Genomic analyses inform on migration events during the peopling of Eurasia

Geographical barriers, environmental challenges, and complex migration events during the peopling of Eurasia

Authors List

Luca Pagani[1,2*†], Daniel John Lawson[3*], Evelyn Jagoda[1,4*], Alexander Mörseburg[1*], Anders Eriksson[5,6*],Mario Mitt[7,8], Florian Clemente[1,9], Georgi Hudjashov[10,11,12], Michael DeGiorgio[13], , Lauri Saag[10], Jeffrey D. Wall[14], Alexia Cardona[1,15], Reedik Mägi[7], Melissa A. Wilson Sayres[16,17], Sarah Kaewert[1], Charlotte Inchley[1], Christiana L. Scheib[1], Mari Järve[10], Monika Karmin[10,18], Guy S. Jacobs[19,20], Tiago Antao[21], Florin Mircea Iliescu[1], Alena Kushniarevich[10,22], Qasim Ayub[23], Chris Tyler-Smith[23], Yali Xue[23], Bayazit Yunusbayev[10,24], Kristiina Tambets[10], Chandana Basu Mallick[10], Lehti Saag[18], Elvira Pocheshkhova[25], George Andriadze[26], Craig Muller[27], Michael C. Westaway[28], David 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], Harilanto Razafindrazaka[44], Herawati Sudoyo[45], Thierry Letellier[44], Murray P. Cox[12], Nikolay A. Barashkov[46,47], Vedrana Skaro[48,49], Lejla Mulahasanovic[50], Dragan Primorac[51,52,53,49], Hovhannes Sahakyan[10,54], Maru Mormina[55], Christina A. Eichstaedt[1,56], Daria V. Lichman[39,57], Syafiq Abdullah[58], Gyaneshwer Chaubey[10], Joseph T. S. Wee [59], Evelin Mihailov[7], Alexandra Karunas[24,60], Sergei Litvinov[24,60,10], Rita Khusainova[24,60], Natalya Ekomasova[60], Vita Akhmetova[24], Irina Khidiyatova[24,60], Damir Marjanovic´[61,62], Levon Yepiskoposyan[54], Doron M. Behar[10], Elena Balanovska[63], Andres Metspalu[7], Miroslava Derenko[64], Boris Malyarchuk[64], Mikhail Voevoda[65,39,57], Sardana A. Fedorova[47,46], 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[72,63], Elza K. Khusnutdinova[24,60], Ene Metspalu[10,18], Mark G. Thomas[67], Andrea Manica[6], Rasmus Nielsen[73], Richard Villems[10,18,74*], Eske Willerslev[27*], Toomas Kivisild[1,10*†], Mait Metspalu[10,18*†]

*These authors contributed equally to this work.

†Corresponding authors: L.P. (lp.lucapagani@gmail.com), T.K. (tk331@cam.ac.uk), M.M. (mait@ebc.ee)

Author Affiliations

1: Department of Biological Anthropology, University of Cambridge, Cambridge, United Kingdom

2: Department of Biological, Geological and Environmental Sciences, University of Bologna, Via Selmi 3, 40126, Bologna, Italy

3: Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol BS8 2BN, UK.

4: Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

5: Integrative Systems Biology Lab, Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia

6: Department of Zoology, University of Cambridge, Cambridge, UK

7: Estonian Genome Center, University of Tartu, Tartu, Estonia

8: Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

9: Institut de Biologie Computationnelle, Université Montpellier 2, Montpellier, France

10: Estonian Biocentre, Tartu, Estonia

11: Department of Psychology, University of Auckland, Auckland, 1142, New Zealand;

12: Statistics and Bioinformatics Group, Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

13: Department of Biology, Pennsylvania State University, University Park, PA, 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

16: School of Life Sciences, Tempe, AZ, 85287 USA

17: Center for Evolution and Medicine, The Biodesign Institute, Tempe, AZ, 85287 USA

18: Department of Evolutionary Biology, Institute of Molecular and Cell Biology, University of Tartu, 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, MT, USA

22: Institute of Genetics and Cytology, National Academy of Sciences, Minsk, Belarus

23: The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom

24: Institute of Biochemistry and Genetics, Ufa Scientific Center of RAS, Ufa , Russia

25: Kuban State Medical University, Krasnodar, Russia

26: Scientific-Research Center of the Caucasian Ethnic Groups, St. Andrews Georgian University, Georgia

27: Center for GeoGenetics, University of Copenhagen, Denmark

28: Research Centre for Human Evolution, Environmental Futures Research Institute, Griffith University, Nathan, Australia

29: Center of Molecular Diagnosis and Genetic Research, University Hospital of Obstetrics and Gynecology, Tirana, Albania

30: Center of High Technology, Academy of Sciences, Republic of Uzbekistan

31: Institute of Bioorganic Chemistry Academy of Science, Republic of Uzbekistan

32: L.N. Gumilyov Eurasian National University, 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, PA, 19104-6145, USA

