deviations are 2.2 times higher on chromosome X, 2.0 times higher in the quintile of lowest B-statistic (closest to functionally important regions), and 0.43 times as high in the quintile of lowest B-statistic (furthest from functional regions).