LETTER

doi:10.1038/nature19792

Genomic analyses inform on migration events during the peopling of Eurasia

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

High-coverage whole-genome sequence studies have so far focused on a limited number[1] of geographically restricted populations[2–5], or been targeted at specific diseases, such as cancer[6]. Nevertheless, the availability of high-resolution genomic data has led to the development of new methodologies for inferring population history[7–9] and refuelled the debate on the mutation rate in humans[10]. Here we present the Estonian Biocentre Human Genome Diversity Panel (EGDP), a dataset of 483 high-coverage human genomes from 148 populations worldwide, including 379 new genomes from 125 populations, which we group into diversity and selection sets. We analyse this dataset to refine estimates of continent-wide patterns of heterozygosity, long- and short-distance gene flow, archaic admixture, and changes in effective population size through time as well as for signals of positive or balancing selection. We find a genetic signature in present-day Papuans that suggests that at least 2% of their genome originates from an early and largely extinct expansion of anatomically modern humans (AMHs) out of Africa. Together with evidence from the western Asian fossil record[11], and admixture between AMHs and Neanderthals predating the main Eurasian expansion[12], our results contribute to the mounting evidence for the presence of AMHs out of Africa earlier than 75,000 years ago.

The paths taken by AMHs out of Africa (OoA) have been the subject of considerable debate over the past two decades. Fossil and archaeological evidence[13,14], and craniometric studies[15] of African and Asian populations, demonstrate that *Homo sapiens* was present outside of Africa ~120–70 thousand years ago (kya)[11]. However, this colonization has been viewed as a failed expansion OoA[16] since genetic analyses of living populations have been consistent with a single OoA followed by serial founder events[17].

Ancient DNA (aDNA) sequencing studies have found support for admixture between early Eurasians and at least two archaic human lineages[18,19], and suggest modern humans reached Eurasia at around 100 kya[12]. In addition, aDNA from modern humans suggests population structuring and turnover, but little additional archaic admixture, in Eurasia over the last 35–45 thousand years[20–22]. Overall, these findings indicate that the majority of human genetic diversity outside Africa derives from a single dispersal event that was followed by admixture with archaic humans[18,23].

We used ADMIXTURE to analyse the genetic structure in our diversity set (Extended Data Figs 1, 2; Supplementary Information 1.1–7). We further compared the individual-level haplotype similarity of our samples using fineSTRUCTURE (Extended Data Fig. 3). Despite small sample sizes, we inferred 106 genetically distinct populations forming 12 major regional clusters, corresponding well to the 148 self-identified population labels. This clustering forms the basis for the groupings used in the scans of natural selection. Similar genetic affinities are highlighted by plotting the outgroup f_3 statistic[9] in the form $f_3(X, Y;$ Yoruba), which here measures shared drift between a non-African population X and any modern or ancient population Y from Yoruba as an African outgroup (Supplementary Information 2.2.6, Extended Data Fig. 4).

Our sampling allowed us to consider geographic features correlated with gene flow by spatially interpolating genetic similarity measures between pairs of populations (Supplementary Information 2.2.2). We considered several measures and report gradients of allele frequencies in Fig. 1, which was compared to gene flow patterns from EEMS[24] as a validation (Extended Data Fig. 5). Controlling for pairwise geographic distance, we find a correlation between these genetic gradients and geographic and climatic features such as precipitation and elevation (inset of Fig. 1, Supplementary Information 2.2.2).

We screened for evidence of selection by first focusing on loci that showed the highest allelic differentiation among groups (Supplementary Information 3). We then performed positive and purifying selection scans (Methods), and found some candidate loci that replicate previously known and functionally supported findings (Supplementary Table 1:3.3.4-I, Supplementary Information 3.1, Extended Data Fig. 6; Supplementary Table 1:3.1-IV,VI). Additionally, we infer more purifying selection in Africans in genes involved in pigmentation (bootstrapping p value (bpv) for $R_{X/Y}$ scores < 0.05) (Extended Data Fig. 6) and immune response against viruses (bpv < 0.05), while further purifying selection was indicated on olfactory receptor genes in Asians (bpv < 0.05) (Supplementary Table 1:3.1.1-II). Our scans for ancient balancing selection found a significant enrichment (FDR < 0.01) of antigen processing/presentation, antigen binding, and MHC and membrane component genes (Supplementary Information 3.2 and 3.3, Supplementary Table 1:3.3.2-I–III). The HLA (*HLA-C*)-associated gene (*BTNL2*) was the top highest scoring candidate in 8 of 12 geographic regions for the HKA test (Supplementary Table 1:3.3.1-I). Our positive selection scans, variant-based analyses (Supplementary Information 3.2 and 3.3) and gene enrichment studies also suggest new candidate loci (Supplementary Information 3.4 and 3.5, Supplementary Table 1:3.5-I–VI), a subset of which is highlighted in Supplementary Table 1:3-I.

Using fineSTRUCTURE, we find in the genomes of Papuans and Philippine Negritos more short haplotypes assigned as African than seen in genomes for individuals from other non-African populations (Extended Data Fig. 7). This pattern remains after correcting for potential confounders such as phasing errors and sampling bias (Supplementary Information 2.2.1). These shorter shared haplotypes would be consistent with an older population split[25]. Indeed, the Papuan–Yoruban median genetic split time (using multiple sequential Markovian coalescent (MSMC)) of 90 kya predates the split of all mainland Eurasian populations from Yorubans at ~75 kya (Supplementary Table 1:2.2.3-I, Extended Data Fig. 4, Fig. 2a). This result is robust to phasing artefacts (Extended Data Fig. 8, see Methods). Furthermore, the Papuan–Eurasian MSMC split time of ~40 kya is only slightly older than splits between west Eurasian and East Asian populations dated at ~30 kya (Extended Data Fig. 4). The Papuan split times from Yoruba and Eurasia are therefore incompatible with a simple bifurcating population tree model.

At least two main models could explain our estimates of older divergence dates for Sahul populations from Africa than mainland Eurasians in our sample: 1) admixture in Sahul with a potentially un-sampled archaic human population that split from modern humans either before or at the same time as did Denisova and Neanderthal; or 2) admixture in Sahul with a modern human population (extinct OoA line; xOoA) that left Africa after the split between modern humans

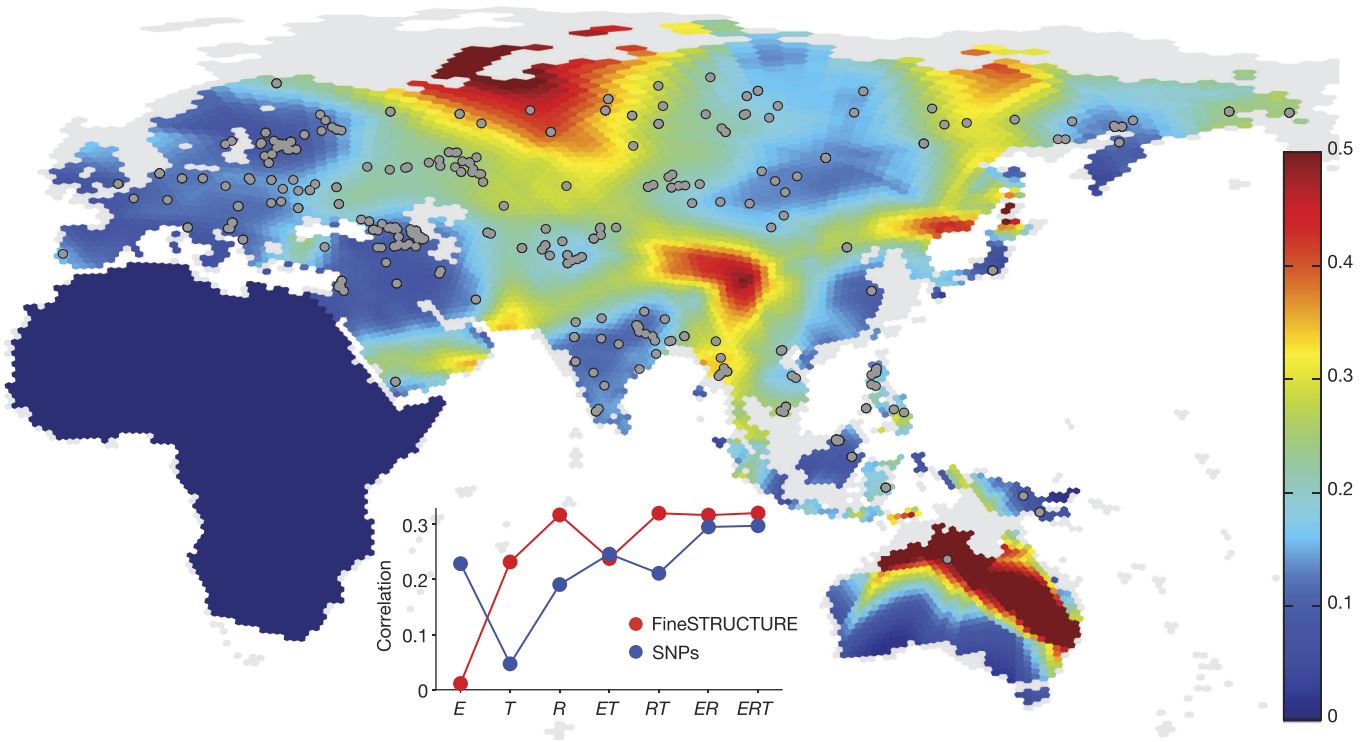


Figure 1 | Genetic barriers across space. Spatial visualization of genetic barriers inferred from genome-wide genetic distances, quantified as the magnitude of the gradient of spatially interpolated allele frequencies (value denoted by colour bar; grey areas have been land during the last glacial maximum but are currently underwater). Here we used a spatial kernel smoothing method based on the matrix of pairwise average heterozygosity and a MATLAB script that plots the hexagons of the grid with a colour coding to represent gradients. Inset, partial correlation between magnitude of genetic gradients and combinations of different geographic factors, elevation (*E*), temperature (*T*) and precipitation (*R*), for genetic gradients from fineSTRUCTURE (red) and allele frequencies (blue). This analysis (Supplementary Information 2.2.2 for details) shows that genetic differences within this region display some correlation with physical barriers such as mountain ranges, deserts, forests, and open water (such as the Wallace line).

and Neanderthals, but before the main expansion of modern humans in Eurasia (main OoA).

We consider support for these two non-mutually exclusive scenarios. Because the introgressing lineage has not been observed with aDNA, standard methods are limited in their ability to distinguish between these hypotheses. Furthermore, we show (Supplementary Information 2.2.7) that single-site statistics, such as Patterson's D[9,18] and sharing of non-African Alleles (nAAs), are inherently affected by confounding effects owing to archaic introgression in non-African populations[23]. Our approach therefore relies on multiple lines of evidence using haplotype-based MSMC and fineSTRUCTURE comparisons (which we show should have power at this timescale[26]; Supplementary Information 2.2.13).

We located and masked putatively introgressed[27] Denisova haplotypes from the genomes of Papuans, and evaluated phasing errors by symmetrically phasing Papuans and Eurasians genomes (Methods). Neither modification (Fig. 2a, Supplementary Information 2.2.9, Supplementary Table 1:2.2.9-I) changed the estimated split time (based on MSMC) between Africans and Papuans (Methods, Supplementary Information 2.2.8, Extended Data Fig. 8, Supplementary Table 1.2.8-I). MSMC dates behave approximately linearly under admixture (Extended Data Fig. 8), implying that the hypothesized lineage may have split from most Africans around 120 kya (Supplementary Information 2.2.4 and 2.2.8).

We compared the effect on the MSMC split times of an xOoA or a Denisova lineage in Papuans by extensive coalescent simulations (Supplementary Information 2.2.8). We could not simulate the large Papuan–African and Papuan–Eurasian split times inferred from the data, unless assuming an implausibly large contribution from a Denisova-like population. Furthermore, while the observed shift in the African–Papuan MSMC split curve can be qualitatively reproduced when including a 4% genomic component that diverged 120 kya from the main human lineage within Papuans, a similar quantity of Denisova admixture does not produce any significant effect (Extended Data Fig. 8). This favours a small presence of xOoA lineages rather than Denisova admixture alone as the likely cause of the observed deep African–Papuan split. We also show (Methods) that such a scenario is compatible with the observed mitochondrial DNA and Y chromosome lineages in Oceania, as also previously argued[13,28].