35: School of Biology, Georgia Institute of Technology, Atlanta, Georgia, USA

36: Departments of Genetics and Biology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

37: DNcode laboratories, Moscow, Russia

38: Institute of Molecular Biology and Medicine, Bishkek, Kyrgyz Republic

39: Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

40: Mongolian Academy of Medical Sciences, Ulaanbaatar, Mongolia

41: Northern State Medical University, Arkhangelsk, Russia

42: Anthony Nolan, London, United Kingdom

43: V. N. Karazin Kharkiv National University, 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, France

45: Genome Diversity and Diseases Laboratory, Eijkman Institute for Molecular Biology, Jakarta, Indonesia

46: Department of Molecular Genetics, Yakut Scientific Centre of Complex Medical Problems, Yakutsk, Russia

47: Laboratory of Molecular Biology, Institute of Natural Sciences, M.K. Ammosov North-Eastern Federal University, Yakutsk, Russia

48: Genos, DNA laboratory, Zagreb, Croatia

49: University of Osijek, Medical School, Osijek, Croatia

50: Center for Genomics and Transcriptomics, CeGaT, GmbH, Tübingen, Germany

51: St. Catherine Speciality Hospital, Zabok, Croatia

52: Eberly College of Science, The Pennsylvania State University, University Park, PA, USA

53: University of Split, Medical School, 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: Thoraxclinic at the University Hospital Heidelberg, Heidelberg, Germany

57: Novosibirsk State University, Novosibirsk, Russia.

58: RIPAS Hospital, Bandar Seri Begawan, Brunei Darussalam

59: National Cancer Centre Singapore, Singapore

60: Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa , Russia

61: Department of Genetics and Bioengineering. Faculty of Engineering and Information Technologies, International Burch University, Sarajevo, Bosnia and Herzegovina

62: Institute for Anthropological Researches, 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, Magadan, Russia

65: Institute of Internal Medicine, Siberian Branch of Russian Academy of Medical Sciences, Novosibirsk, Russia

66: Leverhulme Centre for Human Evolutionary Studies, Department of Archaeology and Anthropology, University of Cambridge, Cambridge, United Kingdom

67: Research Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

68: Department of Archaeology, University of Papua New Guinea, University PO Box 320, NCD, Papua New Guinea

69: College of Arts, Society and Education, James Cook University, PO Box 6811, Cairns QLD 4870, Australia

70: Department of Anthropology, University College London, London, United Kingdom

71: School of Archaeology, University of Oxford, 75 George Street, Oxford, OX1 2BQ, UK

72: Vavilov Institute for General Genetics, Russian Academy of Sciences, Moscow, Russia

73: Department of Integrative Biology, University of California Berkeley, Berkeley 94720, CA, USA

74: Estonian Academy of Sciences, 6 Kohtu Street, Tallinn 10130, Estonia

Abstract

Previous human genetic studies, based on sampling small numbers of populations, have supported a recent Out-of-Africa dispersal model with minor additional input from archaic humans. Here, we present a novel dataset of 379 high-coverage human genomes from 125 populations worldwide. The combination of high spatial and genomic coverage enabled us to refine current knowledge of continent-wide patterns of heterozygosity, long- and short-distance gene flow, archaic admixture, and changes in effective population size. Compared to Eurasians, the examined Papuan genomes show an excess of highly derived modern human haplotypes and deeper split times from Africans. This is compatible with an early and largely extinct expansion of modern humans Out-of-Africa. This is also indicated by the Western Asian fossil record and the recent discovery of modern human and Neanderthal admixture 100,000 years ago, which significantly predates the main Out-of-Africa expansion of modern humans. Our tests of positive and balancing selection highlight a number of new metabolism- and immunity-related loci as candidates for local adaptation.

Introduction

Previous genome-wide sequencing efforts have aimed at characterizing common variants in the human genome by targeting moderate numbers of geographically distinct populations and combining genotyping, low-coverage whole-genome and exome sequencing data[1,2]. High-coverage whole-genome sequence studies have so far been limited to focusing on specific populations[3] and geographic regions[4-7], or targeted at specific diseases, e.g. cancer[8]. Nevertheless, the availability of high-resolution genomic data has led to the development of new methodologies for inferring population history[9-13] and refuelled the debate on the mutation rate in humans[14]. From these initial studies, the unprecedented potential of high-coverage genomic data to reveal geographically specific patterns of genetic diversity has become evident. Here, we present a new dataset of high-coverage human genomes from nearly 150 populations distributed worldwide. This comprehensive population sample, which, among others, includes new samples from Siberia, Island Southeast Asia and Papua New Guinea, allows us to infer human demographic history in finer detail and to investigate signatures of natural selection. We estimate split times among populations, test how the different populations conform to the model of a single expansion out of Africa with archaic admixture (OoA), and assess patterns of neutral and adaptive variation associated with different environments.