We further tested our hypothesized xOoA model by analysing haplotypes in the genomes of Papuans that show African ancestry not found in other Eurasian populations. We re-ran fineSTRUCTURE adding the Denisova, Altai Neanderthal and the Human Ancestral Genome sequences[29] to a subset of the diversity set. FineSTRUCTURE infers haplotypes that have a most recent common ancestor (MRCA) with another individual. Papuan haplotypes assigned as African had, regardless, an elevated level of non-African derived alleles (that is, nAAs fixed ancestral in Africans) compared to such haplotypes in Eurasians. They therefore have an older mean coalescence time with our African samples.

Owing to the deep divergence between the sampled Denisova and the one introgressed into modern humans, it is possible that some archaic haplotypes have a MRCA with an African instead of Denisova and are assigned as 'African'. We can resolve the coalescence time, and hence origin, of these haplotypes by their sequence similarity with modern Africans. To account for the archaic introgression we modelled these genomic segments as a mixture of haplotypes assigned a) as African or b) as Denisova in Eurasians and c) haplotypes assigned as Denisova in Papuans. These haplotypes are modelled (see Methods, Extended Data Fig. 9) in terms of the distribution of length and mutation rate measured as a density of non-African derived alleles. Since Eurasians (specifically Europeans) have not experienced Denisova admixture,

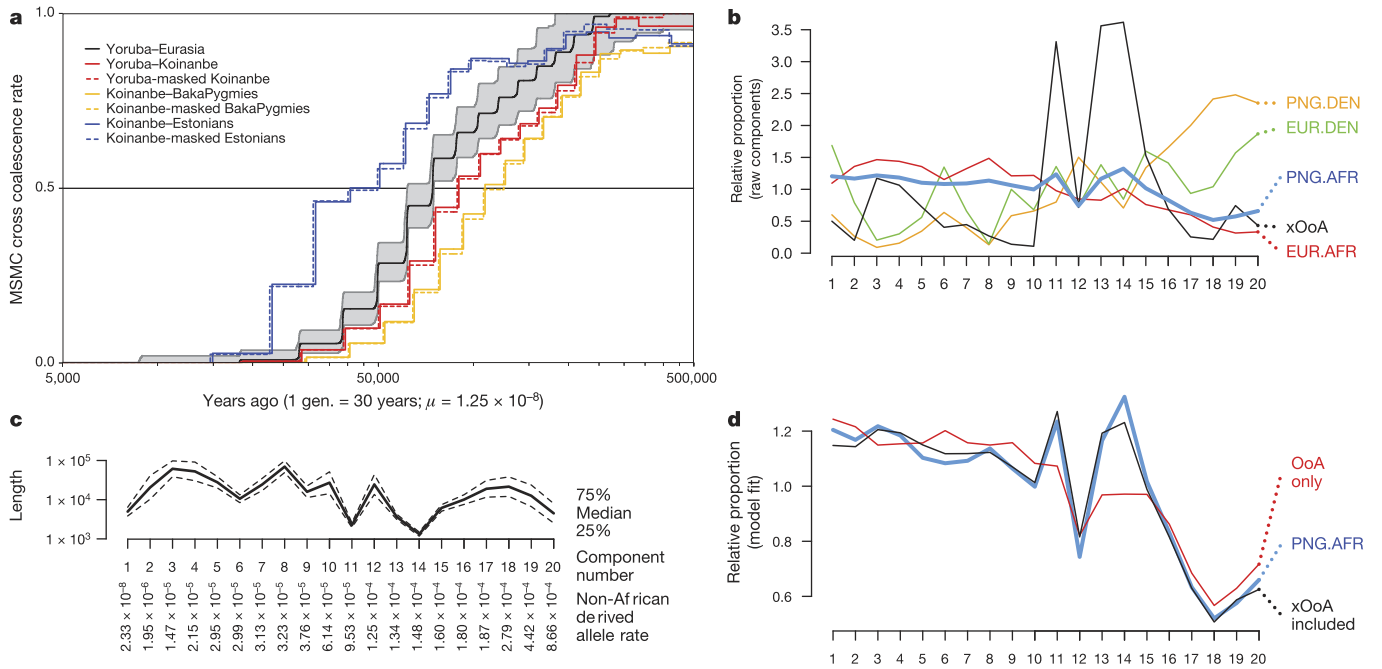


Figure 2 | Evidence of an xOoA signature in the genomes of modern Papuans. a, MSMC split times plot. The Yoruba–Eurasia split curve shows the mean of all Eurasian genomes against one Yoruba genome. The grey area represents top and bottom 5% of runs. We chose a Koinanbe genome as representative of the Sahul populations. **b**–**d**, Decomposition of Papuan haplotypes inferred as African by fineSTRUCTURE. **b**, Semi-parametric decomposition of the joint distribution of haplotype lengths and non-African derived allele rate per SNP, showing the relative proportion of haplotypes in $K = 20$ components of the distribution, ordered by non-African derived allele rate, relative to the overall proportion of haplotypes in each component. The four datasets produced by considering haplotypes inferred as (African/Denisova) in (Europeans/Papuans) are shown with our inferred 'extra Out-of-Africa' (xOoA) component. AFR, African; DEN, Denisova; PNG, Papuans; EUR, Europeans. **c**, The properties of the components in terms of non-African derived allele rate, on which the components are ordered, and length. **d**, The reconstruction of haplotypes inferred as African in the genomes of Papuan individuals, using a mixture of all other data (red) and with the addition of the xOoA signature (black).

this approach disentangles lineages that coalesce before the human/Denisova split from those that coalesce after.

We found that the xOoA signature (Fig. 2b–d; Supplementary Information 2.2.10) was necessary to account for the number of short haplotypes with 'moderate' nAAs density in the data (that is, proportion of non-African-derived sites higher than that of Eurasian haplotypes assigned as African but significantly lower than that of those assigned Denisova in either Eurasians or Papuans). Consistent with our MSMC findings (Supplementary Information 2.2.4), xOoA haplotypes have an estimated MRCA 1.5 times older than the Eurasian haplotypes in Papuan genomes, while the Denisovan haplotypes in Papuans are four times older than the Eurasian haplotypes. Adding up the contributions across the genome (Methods) leads to a genome-wide estimate of 1.9% xOoA (95% confidence interval 1.5–3.3) in Papuans, which we view as a lower bound.

Our results consistently point towards a contribution from a modern human source for derived[29] alleles that are found in the genome sequence of Papuans but not in Africans. Possible confounders could involve a shorter generation time in Papuan and Philippine Negrito populations[30], different recombination processes, or alternative demographic histories that have not been investigated here. We therefore strongly encourage the development of new model-based approaches that can investigate further the haplotype patterns described here.

In conclusion, our results suggest that while the genomes of modern Papuans derive primarily from the main expansion of modern humans out of Africa, we estimate that at least 2% of their genome sequence reflects an earlier, otherwise extinct, dispersal (Extended Data Fig. 10).

The inferred date of the xOoA split time (~120 kya) is consistent with fossil and archaeological evidence for an early expansion of *H. sapiens* from Africa[13,14]. Furthermore, the recently identified modern human admixture into the Altai Neanderthal before 100 kya[12] is consistent with a modern human presence outside Africa well before the main OoA split time (~75 kya). Further studies will confirm whether the Papuan genetic signature reported here and the one observed in Altai Neanderthals reflect the same xOoA human group, as well as clarify the timing and route followed during such an early expansion. The high similarity between Papuans and the Altai Neanderthal reported in Extended Data Fig. 1 may indeed reflect a shared xOoA component. Further studies are needed to explore this model and suggest that understanding human evolutionary history will require the recovery of aDNA from additional fossils, and further archaeological investigations in under-explored geographical regions.

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 20 September 2015; accepted 24 August 2016.
Published online 21 September 2016.

1. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2010).
2. Lachance, J. *et al.* Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150,** 457–469 (2012).
3. Pagani, L. *et al.* Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *Am. J. Hum. Genet.* **96,** 986–991 (2015).
4. Clemente, F. J. *et al.* A selective sweep on a deleterious mutation in CPT1A in Arctic populations. *Am. J. Hum. Genet.* **95,** 584–589 (2014).
5. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47,** 435–444 (2015).
6. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).
7. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).
8. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46,** 919–925 (2014).

9. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192,** 1065–1093 (2012).
10. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13,** 745–753 (2012).
11. Grove, M. *et al.* Climatic variability, plasticity, and dispersal: a case study from Lake Tana, Ethiopia. *J. Hum. Evol.* **87,** 32–47 (2015).
12. Kuhlwilm, M. *et al.* Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530,** 429–433 (2016).
13. Groucutt, H. S. *et al.* Rethinking the dispersal of *Homo sapiens* out of Africa. *Evol. Anthropol.* **24,** 149–164 (2015).
14. Liu, W. *et al.* The earliest unequivocally modern humans in southern China. *Nature* **526,** 696–699 (2015).
15. Reyes-Centeno, H. *et al.* Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proc. Natl Acad. Sci. USA* **111,** 7248–7253 (2014).
16. Mellars, P., Gori, K. C., Carr, M., Soares, P. A. & Richards, M. B. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. *Proc. Natl Acad. Sci. USA* **110,** 10699–10704 (2013).
17. Prugnolle, F., Manica, A. & Balloux, F. Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* **15,** R159–R160 (2005).
18. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
19. Reich, D. *et al.* Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* **89,** 516–528 (2011).
20. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514,** 445–449 (2014).
21. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23,** 553–559 (2013).
22. Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* **534,** 200–205 (2016).
23. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
24. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48,** 94–100 (2016).
25. Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343,** 747–751 (2014).
26. Chapman, N. H. & Thompson, E. A. A model for the length of tracts of identity by descent in finite random mating populations. *Theor. Popul. Biol.* **64,** 141–150 (2003).
27. Wall, J. D. *et al.* Higher levels of Neanderthal ancestry in East Asians than in Europeans. *Genetics* **194,** 199–209 (2013).
28. Posth, C. *et al.* Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a late glacial population turnover in Europe. *Curr. Biol.* **26,** 827–833 (2016).
29. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
30. Migliano, A. B., Vinicius, L. & Lahr, M. M. Life history trade-offs explain the evolution of human pygmies. *Proc. Natl Acad. Sci. USA* **104,** 20216–20219 (2007).

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

Acknowledgements Support was provided by: Estonian Research Infrastructure Roadmap grant no 3.2.0304.11-0312; Australian Research Council Discovery grants (DP110102635 and DP140101405) (D.M.L., M.W. and E.W.); Danish National Research Foundation; the Lundbeck Foundation and KU2016 (E.W.); ERC Starting Investigator grant (FP7 - 261213) (T.K.); Estonian Research Council grant PUT766 (G.C. and M.K.); EU European Regional Development Fund through the Centre of Excellence in Genomics to Estonian Biocentre (R.V.; M.Me. and A.Me.), and Centre of Excellence for Genomics and Translational Medicine Project No. 2014-2020.4.01.15-0012 to EGC of UT (A.Me.) and EBC (M.Me.); Estonian Institutional Research grant IUT24-1 (L.S., M.J., A.K., B.Y., K.T., C.B.M., Le.S., H.Sa., S.L., D.M.B., E.M., R.V., G.H., M.K., G.C., T.K. and M.Me.) and IUT20-60 (A.Me.); French Ministry of Foreign and European Affairs and French ANR grant number ANR-14-CE31-0013-01 (F.-X.R.); Gates Cambridge Trust Funding (E.J.); ICG SB RAS (No. VI.58.1.1) (D.V.L.); Leverhulme Programme grant no. RP2011-R-045 (A.B.M., P.G. and M.G.T.); Ministry of Education and Science of Russia; Project 6.656.2014/K (S.A.F.); NEFREX grant funded by the European Union (People Marie Curie Actions; International Research Staff Exchange Scheme; call FP7-PEOPLE-2012-IRSES-number 318979) (M.Me., G.H. and M.K.); NIH grants 5DP1ES022577 05, 1R01DK104339-01, and 1R01GM113657-01 (S.Tis.); Russian Foundation for Basic Research (grant N 14-06-00180a) (M.G.); Russian Foundation for Basic Research; grant 16-04-00890 (O.B. and E.B); Russian Science Foundation grant 14-14-00827 (O.B.); The Russian Foundation for Basic Research (14-04-00725-a), The Russian Humanitarian Scientific Foundation (13-11-02014) and the Program of the Basic Research of the RAS Presidium "Biological diversity" (E.K.K.); Wellcome Trust and Royal Society grant WT104125AIA & the Bristol Advanced Computing Research Centre (http://www.bris.ac.uk/acrc/) (D.J.L.); Wellcome Trust grant 098051 (Q.A.; C.T.-S. and Y.X.); Wellcome Trust Senior Research Fellowship grant 100719/Z/12/Z (M.G.T.); Young Explorers Grant from the National Geographic Society (8900-11) (C.A.E.); ERC Consolidator Grant 647787 'LocalAdaptatio' (A.Ma.); Program of the RAS Presidium "Basic research for the development of the Russian Arctic" (B.M.); Russian Foundation for Basic Research grant 16-06-00303 (E.B.); a Rutherford Fellowship (RDF-10-MAU-001) from the Royal Society of New Zealand (M.P.C.).