Data description. Our worldwide panel of 483 high-coverage human genomes from 148 populations includes 379 new genomes from 125 populations (Figure 1) (Table S1.7-I). All genomes were sequenced by Complete Genomics Inc. and mapped, called and phased using the same bioinformatic pipeline, thereby minimizing platform and processing bias conflicts (Supplementary Section 1.1). We maximised the number of groups in this study by limiting the number of individuals to three for most populations. Existing SNP-chip information was used in most cases to choose unrelated individuals and to avoid cases of recent admixture between geographically distant populations. For demographic inferences, we combined previously published and new sequences to generate a geographically balanced sample (Figure 1, Diversity Set, N=447). For selection scan analysis, we focussed on well-covered geographic regions,

combining a subset of the Diversity Set with published sequences (Figure 1, Selection Set, N = 396, Supplementary Section 1.7).

The current view on the peopling of Eurasia. The timing and route of human movements out of Africa, as well as the degree to which migrating populations interbred with archaic humans during their expansion across Eurasia, have been the subject of considerable debate over the past two decades[15]. Fossil evidence demonstrates that *Homo sapiens* was present in Levant between ca. 120-70 kya[16]. This colonization has, however, been viewed as a failed expansion OoA[17]. Nevertheless craniometrical studies of African and Asian populations[18] and fossil data from eastern Asia[15], including the very recent reports of human remains in China from before 80 kya[19], admit the possibility of an early dispersal. Moreover, archaeological finds in Arabia and South Asia indicate the presence of human populations in ameliorated environments between 125 and 75 kya[15]. Previous genetic analyses of living populations have revealed a steady decline in genetic diversity with distance from Africa, which is consistent with a serial founder event model[20-22].

Ancient DNA (aDNA) sequencing has further contributed to our understanding of the peopling of Eurasia and revealed admixture with at least two archaic human lineages. Neanderthals have left a genetic signature in all non-Africans from around 55 kya[23], while admixture with Denisova was largely restricted to the ancestors of modern Papuan and Australian populations[24]. In addition aDNA from modern humans indicates population structuring and turnover, but little additional archaic admixture, in Eurasia over the last 35-45 thousand years[25-27]. Overall, these findings provide support for a model[28,29] by which the vast majority of human genetic diversity outside Africa derives from a single dispersal event that was followed by admixture with archaic humans[23,29].

Results

Population structure in Eurasia. We used ADMIXTURE [30] to infer genetic structure and admixture patterns in our Diversity Set (Figure 1 for K=8 and K=14, Supplementary Sections 2.1.1-2 for Ks=2-14). Western Eurasia is characterised by two predominant genetic clusters, whilst the much less

populous Siberia shows evidence of three differentiated clusters (Figure 1, K=14), consistent with previous reports[31]. Island Southeast Asia also exhibits high population structuring. Both these latter two regions have histories of small effective population densities (Figure S2.2.3-I, as inferred by MSMC[10]), which increase genetic drift and local differentiation.

We compared the haplotype similarity of our samples using fineSTRUCTURE[32]. This shows that our sampling strategy retains the power to identify population structure at fine resolution. We inferred 106 genetically distinct populations forming 12 major regional clusters, corresponding well to the 148 self-identified population labels. This clustering is based on an individual level measure of haplotype similarity, which is sensitive to small and recent genetic differentiation, and forms the basis for the groupings used in the scans of natural selection.

The importance of geography. The dense geographic coverage of our samples allowed us to investigate the importance of geographic barriers in shaping gene flow. We did so by interpolating genetic variation spatially, focussing on measures of pairwise similarity between genomes in pairs of populations (Supplementary Section 2.2.2). We considered several similarity measures (Supplementary Section 2.2.2) and report gradients of allele frequencies in Figure 2. We validated the approach using isolation by distance patterns across major gradients and migration surfaces reconstructed using EEMS[33]. The main features are the East-West Eurasian split near the Ural Mountains, and the Tibetan plateau, as expected. To formally link these patterns to geographic features, we quantified the effects of elevation, temperature, and precipitation on genetic gradients while controlling for pairwise geographic distances (Supplementary section 2.2.2). This analysis identifies precipitation and elevation as environmental variables that correlate most strongly with the genetic gradients estimated from allele frequencies (inset of Figure 2).

Differentiation in Eurasia after the expansion out of Africa. We observe the well-documented decrease in the number of heterozygous sites per genome as a function of distance from East Africa (Figure 1); a pattern consistent with a model of serial founder events during the peopling of Eurasia[20,21].

While this pattern is relatively smooth, there are a number of discontinuities that potentially highlight geographic regions that acted as barriers during the expansion. Such discontinuities can be visualised by plotting the outgroup f_3 statistic[13,34] in the form $f_3(X, Y;$ Yoruba), which here measures shared drift between non-African populations X and Y from Yoruba as an African outgroup (Supplementary Section 2.2.6, Figures S2.2.6-I-II). We tested all possible combinations of X and Y within our Diversity Set and 25 published aDNA genomes. While recapitulating the main groupings inferred by ADMIXTURE and fineSTRUCTURE, the outgroup f_3 statistic also flags populations that have experienced additional drift. For example, the f_3 values are similar for comparisons within Caucasus populations and between populations from Europe and Caucasus. The f_3 values for comparisons within Europe, however, are significantly higher. These findings are consistent with a simple model of population splits within the Caucasus dating to approximately the same time as the split between European and Caucasus populations[35].

An excess of old haplotypes in Sahul. Our fineSTRUCTURE analysis highlights an excess of shorter African haplotypes in Papuans, as well as Philippine Negritos, compared to all other non-African populations. This pattern remains after correcting for potential confounders such as phasing errors and sampling bias (Figure S2.2.1-VII, Supplementary Section 2.2.1). A natural interpretation from population genetics theory is that these shorter shared haplotypes reflect an older population split[36].

We further investigated whether Sahul populations differ from other Eurasian populations by estimating population splits using MSMC[10]. We focussed on 23 populations (Supplementary Figure 2.2.3-II), chosen to represent major genetic groups (Supplementary Section 2.2.3) and used a novel method to predict all pairwise split times (Methods, Supplementary Figure 2.2.3-III). The split of all mainland Eurasian populations from Yorubans consistently appears as a gradual process with a median time ~75 kya (Table S2.2.3-I, Figure 3A). Importantly, Papuans are an exception to this broad picture, showing a deeper median split time from Yoruba at around 90 kya; a conclusion robust to phasing artefacts (See Methods). The Papuan-Eurasian MSMC split time of ~40 kya is slightly older (Figure S2.2.3-III) than splits between West Eurasian and East Asian populations

(~30 kya). The Papuan split times from Yoruban and Eurasian are incompatible with a simple bifurcating population tree model, implying that modern Papuan individuals are admixed between different topologies. Some of their genome is an outgroup to most modern Africans and Eurasians, while the rest of their genome shares a history with Eurasia.

Ancient or modern introgression in Sahul? At least two main models could account for Sahul populations having older split dates from Africa than mainland Eurasians in our sample:

a) Admixture in Sahul with an archaic human population that split from modern humans either before or at the same time as did Denisova and Neanderthal. This introgressing population could potentially have diverged from the available aDNA samples more than 350 kya.

b) Admixture in Sahul with a modern human population (xOoA) that left Africa well after the split between modern humans and Neanderthals, but before the main expansion of modern humans in Eurasia (main OoA).

We performed a large number of tests to distinguish these scenarios. Because the introgressing lineage has not been observed with aDNA, standard methods are limited in their ability to distinguish between these hypotheses. Our approach therefore relies on building multiple lines of evidence using haplotype-based MSMC and fineSTRUCTURE comparisons. The two hypotheses are not mutually exclusive and we can only hope to identify the source of the strongest contribution.

Single site statistics cannot identify the source of introgression. We first tried traditional statistical approaches, most notably Patterson's *D* statistic[13,23], which we applied to all possible tree relationships between our samples from Africa, Sahul and Eurasia (Figure S2.2.7-I). The best-supported topology among those tested shows a contribution to the Sahul genome from a population (xOoA) that diverged early from West Africans, Baka and Mbuti. This predates the separation of the ancestors of the modern Africans and Eurasians in our dataset (topology 3 in Figure S2.2.7-I) as previously proposed[37]. However, when including the documented Denisova admixture into the analysis[38] and allowing Denisova introgressed segments to have strongly (350 kya) diverged from the observed Denisova genome, the *D*-based test could not discriminate

between a putative xOoA and the Denisova genomic components (Supplementary Section S2.2.7).

We also counted non-African Alleles (nAAs), i.e. derived alleles present outside Africa, but absent in Africans and also archaic humans (Altai Neanderthal and Denisova genomes) (Figure S2.2.7-II). When compared to Eurasians, both Sahul, including two admixed Australian Aborigine genomes, and Philippine Negrito samples do show an excess of nAAs. This is independent of potential demographic confounders, such as inbreeding or drift (Figure S2.2.7-III). Again, the excess of nAAs could be explained by admixture with xOoA, which had more time to accumulate such alleles. However, simulations show that, when allowing sufficient within-Denisova divergence time, archaic introgression could generate the same pattern. In this case, we fail to fully mask the derived alleles in Papuans originating from the introgressing Denisova by relying only on a single Denisova sample (Figure S2.2.7-IV). Our D-based and nAAs results and related simulations show empirically that these kind of single site statistics lack the power to discriminate between the hypothesised scenarios: either Denisova introgression or a xOoA scenario would result in an increase of non-African derived alleles in Papuans. The extent of such increase, at the genome-wide level, is a function of the admixture proportion and divergence time of the introgressing population from the main human lineage. Therefore, two admixture events with unknown proportions and time depth are equally able to explain the data and cannot be disentangled by single site statistics alone.