Author Contributions R.V., E.W., T.K. and M.Me. conceived the study. A.K., K.T., C.B.M., Le.S., E.P., G.A., C.M., M.W., D.L., G.Z., S.T., D.D., Z.S., G.N.N.S., K.M., J.I., L.D.D., M.G., P.N., I.E., L.At., O.U., F.-X.R., N.B., H.S., T.L., M.P.C., N.A.B., V.S., L.A., D.Pr., H.Sa., M.Mo., C.A.E., D.V.L., S.A., G.C., J.T.S.W., E.Mi., A.Ka., S.L., R.K., N.T., V.A., I.K., D.M., L.Y., D.M.B., E.B., A.Me., M.D., B.M., M.V., S.A.F., L.P.O., M.Mi., M.L., A.B.M., O.B., E.K.K, E.M., M.G.T. and E.W. conducted anthropological research and/or sample collection and management. J.L. and S.Ti. provided access to data. L.P., D.J.L, E.J., A.Mo., A.E., M.Mi., F.C., G.H., M.D., L.S., J.W., A.C., R.M., M.A.W.S., S.K., C.I., C.L.S., M.J., M.K., G.S.J., T.A., F.M.I., A.K., Q.A., C.T.-S., Y.X., B.Y., C.B.M., T.K. and M.Me. analysed data. L.P., D.J.L., E.J., A.Mo., L.S., M.K., K.T., C.B.M., Le.S., G.C., M.Mi., P.G., M.L., A.B.M., M.P., E.M., M.G.T., A.Ma., R.N., R.V., E.W., T.K. and M.Me. contributed to the interpretation of results. L.P., D.J.L., E.J., A.Mo., A.E., F.C., G.H., M.D., A.C., M.A.W.S., B.Y., J.L., S.Ti., M.Mi., P.G., M.L., A.B.M., M.P., M.G.T., A.Ma., R.N., R.V., E.W., T.K. and M.Me. wrote the manuscript.

Additional Information The newly sequenced genomes are part of the Estonian Biocentre human Genome Diversity Panel (EGDP) and were deposited in the ENA archive under accession number PRJEB12437 and are also freely available through the Estonian Biocentre website (www.ebc.ee/free_data). 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 L.P. (lp.lucapagani@gmail.com), T.K. (tk331@cam.ac.uk) or M.Me. (mait@ebc.ee).

Reviewer Information *Nature* thanks R. Dennell and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Luca Pagani[1,2,3]*, Daniel John Lawson[4]*, Evelyn Jagoda[2,5]*, Alexander Mörseburg[2]*, Anders Eriksson[6,7]*, Mario Mitt[8,9], Florian Clemente[2,10], Georgi Hudjashov[1,11,12], Michael DeGiorgio[13], Lauri Saag[1], Jeffrey D. Wall[14], Alexia Cardona[2,15], Reedik Mägi[8], Melissa A. Wilson Sayres[16,17], Sarah Kaewert[2], Charlotte Inchley[2], Christiana L. Scheib[2], Mari Järve[1], Monika Karmin[1,11,18], Guy S. Jacobs[19,20], Tiago Antao[21], Florin Mircea Iliescu[2], Alena Kushniarevich[1,22], Qasim Ayub[23], Chris Tyler-Smith[23], Yali Xue[23], Bayazit Yunusbayev[1,24], Kristiina Tambets[1], Chandana Basu Mallick[1], Lehti Saag[18], Elvira Pocheshkhova[25], George Andriadze[26], Craig Muller[27], Michael C. Westaway[28], David M. Lambert[28], Grigor Zoraqi[29], Shahlo Turdikulova[30], Dilbar Dalimova[31], Zhaxylyk Sabitov[32], Gazi Nurun Nahar Sultana[33], Joseph Lachance[34,35], Sarah Tishkoff[36], Kuvat Momynaliev[37], Jainagul Isakova[38], Larisa D. Damba[39], Marina Gubina[39], Pagbajabyn Nymadawa[40], Irina Evseeva[41,42], Lubov Atramentova[43], Olga Utevska[43], François-Xavier Ricaut[44], Nicolas Brucato[44], Herawati Sudoyo[45], Thierry Letellier[44], Murray P. Cox[12], Nikolay A. Barashkov[46,47], Vedrana Škaro[48,49], Lejla Mulahasanović[50], Dragan Primorac[49,51,52,53], Hovhannes Sahakyan[1,54], Maru Mormina[55], Christina A. Eichstaedt[2,56], Daria V. Lichman[39,57], Syafiq Abdullah[58], Gyaneshwer Chaubey[1], Joseph T. S. Wee[59], Evelin Mihailov[8], Alexandra Karunas[24,60], Sergei Litvinov[1,24,60], Rita Khusainova[24,60], Natalya Ekomasova[60], Vita Akhmetova[24], Irina Khidiyatova[24,60], Damir Marjanović[61,62], Levon Yepiskoposyan[54], Doron M. Behar[1], Elena Balanovska[63], Andres Metspalu[8,9], Miroslava Derenko[64], Boris Malyarchuk[64], Mikhail Voevoda[39,57,65], Sardana A. Fedorova[46,47], Ludmila P. Osipova[39,57], Marta Mirazón Lahr[66], Pascale Gerbault[67], Matthew Leavesley[68,69], Andrea Bamberg Migliano[70], Michael Petraglia[71], Oleg Balanovsky[63,72], Elza K. Khusnutdinova[24,60], Ene Metspalu[1,18], Mark G. Thomas[67], Andrea Manica[7], Rasmus Nielsen[27,73], Richard Villems[1,18,74]*, Eske Willerslev[27]*, Toomas Kivisild[1,2]* & Mait Metspalu[1]*

[1]Estonian Biocentre, 51010 Tartu, Estonia. [2]Department of Archaeology and Anthropology, University of Cambridge, Cambridge CB2 1QH, UK. [3]Department of Biological, Geological and Environmental Sciences, University of Bologna, Via Selmi 3, 40126 Bologna, Italy. [4]Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol BS8 2BN, UK. [5]Department of Human Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [6]Integrative Systems Biology Lab, Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. [7]Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK. [8]Estonian Genome Center, University of Tartu, 51010 Tartu, Estonia. [9]Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, 51010 Tartu, Estonia. [10]Institut de Biologie Computationnelle, Université Montpellier 2, 34095 Montpellier, France. [11]Department of Psychology, University of Auckland, Auckland 1142, New Zealand. [12]Statistics and Bioinformatics Group, Institute of Fundamental Sciences, Massey University, 4442 Palmerston North, New Zealand. [13]Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [14]Institute for Human Genetics, University of California, San Francisco, California 94143, USA. [15]MRC Epidemiology Unit, University of Cambridge, Institute of Metabolic Science, Box 285, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK. [16]School of Life Sciences, Arizona State University, Tempe, Arizona 85287, USA. [17]Center for Evolution and Medicine, The Biodesign Institute, Tempe, Arizona 85287, USA. [18]Department of Evolutionary Biology, Institute of Molecular and Cell Biology, University of Tartu, 51010 Tartu, Estonia. [19]Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK. [20]Institute for Complex Systems Simulation, University of Southampton, Southampton SO17 1BJ, UK. [21]Division of Biological Sciences,

University of Montana, Missoula, Montana 59812, USA. [22]Institute of Genetics and Cytology, National Academy of Sciences, BY-220072 Minsk, Belarus. [23]The Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. [24]Institute of Biochemistry and Genetics, Ufa Scientific Center of RAS, 450054 Ufa, Russia. [25]Kuban State Medical University, 350040 Krasnodar, Russia. [26]Scientific Research Center of the Caucasian Ethnic Groups, St. Andrews Georgian University, 0162 Tbilisi, Georgia. [27]Center for GeoGenetics, University of Copenhagen, 1350 Copenhagen, Denmark. [28]Research Centre for Human Evolution, Environmental Futures Research Institute, Griffith University, Nathan, Queensland 4111, Australia. [29]Center of Molecular Diagnosis and Genetic Research, University Hospital of Obstetrics and Gynecology, 1000 Tirana, Albania. [30]Center of High Technology, Academy of Sciences, 100047 Tashkent, Uzbekistan. [31]Institute of Bioorganic Chemistry Academy of Science, 100047 Tashkent, Uzbekistan. [32]L.N. Gumilyov Eurasian National University, 010008 Astana, Kazakhstan. [33]Centre for Advanced Research in Sciences (CARS), DNA Sequencing Research Laboratory, University of Dhaka, Dhaka-1000, Bangladesh. [34]Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6145, USA. [35]School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia 30332, USA. [36]Departments of Genetics and Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6313, USA. [37]DNcode laboratories, 117623 Moscow, Russia. [38]Institute of Molecular Biology and Medicine, 720040 Bishkek, Kyrgyzstan. [39]Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 630090 Novosibirsk, Russia. [40]Mongolian Academy of Medical Sciences, 210620 Ulaanbaatar, Mongolia. [41]Northern State Medical University, 163000 Arkhangelsk, Russia. [42]Anthony Nolan, The Royal Free Hospital, Pond Street, London NW3 2QG, UK. [43]V. N. Karazin Kharkiv National University, 61022 Kharkiv, Ukraine. [44]Evolutionary Medicine group, Laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse, UMR 5288, Centre National de la Recherche Scientifique, Université de Toulouse 3, Toulouse 31073, France. [45]Genome Diversity and Diseases Laboratory, Eijkman Institute for Molecular Biology, 10430 Jakarta, Indonesia. [46]Department of Molecular Genetics, Yakut Scientific Centre of Complex Medical Problems, 677027 Yakutsk, Russia. [47]Laboratory of Molecular Biology, Institute of Natural Sciences, M.K. Ammosov North-Eastern Federal University, 677027 Yakutsk, Russia. [48]Genos DNA laboratory, 10000 Zagreb, Croatia. [49]University of Osijek, Medical School, 31000 Osijek, Croatia. [50]Center for Genomics and Transcriptomics, CeGaT, GmbH, D-72076 Tübingen, Germany. [51]St. Catherine Specialty Hospital, 49210 Zabok and 10000 Zagreb, Croatia. [52]Eberly College of Science, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [53]University of Split, Medical School, 21000 Split, Croatia. [54]Laboratory of Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences, Republic of Armenia, 7 Hasratyan Street, 0014 Yerevan, Armenia. [55]Department of Applied Social Sciences, University of Winchester, Sparkford Road, Winchester SO22 4NR, UK. [56]Thoraxklinik Heidelberg, University Hospital Heidelberg, 69120 Heidelberg, Germany. [57]Novosibirsk State University, 630090 Novosibirsk, Russia. [58]RIPAS Hospital, Bandar Seri Begawan, BE1518 Brunei. [59]National Cancer Centre Singapore, 169610 Singapore. [60]Department of Genetics and Fundamental Medicine, Bashkir State University, 450000 Ufa, Russia. [61]Department of Genetics and Bioengineering. Faculty of Engineering and Information Technologies, International Burch University, 71000 Sarajevo, Bosnia and Herzegovina. [62]Institute for Anthropological Researches, 10000 Zagreb, Croatia. [63]Research Centre for Medical Genetics, Russian Academy of Sciences, Moscow 115478, Russia. [64]Genetics Laboratory, Institute of Biological Problems of the North, Russian Academy of Sciences, 685000 Magadan, Russia. [65]Institute of Internal Medicine, Siberian Branch of Russian Academy of Medical Sciences, 630009 Novosibirsk, Russia. [66]Leverhulme Centre for Human Evolutionary Studies, Department of Archaeology and Anthropology, University of Cambridge, Cambridge CB2 1QH, UK. [67]Research Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK. [68]Department of Archaeology, University of Papua New Guinea, University PO Box 320, 134 NCD, Papua New Guinea. [69]College of Arts, Society and Education, James Cook University, PO Box 6811, Cairns, Queensland 4870, Australia. [70]Department of Anthropology, University College London, London WC1H 0BW, UK. [71]Max Planck Institute for the Science of Human History, Kahlaische Strasse 10, D-07743 Jena, Germany. [72]Vavilov Institute for General Genetics, Russian Academy of Sciences, 119333 Moscow, Russia. [73]Department of Integrative Biology, University of California Berkeley, Berkeley 94720, California, USA. [74]Estonian Academy of Sciences, 6 Kohtu Street, Tallinn 10130, Estonia.
*These authors contributed equally to this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Data preparation. We analyse a set of genomes sequenced by the same technology (Complete Genomics Inc.) which results in minimal platform differences between batches of samples analysed by slight modifications of CG proprietary pipeline (Extended Data Fig. 2; Supplementary Information 1.6). Informed consent forms and REC approvals were obtained for all samples newly collected for this study. We see good concordance between CG sequence and Illumina genotyping array results for the same samples with minor reference bias in the latter data (Extended Data Fig. 2; Supplementary Information 1.6). In the final dataset, we retained only one second-degree (Australians, to make use of all the available samples) and five third-degree relatives pairs (Supplementary Table 1:1.7-I). All genomes were annotated against the Ensembl GRCh37 database and compared to dbSNP Human Build 141 and Phase 1 of the 1000 Genomes Project dataset[29] (Supplementary Information 1.1–1.6). We found 10,212,117 new SNPs, 401,911 of which were exonic. As expected from our sampling scheme, existing lists of variable sites have been extended mostly by the Siberian, Southeast Asian and South Asian genomes, which contribute 89,836 (22.4%), 63,964 (15.9%) and 40,758 (10.1%) of the new exonic variants detected in this study.