Haplotype-based analyses indicate an early modern human expansion signature in Sahul. Using a previously published method[39], we located and masked putatively introgressed Denisova haplotypes from the genomes of Papuans. We also tried symmetrically phasing Papuans and Eurasians (see Methods) to evaluate the contribution of phasing errors to the observed shift in MSMC split dates. Neither modification (Figure 3A, Supplementary Section 2.2.9, Table S2.2.9-I) changed the estimated split time (based on MSMC) between Africans and Papuans, suggesting that Denisova admixture or phasing artefacts are not the main driver of this pattern (See Methods, Supplementary Section 2.2.8, Figure S.2.2.8-I, Table 2.2.8-I). We further tested the possible role of Denisova admixture by extensive coalescent

simulations (Figures S2.2.8-I-II). Without assuming an implausibly large contribution from a Denisova-like population, we could not simulate the large Papuan-African and Papuan-Eurasian split times inferred from the data. Assuming that MSMC dates behave linearly under admixture, the results also indicate that the hypothesised xOoA lineage may have split from most Africans as early as 120 kya. This assumption is validated in Supplementary Section 2.2.4 by checking that split dates behave as a mixture in known admixture events. However, for very old divergences the linearity does not hold true as we demonstrate in Supplementary Section 2.2.8. Here we show with additional simulations that 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, but that a similar quantity of Denisova admixture does not produce any significant effect (Figure S2.2.8-III). Together with the previous simulations, 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 further tested our hypothesised xOoA model by focussing on genomic regions in Papuans that have African ancestry not found in other Eurasian populations. We reran fineSTRUCTURE on an "ancient diversity panel", a subset of the Diversity Set with the addition of the Denisova, Altai Neanderthal and the Human Ancestral Genome sequences[2], with sites that are heterozygous in archaic humans removed. FineSTRUCTURE infers chunks of the genome that have a single inferred most recent common ancestor (MRCA). An MRCA between different populations occurs either because the lineage first coalesces before two populations split, or because of a more recent introgression event. Papuan genomic chunks that have an African MRCA assignment in the sample, like the genome-wide nAAs results above, had an elevated level of non-African derived alleles compared to such chunks in Eurasians. They therefore have an older mean coalescence time with our African samples, as would be expected if Papuans contained genetic contributions from a xOoA lineage.

On the other hand there may also be a deep divergence between the sampled Denisova and the one introgressing into modern humans. We were hence concerned that some introgressed archaic haplotypes have an MRCA with

Africans due to coalescence in the ancestral population, and hence are assigned to be African. However, we can resolve the age and hence origin of these chunks by their sequence similarity with modern Africans. To account for the archaic introgression we modelled these genomic portions as a mixture of chunks assigned African or Denisova in Eurasians, as well as chunks assigned Denisova in Papuans. Chunks are modelled (see Methods) in terms of the distribution of length and mutation rate, which is characterised in terms of the density of non-African derived alleles, which are nAAs that are fixed ancestral in our Africans.

This approach captures lineages that coalesce before the human/Denisova split since the properties of these chunks should not depend on the population they were found in, and since Eurasians (specifically Europeans) have not experienced Denisova admixture. By this way we could disentangle the various introgressing lineages by looking at their mutation density. From the discrepancy between the distribution of Papuan chunks assigned to Africans and the fitted distribution (Figure 3B-D) we can identify the characteristics of xOoA chunks (Supplementary Section 2.2.10). Including a xOoA component was necessary to account for the number of short chunks with "moderate" mutation density, i.e. higher than Eurasian chunks assigned African but significantly lower than those assigned Denisova in either Eurasians or Papuans. Inferred xOoA chunks have 1.5 times more nAAs than that observed in chunks assigned to be Eurasian, compared to 4 times for chunks assigned to be Denisova. These proportions can be interpreted as a relative mean time to the most recent common ancestor, implying a xOoA-Africa split 1.5 times older than the main OoA, consistent with our MSMC findings (Supplementary Section 2.2.4).

We went on to estimate the proportion of xOoA in Papuan chunks assigned as both Eurasian (0.1%, 95% CI 0-2.6) and Papuan (4%, 95% CI 2.9-4.5) (Supplementary Section 2.2.10), by using the estimated mutation density in xOoA. To do this we used the same mixture model as above (additionally considering Eurasian chunks assigned to be Eurasian) to obtain a xOoA-free prediction. When this predicted too few mutations, we assumed that the difference is due to the xOoA admixture. Adding up the contributions from all assignments of chunks leads to a genome-wide estimate of 1.9% xOoA (95% CI 1.5-3.3) in Papuans.

Our results consistently point towards a predominantly modern human source for the abundance of alleles found in Papuans that are absent in Africans and are derived according to the ancestral human sequence. It follows that the genome of modern Papuans is best described as consisting of two human components. The predominant component is an early split from the major migration out of Africa that colonized Eurasia while the lesser component is derived from an earlier, otherwise extinct, dispersal.

Adaptation outside Africa

Humans faced a number of ecological challenges as they encountered new environments outside Africa. To study the nature and extent of any resultant adaptation, we explored the distribution of functional variants among populations, performed tests of purifying, balancing and positive selection and, finally, identified loci that showed the highest allelic differentiation among groups (Supplementary Section 3). It is important to emphasise that our sampling strategy may be underpowered to detect certain types of selection. Despite this, strong signals are present in the data.

Relationship to other findings. The results of our positive selection tests corroborated the identification of a number of selective sweeps that are well supported by functional evidence (Table S3.3.4-I), suggesting that, regardless of our sample pooling strategy, our dataset is able to detect region-specific signals of haplotype homozygosity and allelic differentiation. Our tests for purifying selection are also consistent with previous studies[2,40,41], in terms of both the lack of differential purifying selection between Africans and non-Africans, as well as the distribution of alleles across frequency classes and populations (Supplementary Section 3.1, Figure S3.1-I,II; Table S3.1-IV,VI).

Novel findings. Our results show novel signals of purifying, balancing and positive selection. With regard to purifying selection, we report evidence for significant differences in the strength of selection in systematically defined phenotype-related sets of genes. We infer more purifying selection in Africans in genes involved in pigmentation (bootstrapping p value for $R_{X/Y}$-scores < 0.05) (Figure S3.5-II) and immune response against viruses ($p < 0.05$), whilst more

purifying selection was indicated on olfactory receptor genes in Asians ($p < 0.01$ in the Southeast Asia Island population, $p < 0.05$ in the Southeast Asia Mainland, South American and Northeast Siberia populations) (Table S3.1.1-II). A genome-wide scan for ancient balancing selection in populations grouped into 12 geographical regions according to their genetic clustering (Supplementary Section 3.2) revealed a significant enrichment (false discovery rate q-value < 0.01) for antigen processing/presentation, antigen binding, and MHC and membrane component genes (Tables S3.3.2-I-III). The HLA (*HLA-C*)-associated gene (*BTNL2*) was the top candidate in eight of 12 geographic regions (Table S3.3.1-I).

Our positive selection scans and variant-based analyses (Supplementary Sections 3.2 and 3.2) revealed many novel signals, especially in the less-studied populations, a subset of which is highlighted in Table 1. Benefiting from the availability of high resolution sequencing information, we were also able to identify new potentially causal variants in both novel and previously-detected positive selection signals.

Given the geographic distribution of our samples, we were particularly interested in assessing whether genes associated with phenotypes highly-correlated to local environmental features, such as temperature, UV exposure, diet, and pathogen load, are systematically overrepresented in the signals of positive selection in the sampled populations (Supplementary Section 3.4; Tables S3.5-I-VI). All categories reported as enriched have chi-square p-values less than 0.01. We observed that genomic regions containing pigmentation-related genes were overrepresented in some of our positive selection tests in West Eurasian populations (Table S3.5-I), while those containing genes relating to thermoregulation were enriched, albeit for different genes, in Africans and Central Siberians (Table S3.5-II). Unlike Khrameeva and colleagues[42], we do not observe an enrichment of fatty acid metabolism (or specifically lipid catabolism) genes in the positive selection tests for our European samples. We do, however, observe enrichment of such genes in Island Southeast Asian and Central Siberian populations (Table S3.5-IV, Figure S3.5-IV).

With regard to immunity, we found enrichment of bacterial immunity genomic windows in Island Southeast Asians (Table S3.5-V), which was lost after the

exclusion of Philippine Negritos from the tests, suggesting that the observation partially reflects elevated selection in these hunter-gatherer groups. Furthermore, both western Asian and the South Asian groups showed significant enrichment in innate immune response annotations based on Tajima's *D* statistic (Table S3.5-VI, Figure S3.5-V), which was the only category that showed any enrichment by that test. This is consistent with selection represented by these signatures being older than those detected by the haplotype homozygosity tests. The fact that most innate immunity signals are shared between at least two populations supports this interpretation.

Discussion

A valuable resource. The collection of worldwide high-coverage genomes presented here has allowed us to: (i) provide a finer resolution description of human genetic diversity; (ii) identify the genetic trace of a so-far unidentified component in Sahul populations; and (iii) increase the number of candidate genome regions that have been subjected to distinct selective pressures on physiological processes. The latter is key to unravelling our adaptation history. The data and inferences presented here provide the groundwork to refine hypotheses about human evolution that are essential to the understanding of modern patterns of genetic diversity, disease vulnerability and distribution.

Methodological difficulties. Existing methods based on single-site analyses seemed unable to resolve our hypotheses about Sahul and could not be used to distinguish between a small fraction of ancient admixture and a larger fraction of more recent admixture. The power of these approaches in practice depends on appropriate ancient samples being available. The behaviour of haplotype-based inference approaches are relatively poorly characterised and there is no formal inferential framework available to address our hypotheses. However, haplotypes preserve more information on our evolution as they can persist for long periods in finite populations[43] at lengths that are detectable with sequence variation data (Supplementary Section 2.2.13). They allow us to calibrate drift by considering the rate of non-African alleles accumulated in

segments of known length, providing us with a way to estimate the age of splits from Africa.