Compared to the genome-wide average, we see fewer heterozygous sites on chromosomes 1 and 2, and an excess on chromosomes 16, 19 and 21 (Extended Data Fig. 2). This pattern is independent of simple potential confounders, such as rough estimates of recombination activity and gene density (Supplementary Information 1.8), and mirrors the inter-chromosomal differences in divergence from chimpanzee[31], suggesting large-scale differences in mutation rates among chromosomes. We confirmed this general pattern using 1000 Genomes Project data (Supplementary Information 1.8).

The 'ancient genome diversity panel' consisted of 106 samples from the main Diversity panel along with Altai Neanderthal, Denisova and the Modern Human reference genome. Sites that are heterozygous in archaic humans were removed.

Geographic gradient analyses. We used a Gaussian kernel smoothing (based on the shortest distance on land to each sample) to interpolate genetic patterns across space. Averaging over all markers, we obtained an expression for the mean square gradient of allele frequencies in terms of the matrix of genetic distance between pairs of samples (Supplementary Information 2.2.2). This provides a simple way to identify spatial regions that contribute strongly to genetic differences between samples, and can be used, in principle, for any measure of genetic difference (for fineSTRUCTURE data, we used negative shared haplotype length as a measure of differentiation).

To quantify the link between the magnitude of genetic gradients (from fineSTRUCTURE and allele frequency data) and geographic factors, we fitted a generalized linear model to the sum of genetic magnitude gradients on the shortest paths between samples to elevation, minimum quarterly temperature, and annual precipitation summed in the same way, controlling for path length and spatial random effects (Supplementary Information 2.2.2), and calculated partial correlations between genetic gradient magnitudes and geographic factors.

FineSTRUCTURE analysis. FineSTRUCTURE[32] was run as described in Supplementary Information 2.2.1. Within the 106 genetically distinct genetic groups, labels were typically genetically homogeneous—113 of the 148 population labels (76%) were assigned to only one 'genetic cluster'. Similarly, genetic clusters were typically specific to a label, with 66 of the 106 'genetic clusters' (62%) containing only one population label.

Correction for phasing errors. To check whether phasing errors could produce the shorter Papuan haplotypes, we focused on regions of the genome that had an extended (>500 kb) run of homozygosity. We ran ChromoPainter for each individual on only these regions, meaning each individual was only painted where it had been perfectly phased. This did not change the qualitative features (Supplementary Information 2.2.1).

Removal of similar samples. Papuans are genetically distinct from other populations due to tens of thousands of years of isolation. We wanted to check whether the length of haplotypes assigned as African was biased by the inclusion of a large number of relatively homogeneous Eurasians with few Papuans. To do this we repeated the $n = 447$ painting allowing only donors from dissimilar populations, including only individuals who donated < 2% of a genome in the main painting. This did not change the qualitative haplotype length features (Supplementary Information 2.2.1).

Inclusion of ancient samples. We ran our smaller individual panel with ($n = 109$) and without ($n = 106$) ancient samples (Denisova, Neanderthal and ancestral human). This did not change the qualitative haplotype length features (Supplementary Information 2.2.1).

Selection analyses. We investigated balancing, positive and purifying selection for a part of the dataset with larger group sizes which was defined as the Selection subset (Supplementary Table 1:3.1-I and 3.2-I) using a wide range of window-based as well as variant-based approaches. Furthermore, we investigated how these signals relate to shared demographic history. Where possible we contextualized our findings by integrating them with information from various functional databases. Detailed descriptions of all methods used are available in Supplementary Information section 3.

MSMC, Denisova masking, simulations of alternative scenarios and assessment of phasing robustness. Genetic split times were initially calculated following the standard MSMC procedure[8], and subsequently modified as follows. To estimate the effect of archaic admixture, putative Denisova haplotypes were identified in Papuans using a previously published method[27] and masked from all the analysed genomes. Particularly, whether a putative archaic haplotype was found in heterozygous or homozygous state within the chosen Papuan genome, the 'affected' locus was inserted into the MSMC mask files and, hence, removed from the analysis.

We note that a fraction of the Denisova and Neanderthal contributions to the Papuan genomes may be indistinguishable, owing to the shared evolutionary history of these two archaic populations. As a result, some of the removed 'Denisova' haplotypes may have actually entered the genome of Papuans through Neanderthal. Regardless of this, our exercise successfully shows that the MSMC split time estimates are not affected by the documented presence of archaic genomic component (whether coming entirely from Denisova or partially shared with Neanderthal).

We further excluded the role of Denisova admixture in explaining the deeper African–Papuan MSMC split times through coalescent simulations (using ms to generate 30 chromosomes of 5 Mbp each, and simulating each scenario 30 times). These showed that the addition of 4% Denisova lineages to the Papuan genomes does not change the MSMC results, while the addition of 4% xOoA lineages recreates the qualitative shift observed in the empirical data.

Phasing artefacts were also taken into account as putative confounders of the MSMC split time estimates. We re-ran MSMC after re-phasing one Estonian, one Papuan and 20 West African and Pygmy genomes in a single experiment. This way we ruled out potential artefacts stemming from the excess of Eurasian over Sahul samples during the phasing process. Both the archaic and phasing corrections yielded the same split time as of the standard MSMC runs.

Emulation of all pairwise MSMC split times. We confirmed that none of the other populations behaved as an outlier from those identified in the $n = 22$ full pairwise analysis by estimating the MSMC split times between all pairs. We chose 9 representative populations (including Papuan, Yoruba and Baka) from the 22, and compared each of the 447 diversity panel genomes to them. For each individual l not in our panel, we obtain the positive mixture weights α_k using the model

$$\hat{t}_{lj} = \sum_{k=1}^{9} \alpha_{lk} t_{lj} \text{ for } j \in (1..9)$$

The parameters are estimated using the $j \in (1..9)$ observations for which we have data using a quadratic loss function. We can then predict the unobserved values

$$\hat{t}_{li} = \sum_{k=1}^{9} \alpha_k t_{ki}$$

Examination of this matrix (Supplementary Information 2.2.3, Supplementary Table 1:2.2.3-III) implies no other populations are expected to have unusual MSMC split times from Africa.

Mixture model for African haplotypes in Papuans. *Obtaining haplotypes from painting.* We define African or Archaic haplotypes in Eurasians or Papuans as genomic loci spanning at least 1,000 bp, and showing SNPs that were assigned by chromopainter a ≥50% chance of copying from either an African or Archaic genome, respectively. For each haplotype we then calculated the number of non-African mutations, defined as sites found in derived state in a given haplotype and in ancestral state in all of the African genomes included in the present study.

Modelling. We used a non-parametric model for the joint distribution of length and non-African derived allele mutation rate in haplotypes. We fit $K = 20$ components to the joint distribution. Each component has a characteristic length l_k, variability σ_k and mutation rate μ_k. A haplotype of length l_i with X_i such mutations from component $I_i = k$ has the following distribution:

$$l_i | \{l_k, \sigma_k^2, I_i = k\} \sim \text{log-Normal}(l_k, \sigma_k^2)$$

$$X_i | \{l_k, \mu_k, I_i = k\} \sim \text{Binomial}(l_k, \mu_k)$$

This model for haplotype lengths is motivated by the extreme age of the split times we seek to model. Recent splits would lead to an exponential distribution of haplotype lengths. However, owing to haplotype fixation caused by finite population size, very old splits have finite (non-zero) haplotype lengths. Additionally, the data are left-censored since we cannot reliably detect haplotypes that are very short. We note that while this makes a single component a reasonable fit to the data, as K increases the specific choice becomes less important.

We then impose the prior $p(I_i = k) = 1/K$ and use the expectation-maximization algorithm to estimate the mixture proportions $\pi_{ik} = E(I_{ik}|l_i, X_i)$ along with the maximum likelihood parameter estimates $\{l_k, \sigma_k^2, \mu_k\}$. We do this for the four combinations of haplotypes assigned as African (AFR) and Denisova (DEN) found in Papuans (PNG) or Europeans (EUR), in order to learn the parameters. Supplementary Information 2.2.10 describes this in more detail. We then describe the distribution of haplotypes for each class c of haplotype in terms of the expected proportion of haplotypes found in each component,

$$\pi_{ck} = \frac{\pi'_{ck}}{\sum_{k=1}^{K} \pi'_{ck}} \text{ where } \pi'_{ck} = \sum_{i=1}^{N_c} \pi_{cik}$$

where N_c is the number of haplotypes of class c. π_c is a vector of the proportions from each of the K components.

Single-out-of-Africa model. We fit haplotypes assigned as African in Papuans as a mixture of the others in a second layer of mixture modelling:

$$\pi_{\text{PNG.AFR}} = \sum_{c \in \{\text{PNG.DEN,EUR.AFR,EUR.DEN}\}} \alpha_c \, \pi_c$$

where α_c sum to 1. This is straightforward to fit.

xOoA model. We jointly estimate an additional component π_{xOoA} and the mixture contributions β_c under the mixture

$$\pi_{\text{PNG.AFR}} = \sum_{c \in \{\text{PNG.DEN,EUR.AFR,EUR.DEN,xOoA}\}} \beta_c \, \pi_c$$

This is non-trivial to fit. We use a penalization scheme to simultaneously ensure we a) obtain a valid mixture for β_c; b) give a prediction x_k that is also a valid mixture; c) leave little signal in the residuals; and d) obtain a good fit. Cross-validation is used to obtain the optimal penalization parameters (A and B) with the loss function:

$$\text{loss} = \sum_{k=1}^{K} e_k^2 + AP_A + BP_B,$$

where e_k are the residuals in each component, $P_A = \left|\left(\sum_c \beta_c\right) - 1\right| + \left|\left(\sum_k x_k\right) - 1\right|$ (for a valid mixture) and $P_B = s \,.\, d(e_k)$ (for requirement c, good solutions will have similar residuals across components). The loss is minimized via standard optimization techniques. Supplementary Information 2.2.10 details how initial values are found and explores the robustness of the solution to changes in A and B—the results do not change qualitatively for reasonable choices of these parameters, and the mixtures are valid to within numerical error.