A further confounder is that detecting Denisova and Neanderthal introgression mostly relies on matching to the aDNA data available, which may be a poor proxy for the actual introgressed DNA. Other possible confounders could involve a shorter generation time in Papuan and Philippine Negrito populations[44], 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 explain the haplotype patterns described here.

Evidence for an earlier exit out of Africa? Our estimate of the split between African and Eurasians is in broad agreement with previous reports based on mtDNA and Y chromosome[45-47] and full genome sequencing data[5,10], and is consistent with a major OoA expansion (likely through the Levant[5] and/or Arabia[15]) after that date. Other methods rescaled to the lower mutation rate used here[14] suggest slightly older dates for that split[28,48]. A recent IBS tract sharing-based method[11], when similarly rescaled, yields a remarkably similar split time of ~80 kya.

Our analyses, however, provide clear evidence that the Sahul populations sampled here, and possibly other populations from the region that were not included in our study design, possess an additional genetic signal of introgression from an uncharacterised hominin. We used a series of tests to try to identify whether this hominin came from a) an archaic lineage or b) an earlier out-of-Africa, modern human branch. Current single-site approaches could not distinguish these hypotheses, but our haplotype-based approaches all point towards a small amount of admixture (at least 2%) from an earlier modern human dispersal out-of-Africa around 120 kya (Figure 4) whose genetic signature has not been identified in any other extant population. We also show (see Methods) that this is not at odds with evidence that show that Sahul shares Y chromosome and mtDNA lineages with Eurasians, as there is a high probability that older Y and mtDNA lineages would be lost as a result of random genetic drift, as was also argued by Groucutt and others colleagues[15,49].

The inferred xOoA split time (~120 kya) corresponds with fossil and archaeological evidence for an early expansion of *Homo sapiens* from Africa[15,19]. Furthermore, Kuhlwilm and colleagues[50] recently identified modern human admixture into the Altai Neanderthal before 100 kya. This is consistent with modern human presence outside of Africa well before the main OoA expansion after 75 kya. Further studies will confirm if the xOoA we propose here and the early modern humans that admixed with ancestors of Altai Neanderthals were part of the same early expansion out of Africa. Similarly, we are agnostic to the geographic extent of such an early event. Indeed, archaeological evidence for modern human colonization of Sahul is no earlier than ca. 60-50 kya[51], and perhaps as late as ca. 47 kya[52]. The preponderance of genomic evidence, in fact, indicates that early human expansions did not leave detectable genetic traces in most contemporary Eurasian populations, perhaps as a consequence of substantial population replacements, as indicated by aDNA from Oase, Romania[53]. Climatic changes over the last 120 thousand years, including glacial advances and significant fluctuations of wet and dry environmental cycles, likely influenced population structure across Eurasia[54], perhaps leading to lineage extinctions and regional extirpations. The unexpected genetic traces of xOoA in Papuans, shown here for the first time, suggest that unravelling the evolutionary history of our own species will require the recovery of aDNA from additional fossils, and further archaeological investigations in under-explored regions of Eurasia.

Data availability

The newly sequenced genomes were deposited in the ENA archive under accession number ENAXXXX and are also freely available through the Estonian Biocentre website (www.ebc.ee/free_data).

Acknowledgements

Support was provided by: Australian Research Council (M.W.; D.L.); Danish National Research Foundation; the Lundbeck Foundation and KU2016 (E.W.); ERC Starting Investigator grant (FP7 - 261213) (T.K.); Estonian Institutional

Research grant IUT24-1 (T.K.); Estonian Research Council grant PUT766 (G.C.; M.K.); Estonian Science Foundation grant 8973 (M.M.); EU European Regional Development Fund through the Centre of Excellence in Genomics to Estonian Biocentre; 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.; M.M.); 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. & 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) (G.H.); 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.; E.B); Russian Science Foundation grant 14-14-00827 (O.B.); The Russian Foundation for Basic Research (11-04-00725-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 (D.J.L); Wellcome Trust grant 098051 (Q.A.; C.T.-S.; 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’ (AM); Program of the RAS Presidium "Basic research for the development of the Russian Arctic" (B.M.).

Figure and Table Legends

Table 1 Subset of novel positive selection Findings in our 12 macro-regional groups defined using fineSTRUCTURE.

Figure 1 Panel A: Map of samples location highlighting Diversity/Selection Set; Panel B: ADMIXTURE plot (K=8 and 14) which relates general visual inspection of genetic structure to studied populations and their region of origin; Panel C: 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.