Genome-wide xOoA estimation. We used the estimated xOoA derived allele mutation rate estimate θ_{xOoA} to estimate the xOoA contribution in haplotypes classed as Eurasian or Papuan by ChromoPainter. First we obtained estimates of $\pi_{\text{PNG.EUR}}$ and $\pi_{\text{PNG.PNG}}$ using the single out-of-Africa model above, additionally allowing for a EUR.EUR contribution. We then estimate α_{xOoA} using the observed mutation rate θ_{obs} and that predicted under the mixture model θ_{mix} by rearranging the mixture:

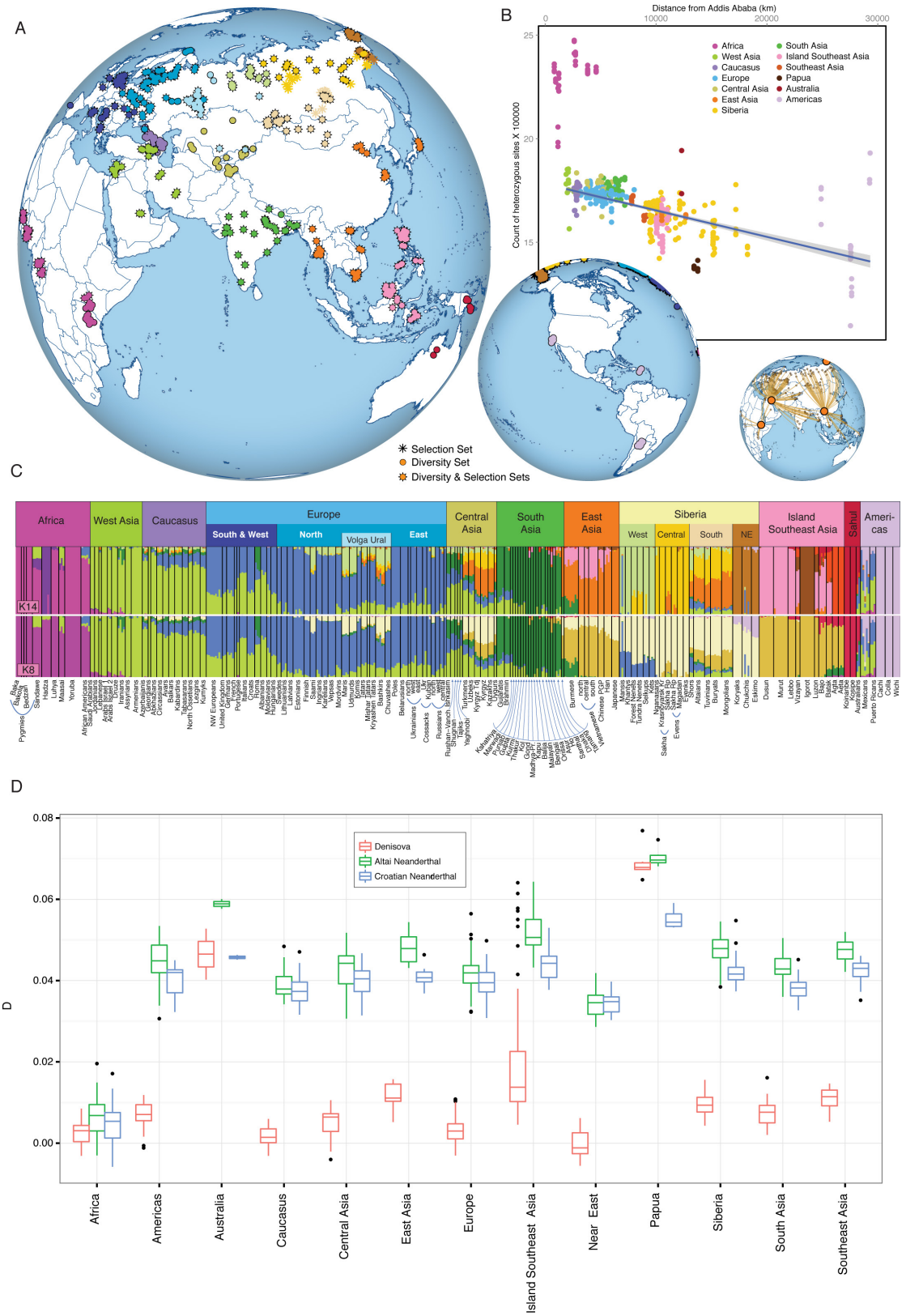
$$\theta_{\text{obs}} = \alpha_{\text{xOoA}}\theta_{\text{xOoA}} + (1 - \alpha_{\text{xOoA}})\theta_{\text{mix}}$$

Estimates less than 0 are set to 0. The genome-wide estimate is obtained by weighting each θ by the proportion of the genome that was painted with that donor. Neanderthal and Denisova haplotypes were assumed to be proxied by PNG.DEN (0% xOoA by assumption); African haplotypes by PNG.AFR; Papuan and Australian by PNG.PNG and all other haplotypes by PNG.EUR. We obtain confidence intervals by bootstrap resampling of haplotypes for each donor/recipient pair.

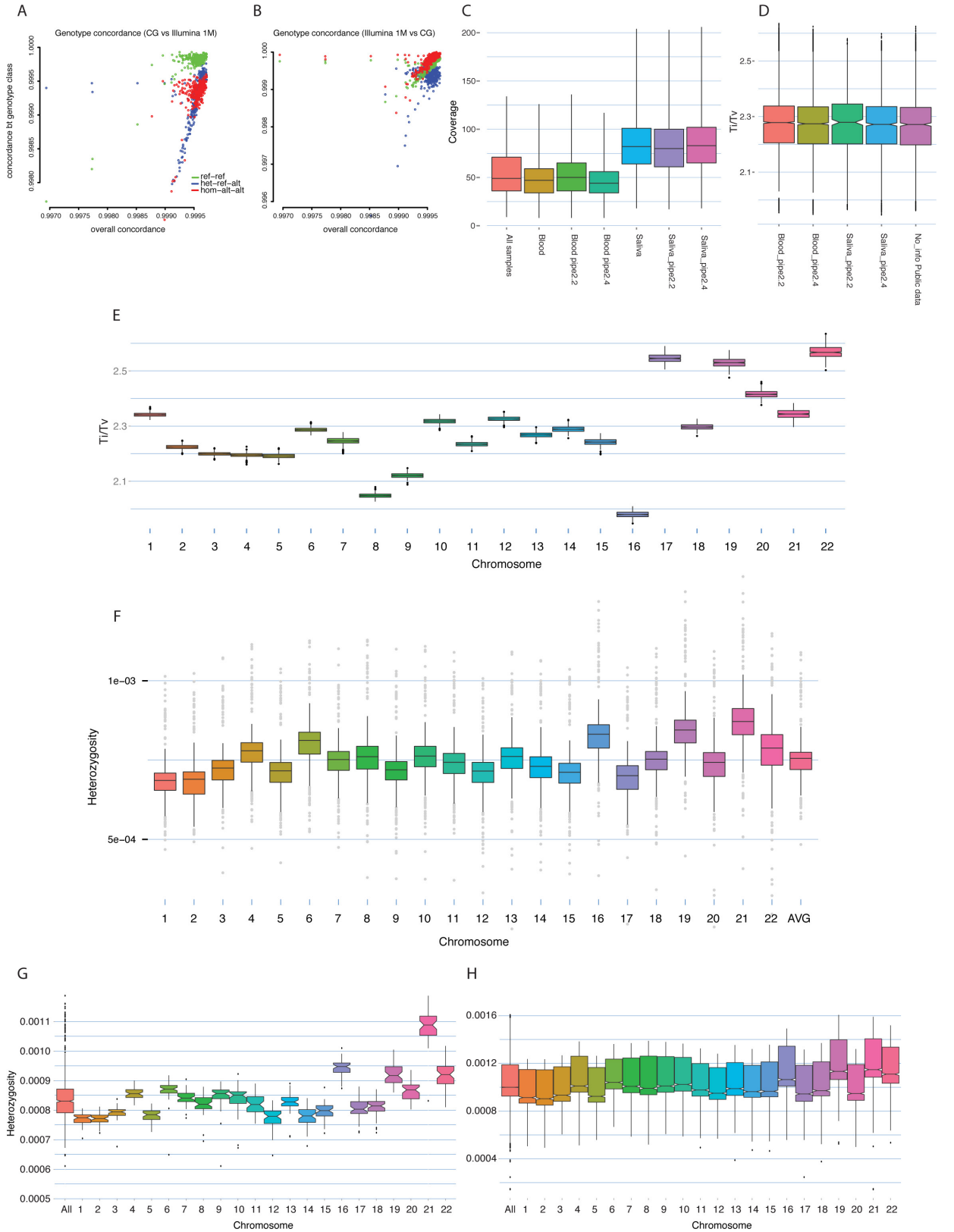
We estimate the proportion of xOoA in Papuan haplotypes assigned as both Eurasian (0.1%, 95% CI 0–2.6) and Papuan (4%, 95% CI 2.9–4.5) (Supplementary Information 2.2.10), by using the estimated mutation density in xOoA.

Y chromosome and mtDNA haplopgroup analysis. The presence of an extinct xOoA trace in the genome of modern Papuans may seem at odds with analyses of mtDNA and Y chromosome phylogenies, which point to a single, recent origin for all non-African lineages (mtDNA L3, which gives rise to all mtDNA lineages outside Africa has been dated at ~70,000 years old[33,34]). However, uniparental markers inform on a small fraction of our genetic history, and a single origin for all non-African lineages does not exclude multiple waves OoA from a shared common ancestor. We show analytically (Supplementary Information 2.2.12) that, if the xOoA signature entered the genome of Papuan individuals > 40 kya, their mtDNA and Y lineages could have been lost by genetic drift even assuming an initial xOoA mixing component of up to 35%. Similar findings have been reported recently[13].

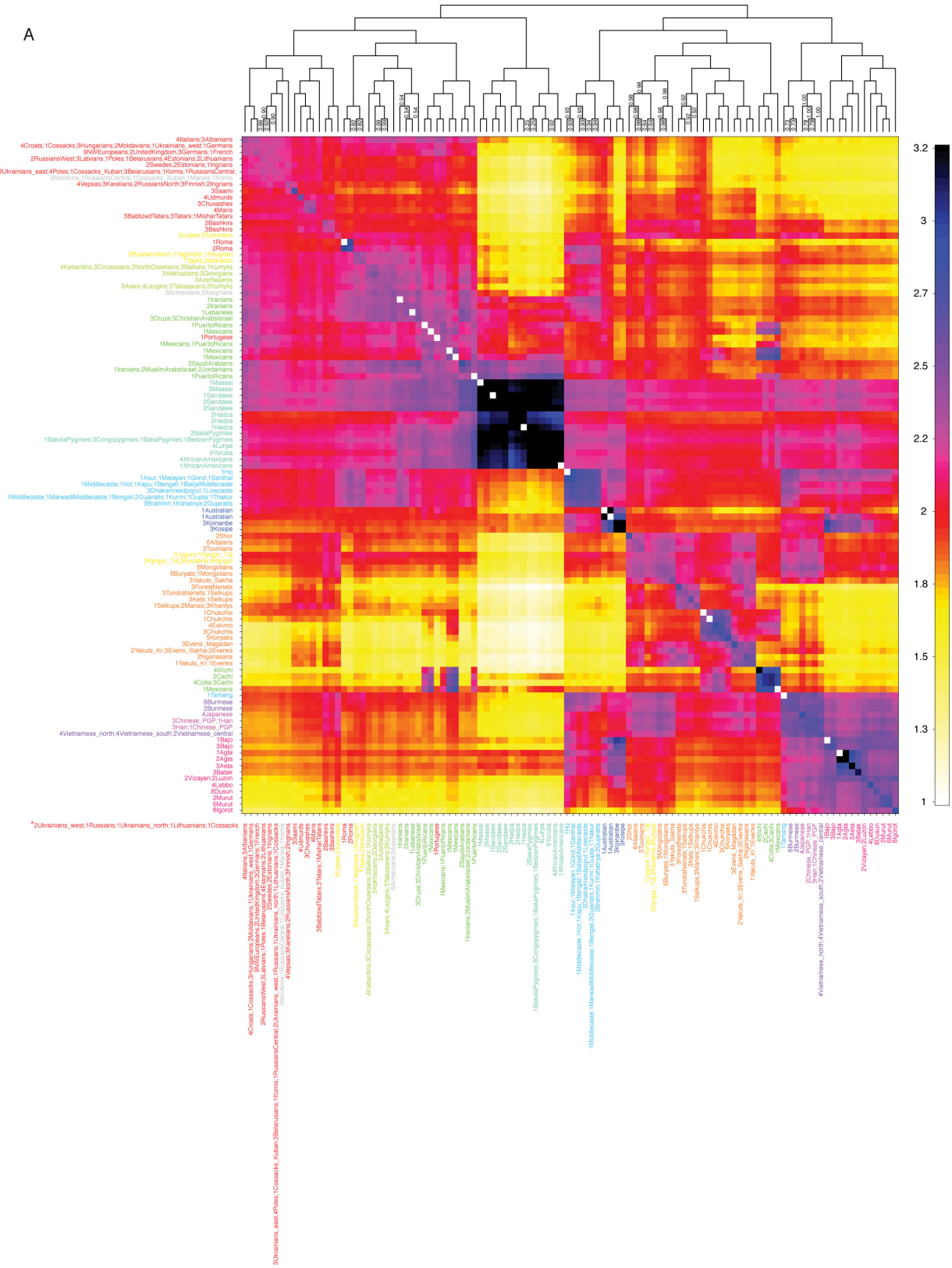
31. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
32. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8,** e1002453 (2012).
33. Behar, D. M. *et al.* A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90,** 675–684 (2012).
34. Soares, P. *et al.* The archaeogenetics of Europe. *Curr. Biol.* **20,** R174–R183 (2010).



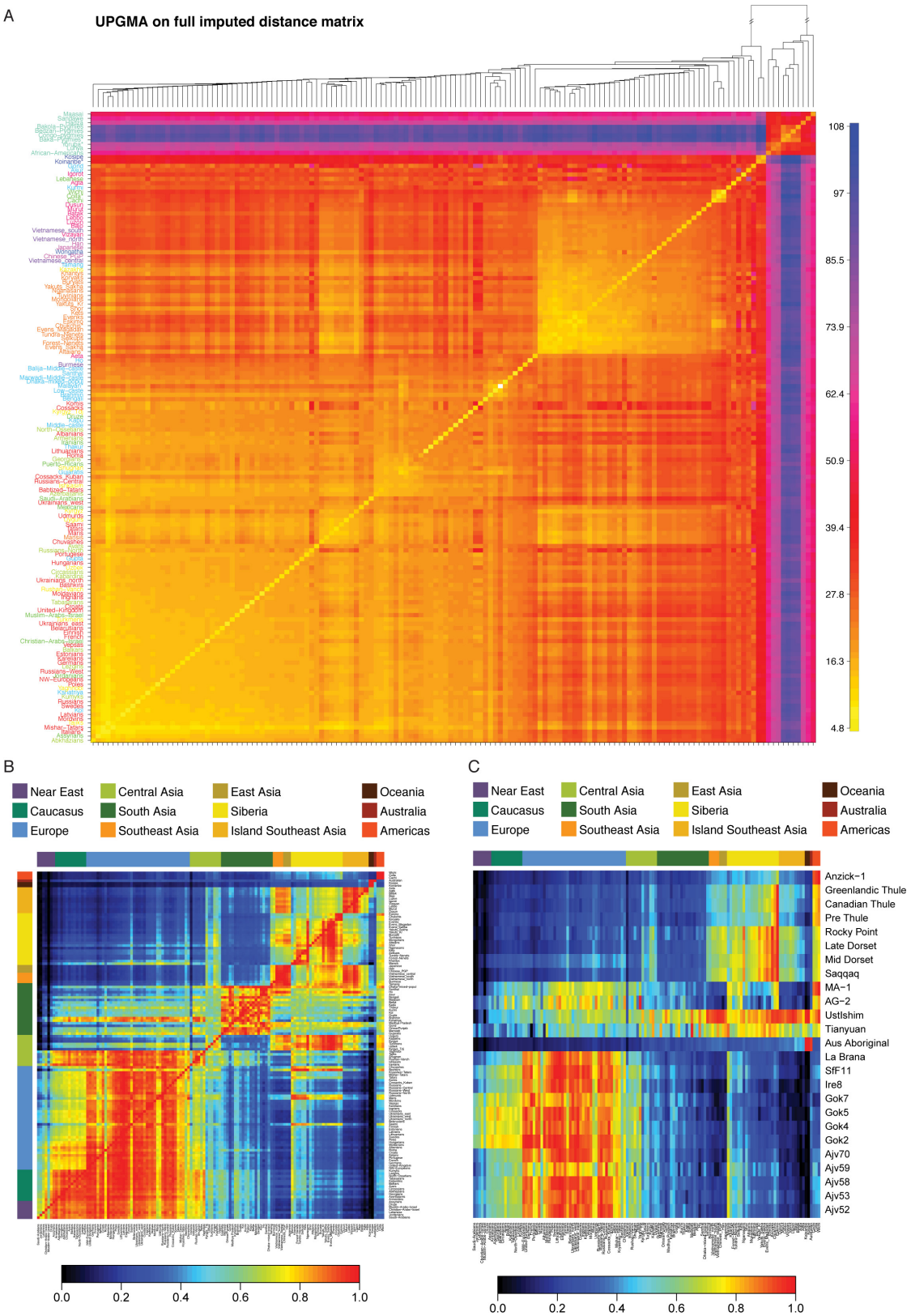
Extended Data Figure 1 | Sample Diversity and Archaic signals. **a**, Map of location of samples highlighting the diversity/selection sets. **b**, Sample-level heterozygosity is plotted against distance from Addis Ababa. The trend line represents only non-African samples. The inset shows the waypoints used to arrive at the distance in kilometres for each sample. **c**, ADMIXTURE plot ($K = 8$ and 14) which relates general visual inspection of genetic structure to studied populations and their region of origin. **d**, Box plots were used to visualize the Denisova (red), Altai (green) and Croatian Neanderthal (blue) D distribution for each regional group of samples. Oceanian Altai D values show a remarkable similarity with the Denisova D values for the same region, in contrast with the other groups of samples where the Altai box plots tend to be more similar to the Croatian Neanderthal ones. Boxes show median, first and third quartiles, with 1.5× interquartile range whiskers and black dots as outliers.



Extended Data Figure 2 | Data quality checks and heterozygosity patterns. a, **b**, Concordance of DNA sequencing (Complete Genomics Inc.) and DNA genotyping (Illumina genotyping arrays) data (ref-ref; het-ref-alt and hom-alt-alt, see Supplementary Information 1.6) from chip (**a**) and sequence data (**b**). **c**, Coverage (depth) distribution of variable positions, divided by DNA source (blood or saliva) and complete genomic calling pipeline (release version). **d**, Genome-wide distribution of transition/transversion ratio subdivided by DNA source (saliva or blood) and by complete genomic calling pipeline. **e**, Genome-wide distribution of transition/transversion ratio subdivided by chromosomes. **f**, Inter-chromosome differences in observed heterozygosity in 447 samples from the diversity set. **g**, Inter-chromosome differences in observed heterozygosity in a set of 50 unpublished genomes from the Estonian Genome Center, sequenced on an Illumina platform at an average coverage exceeding 30×. **h**, Inter-chromosome differences in observed heterozygosity in the phase 3 of the 1000 Genomes Project. The total number of observed heterozygous sites was divided by the number of accessible base pairs reported by the 1000 Genomes Project.

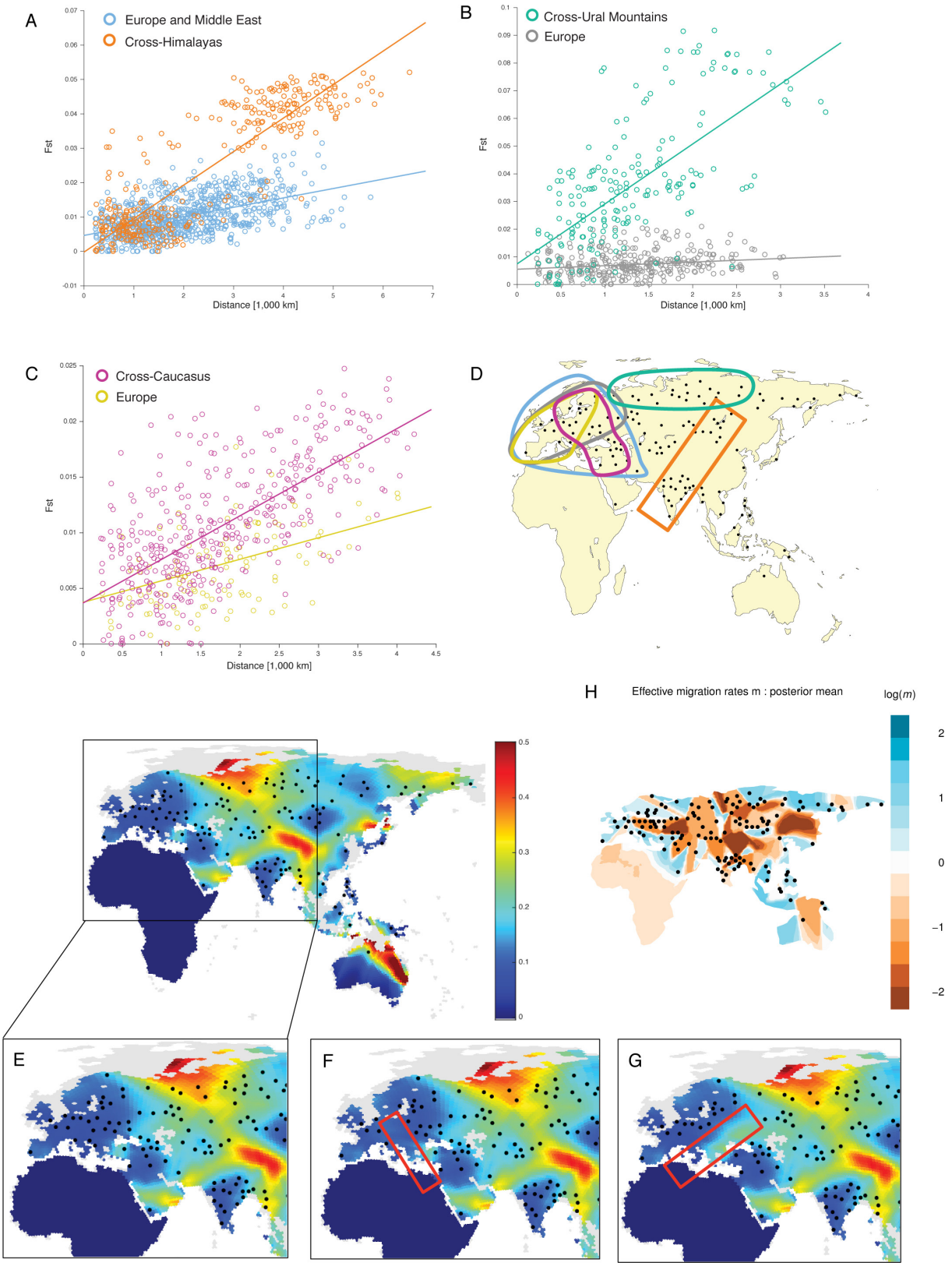


Extended Data Figure 3 | FineSTRUCTURE shared ancestry analysis. ChromoPainter and FineSTRUCTURE results, showing both inferred populations and the underlying (averaged) number of haplotypes that an individual in a population receives (rows) from donor individuals in other populations (columns). 108 populations are inferred by FineSTRUCTURE. The dendrogram shows the inferred relationship between populations. The numbers on the dendrogram give the proportion of MCMC iterations for which each population split is observed (where this is less than 1). Each 'geographical region' has a unique colour from which individuals are labelled. The number of individuals in each population is given in the label; for example, '4Italians; 3Albanians' is a population of size 7 containing 4 individuals from Italy and 3 from Albania.

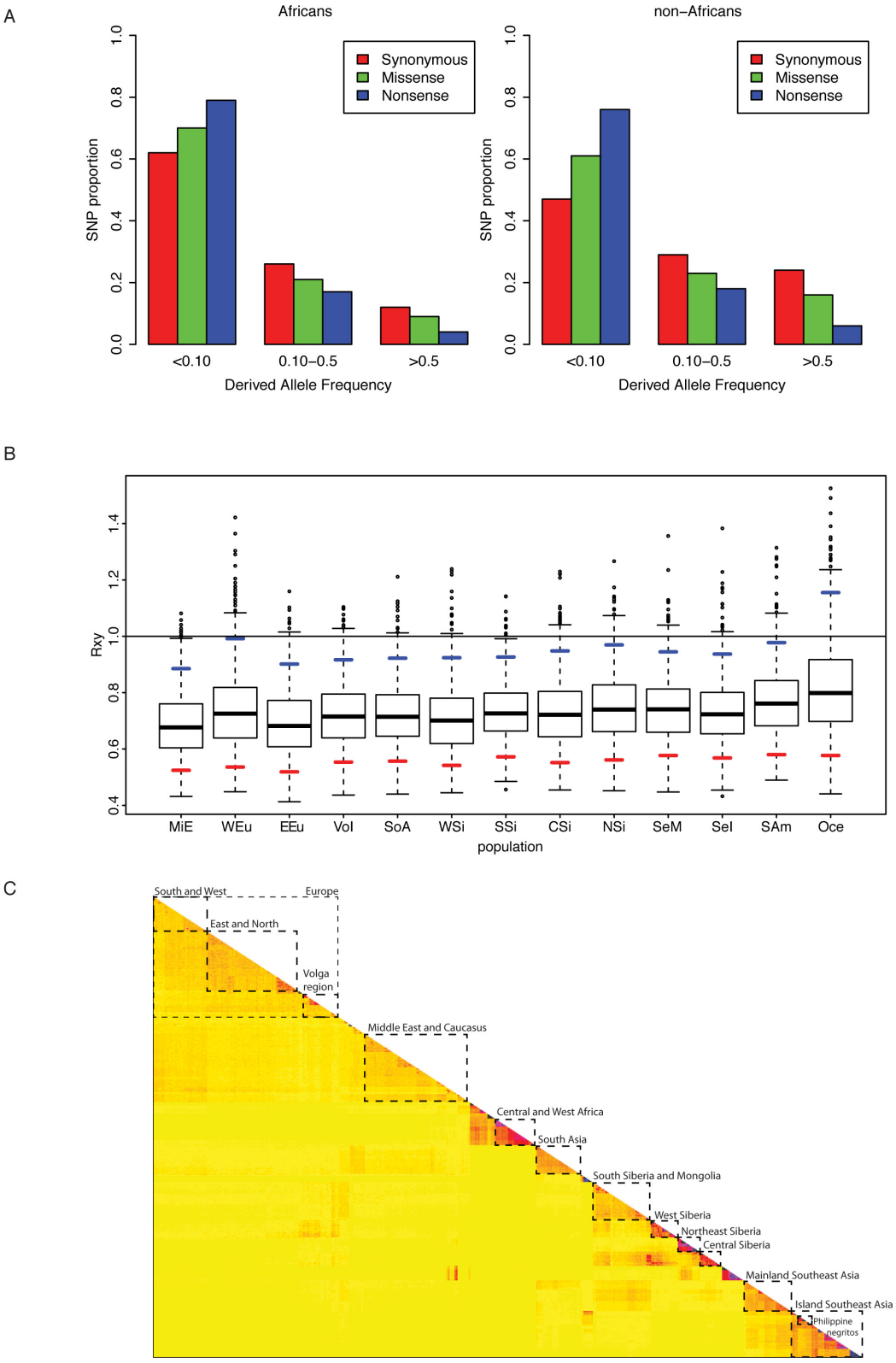


Extended Data Figure 4 | See next page for caption.

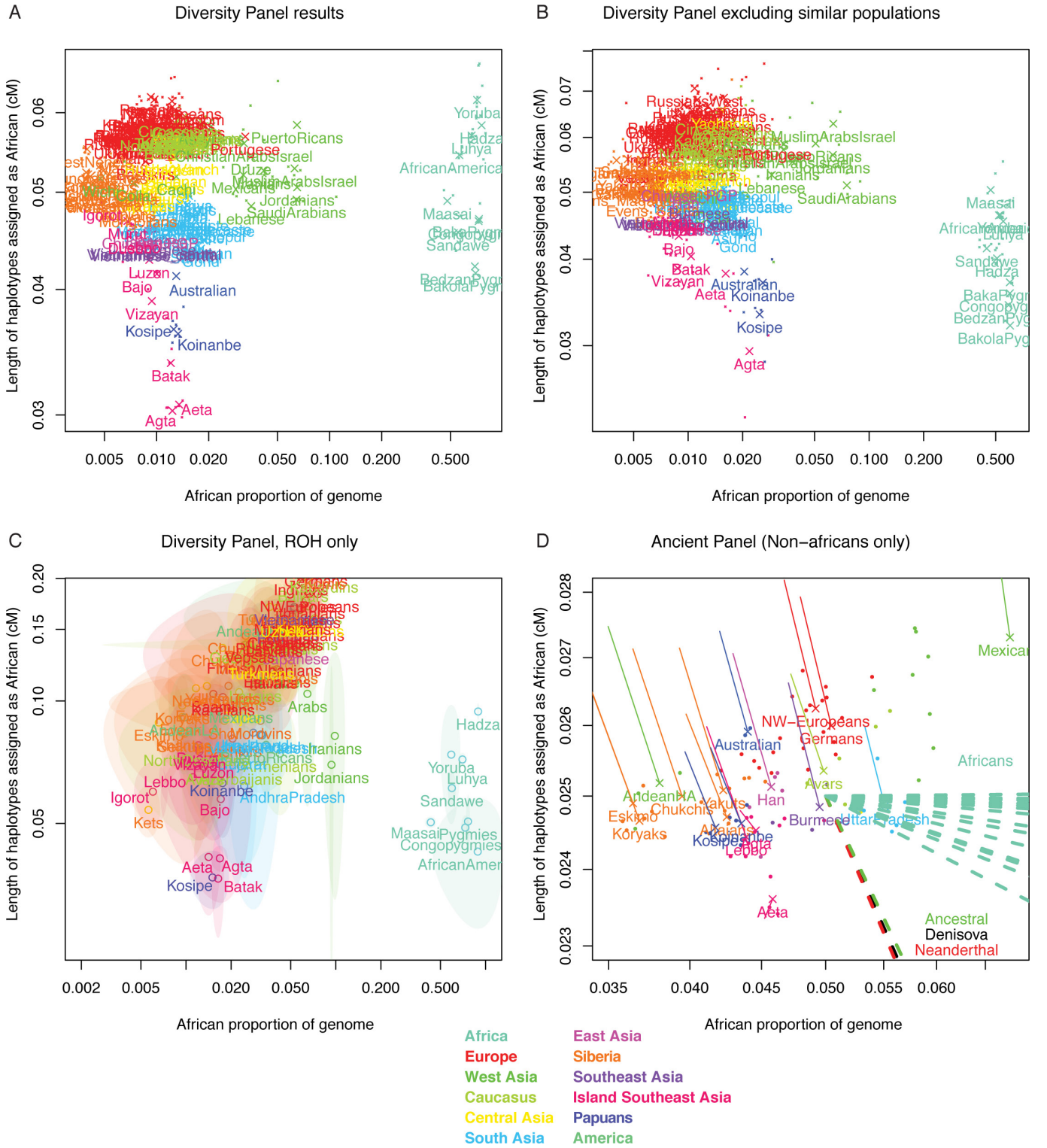
Extended Data Figure 4 | MSMC genetic split times and outgroup f_3 results. a, The MSMC split times estimated between each sample and a reference panel of nine genomes were linearly interpolated to infer the broader square matrix. **b**, **c**, Summary of outgroup f_3 statistics for each pair of non-African populations or an ancient sample using Yoruba as an outgroup. Populations are grouped by geographic region and are ordered with increasing distance from Africa (left to right for columns and bottom to top for rows). Colour bars at the left and top of the heat map indicate the colour coding used for the geographical region. Individual population labels are indicated at the right and bottom of the heat map. The f_3 statistics are scaled to lie between 0 and 1, with a black colour indicating those close to 0 and a red colour indicating those close to 1. Let m and M be the minimum and maximum f_3 values within a given row (that is, focal population). That is, for focal population X (on rows), $m = \min_{Y, Y \neq X} f_3(X, Y; \text{Yoruba})$ and $M = \max_{Y, Y \neq X} f_3(X, Y; \text{Yoruba})$. The scaled f_3 statistic for a given cell in that row is given by $f_{3\text{scaled}} = (f_3 - m)/(M - m)$, so that the smallest f_3 in the row has value $f_{3\text{scaled}} = 0$ (black) and the largest has value $f_{3\text{scaled}} = 1$ (red). By default, the diagonal has value $f_{3\text{scaled}} = 1$ (red). The heat map is therefore asymmetric, with the population closest to the focal population at a given row having value $f_{3\text{scaled}} = 1$ (red colour) and the population farthest from the focal population at a given row having value $f_{3\text{scaled}} = 0$ (black colour). Therefore, at a given row, scanning the columns of the heat map reveals the populations with the most shared ancestry with the focal population of that row in the heat map.



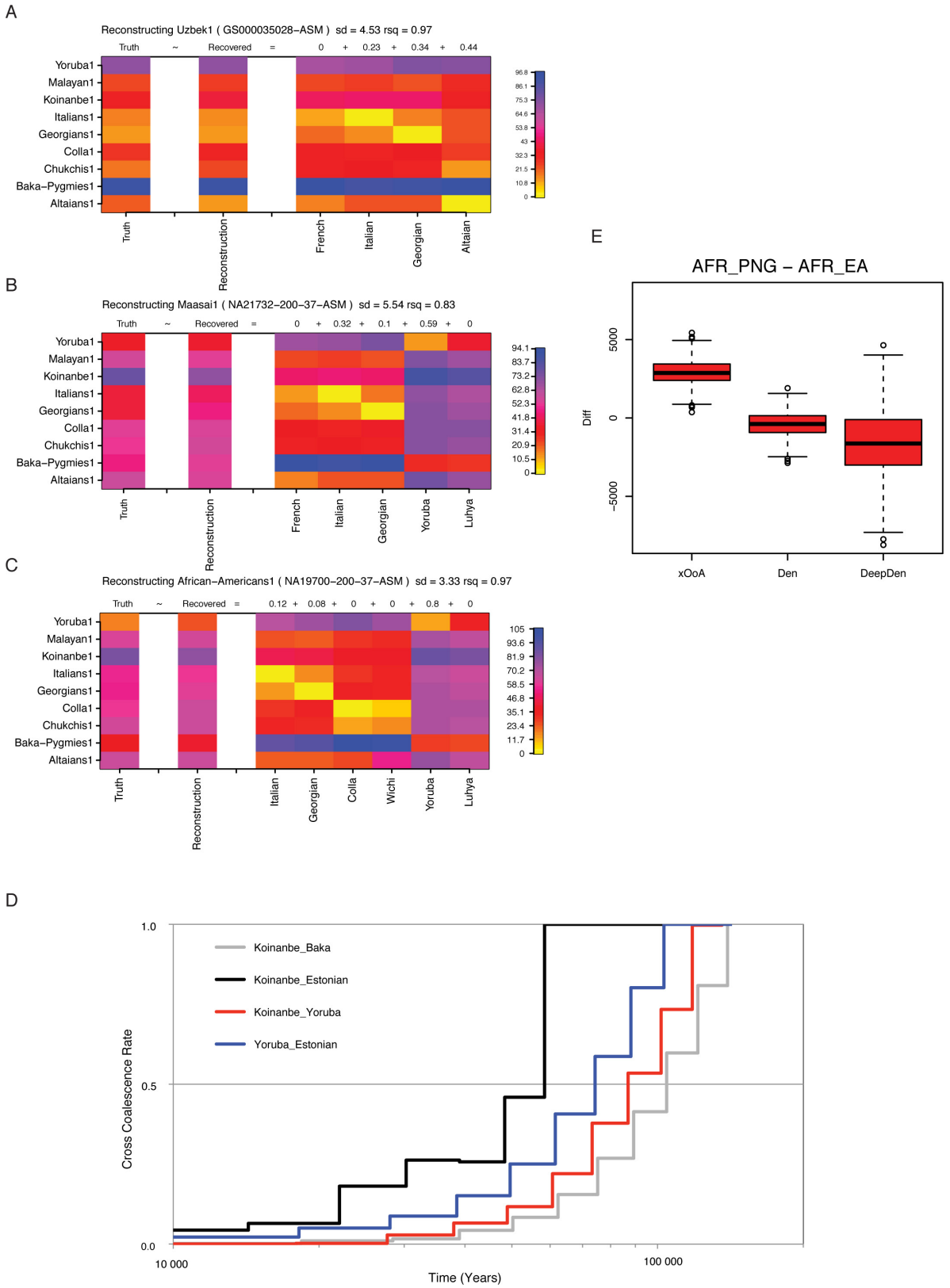
Extended Data Figure 5 | Geographical patterns of genetic diversity. Isolation by distance pattern across areas of high genetic gradient, using Europe as a baseline. The samples used in each analysis are indicated by coloured lines on the maps to the right of each plot. **a**–**d**, The panels show F_{ST} as a function of distance across the Himalayas (**a**), the Ural mountains (**b**), and the Caucasus (**c**) as reported on the colour-coded map (**d**). **e**, Effect of creating gaps in the samples in Europe. **f**, **g**, We tested the effect of removing samples from stripes, either north to south (**f**) or west to east (**g**), to create gaps comparable in size to the gaps in samples in the dataset. **h**, Effective migration surfaces inferred by EEMS.



Extended Data Figure 6 | Summary of positive selection results. a, Bar plot comparing frequency distributions of functional variants in Africans and non-Africans. The distribution of exonic SNPs according to their functional impact (synonymous, missense and nonsense) as a function of allele frequency. Note that the data from both groups was normalized for a sample size of $n = 21$ and that the Africans show significantly (χ^2 $P < 1 \times 10^{-15}$) more rare variants across all sites classes. **b**, Result of 1,000 bootstrap replica of the $R_{X/Y}$ test for a subset of pigmentation genes highlighted by Genome Wide Association Studies (GWAS, $n = 32$). The horizontal line provides the African reference ($x = 1$) against which all other groups are compared. The blue and red marks show the 95th and the 5th percentile of the bootstrap distributions respectively. If the 95th percentile is below 1, then the population shows a significant excess of missense variants in the pigmentation subset relative to the Africans. Note that this is the case for all non-Africans except the Oceanians. **c**, Pools of individuals for selection scans. fineSTRUCTURE-based co-ancestry matrix was used to define twelve groups of populations for the downstream selection scans. These groups are highlighted in the plot by boxes with broken line edges. The number of individuals in each group is reported in Supplementary Table 1:3.2-I.



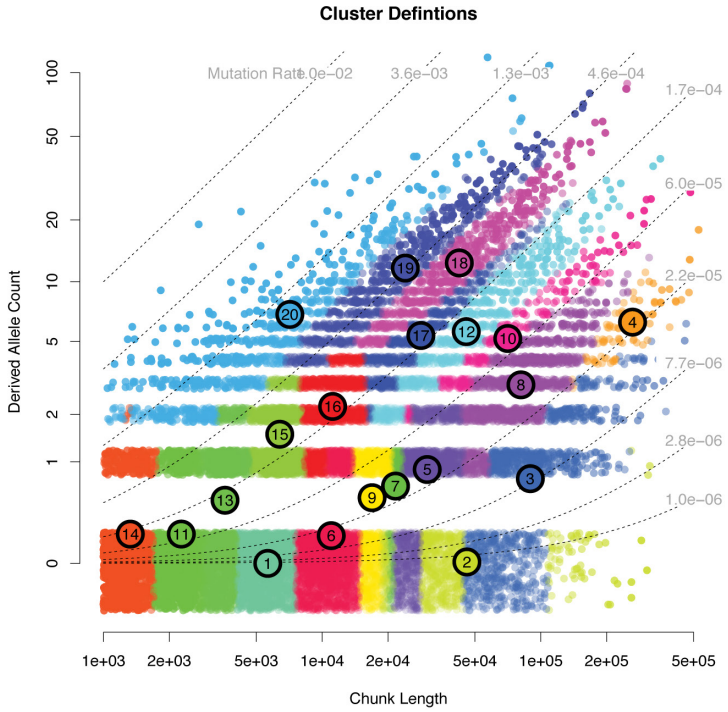
Extended Data Figure 7 | Length of haplotypes assigned as African by fineSTRUCTURE as a function of genome proportion. a, 447 Diversity Panel results, showing label averages (large crosses) along with individuals (small dots). **b**, Relative excluded Diversity Panel results, to check for whether including related individuals affects African genome fraction. Individuals that shared more than 2% of genome fraction were forbidden from receiving haplotypes from each other, and the painting was re-run on a large subset of the genome (all run of homozygosity (ROH) regions from any individual). **c**, ROH-only African haplotypes. To guard against phasing errors, we analysed only regions for which an individual was in a long (>500 kb) run of homozygosity using the PLINK command '–homozyg-window-kb 500000–homozyg-window-het 0–homozyg-density 10'. Because there are so few such regions, we report only the population average for populations with two or more individuals, as well as the standard error in that estimate. Populations for which the 95% confidence interval passed 0 were also excluded. Note the logarithmic axis. **d**, Ancient DNA panel results. We used a different panel of 109 individuals which included three ancient genomes. We painted chromosomes 11, 21 and 22 and report as crosses the population averages for populations with two or more individuals. The solid thin lines represent the position of each population when modern samples only are analysed. The dashed lines lead off the figure to the position of the ancient hominins and the African samples.



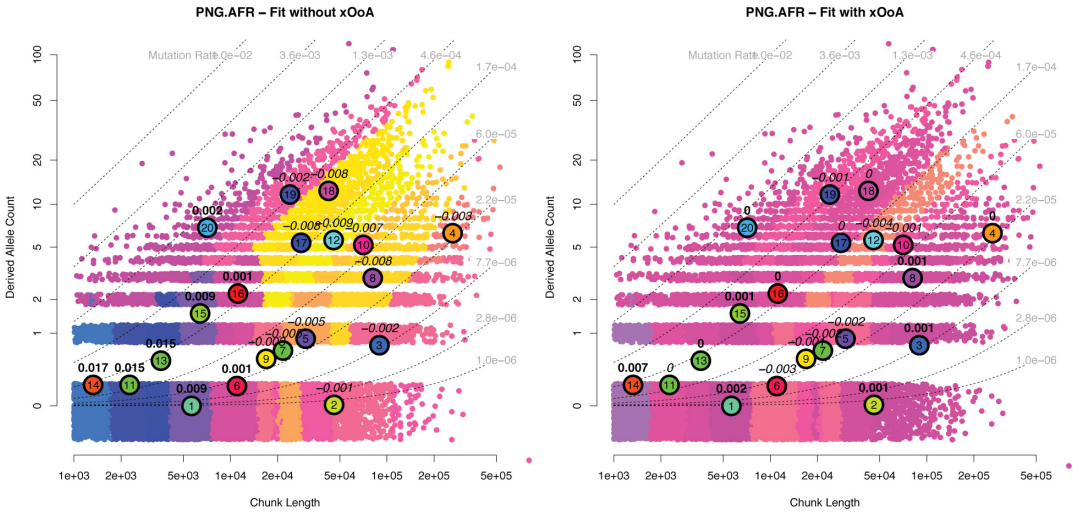
Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | MSMC Linear behaviour of MSMC split estimates in presence of admixture. a–c, The examined Central Asian (**a**), East African (**b**), and African–American (**c**) genomes yielded a signature of MSMC split time (truth, left-most column) that could be recapitulated (reconstruction, second left-most column) as a linear mixture of other MSMC split times. The admixture proportions inferred by our method (top of each admixture component column) were remarkably similar to the ones previously reported from the literature. **d**, MSMC split times calculated after re-phasing an Estonian and a Papuan (Koinanbe) genome together with all the available West African and Pygmy genomes from our dataset to minimize putative phasing artefacts. The cross coalescence rate curves reported here are quantitatively comparable with the ones of Fig. 2a, hence showing that phasing artefacts are unlikely to explain the observed past-ward shift of the Papuan–African split time. **e**, Box plot showing the distribution of differences between African–Papuan and African–Eurasian split times obtained from coalescent simulations assembled through random replacement to make 2,000 sets of 6 individuals (to match the 6 Papuans available from our empirical dataset), each made of 1.5 Gb of sequence. The simulation command line used to generate each chromosome made of 5 Mb was as follows, where x is the variable for the divergence time used. $x = 0.064$, 0.4 or 0.8 for the xOoA, Denisova (Den) and Divergent Denisova (DeepDen) cases, respectively. ms0ancient2 10 1. 065.05 -t 5000. -r 3000. 5000000 -I 7 1 1 1 1 2 2 2 -en 0. 1 .2 -en 0. 2 .2 -en 0. 3 .2 -en 0. 4 .2 -es .025 7.96 -en .025 8.2 -ej.03 7 6 -ej.04 6 5 -ej.060 8 3 -ej.061 4 3 -ej.062 2 1 -ej.063 3 1 -ej *x* 1 5.

A



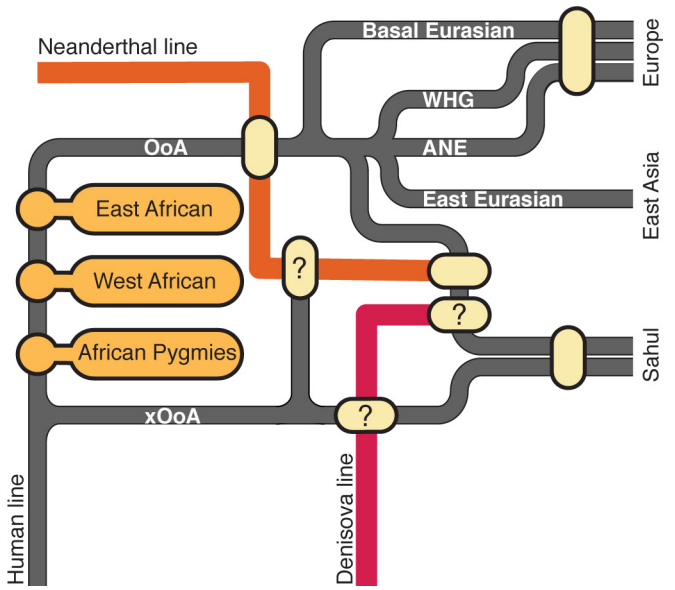
B



C

	PNG.DEN	EUR:DEN	PNG:AFR	EUR:AFR	xOoA
Average Length	19518	17554	28499	32824	17783
Non-African Derived allele mutation rate	0.000202	0.000142	0.000061	0.000049	0.000074
Relative age to OoA (mutational clock)	4.11	2.88	1.24	1	1.51

Extended Data Figure 9 | Modelling the xOoA components with FineSTRUCTURE. a, Joint distribution of haplotype lengths and derived allele count, showing the median position of each cluster and all haplotypes assigned to it in the maximum a posteriori (MAP) estimate. Note that although a different proportion of points is assigned to each in the MAP, the total posterior is very close to $1/K$ for all. The dashed lines show a constant mutation rate. Haplotypes are ordered by mutation rate from low to high. **b**, Residual distribution comparison between the two-component mixture using EUR.AFR and EUR.PNG (left), and the three-component mixture including xOoA (using the same colour scale) (right). The root mean square error (RMSE) residuals without xOoA are larger (RMSE = 0.0055 compared to RMSE = 0.0018) but more importantly, they are also structured. **c**, Assuming a mutational clock and a correct assignment of haplotypes, we can estimate the relative age of the splits from the number of derived alleles observed on the haplotypes. This leads to an estimate of 1.5 times older for xOoA compared to the Eurasian–Africa split.



Extended Data Figure 10 | Proposed xOoA model. A schematic illustrating, as suggested by the results presented here, a model of an early, extinct Out-of-Africa (xOoA) signature in the genomes of Sahul populations at their arrival in the region. Given the overall small genomic contribution of this event to the genomes of modern Sahul individuals, we could not determine whether the documented Denisova admixture (question marks) and putative multiple Neanderthal admixtures took place along this extinct OoA. We also speculate (question mark) people who migrated along the xOoA route may have left a trace in the genomes of the Altai Neanderthal as reported by Kuhlwilm and colleagues[12].