Figure 2 Spatial visualisation 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 under water). Here we used a novel spatial kernel smoothing method based on the matrix of pairwise average heterozygosity. **Inset:** partial correlation between magnitude of genetic gradients and combinations of different geographic factors, elevation (E), temperature (T) and precipitation (P), for genetic gradients from fineSTRUCTURE (red) and allele frequencies (blue). This analysis (see Supplementary Section 2.2.2 for details) shows that despite the large number of prehistoric movements across Eurasia, genetic differences within this region have been strongly shaped by physical barriers such as mountain ranges, deserts, and open water (such as the Wallace line).

Figure 3 Panel 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. Panels B-D: Decomposition of the ChromoPainter inferred African chunks in Papuans. Panel **B**: Semi-parametric decomposition of the joint distribution of chunk lengths and non-African derived allele rate per SNP, showing the relative proportion of chunks in K=20 components of the distribution, ordered by non-African derived allele rate, relative to the overall proportion of chunks in each component. The four datasets produced by considering (African/Denisova) chunks in (Europeans/Papuans) are shown with our inferred "extra Out-of-Africa xOoA" component. Panel **C**: The reconstruction of African chunks in Papuans using a mixture of the other data (red) and adding the xOoA component (black). Panel **D**: The properties of the components in terms of non-African derived allele rate, on which the components are ordered, and length.

Figure 4 A subway map figure illustrating, as suggested by the novel results presented here, the model of an early, extinct Out-of-Africa (xOoA) entering the genome of Sahul populations at their arrival in the region. Given the overall small genomic contribution of this event to the genome of modern Sahul, we could not determine whether the documented Denisova admixture (question marks) and putative multiple Neanderthal admixtures took place along this extinct OoA.

Methods

Data Preparation: In the final dataset, we retained only one second (Australians, to make use of all the available samples)- and five third-degree relatives pairs (Table S1.7-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[2] (Supplementary Sections 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, South-East 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. This pattern is independent of simple potential confounders, such as rough estimates of recombination activity and gene density (Supplementary Section 1.8), and mirrors the inter-chromosomal differences in divergence from chimpanzee[55], suggesting large-scale differences in mutation rates among chromosomes. We confirmed this general pattern using 1000Genomes Project data (Supplementary Section 1.8).

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 Section 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 generalised 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 Section 2.2.2), and calculated partial correlations between genetic gradient magnitudes and geographic factors.

Finestructure Analysis. FineSTRUCTURE[32] was run as described in Supplementary Section S2.2.1. 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 chunks, we focussed on regions of the genome that had an extended (>500Kb) 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 Section 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 African chunk lengths were 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 chunk length features (Supplementary Section 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 chunk length features (Supplementary Section 2.2.1).

MSMC, Denisova masking, simulations of alternative scenarios and assessment of phasing robustness. Genetic split times were initially calculated following the standard MSMC procedure[10], and subsequently modified as follows. To estimate the effect of archaic admixture, putative Denisova

haplotypes were identified in Papuans using a previously published method[39] 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, due 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-run MSMC after re-phasing one Estonian, one Papuan and 20 West African and Pygmies genomes in a single experiment. By 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. We learn a model for each individual l not in our panel,

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

where the positive mixture weights α_k sum to 1 and are otherwise learned from the $j \in (1..9)$ observations which we have data under quadratic loss. We can then predict the unobserved values

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

Examination of this matrix (Supplementary Section S2.2.3, Table S2.2.3-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 as African or Archaic chunk in Eurasians or Papuans a genomic locus spanning at least 1000bp, and showing SNPs that were assigned by chromopainter a 50% chance of copying from either an African or Archaic genome, respectively. For each chunk we then calculated the number of non-African mutations, defined as sites found in derived state in a given chunk 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 of chunks. 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 chunk 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 chunk 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, due 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 chunks that are very short. We note that whilst 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 $\pi_{ik} = \mathbb{E}(I_{ik}|l_i, X_i)$

along with the maximum likelihood parameter estimates $\{l_k, \sigma_k^2, \mu_k\}$. We do this for the four combinations of African (AFR) and Denisova (DEN) chunks found in Papuans (PNG) or Europeans (EUR), in order to learn the parameters. Supplementary Section S2.2.10 describes this in more detail. We then describe the distribution of chunks for each class c of chunk in terms of the expected proportion of chunks found in each component,

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

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

<u>*Single-out-of-Africa model:*</u> We fit African chunks in Papuans as a mixture of the others in a second layer of mixture modelling:

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

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

<u>*xOoA model:*</u> We jointly estimate an additional component π_{xOoA} and the mixture contributions β_c under the mixture

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

This is non-trivial to fit. We use a penalisation 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 penalisation 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, $P_A = |(\sum_c \beta_c) - 1| + |(\sum_k x_k) - 1|$ (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 minimised via standard optimization techniques. Supplementary Section S2.2.10 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 $\pi_{PNG.EUR}$ and $\pi_{PNG.PNG}$ using the single out-of-Africa model above, additionally allowing 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:

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

Estimates less than zero 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 chunks were assumed to be proxied by PNG.DEN (0% xOoA by assumption); African chunks by PNG.AFR; Papuan and Australian by PNG.PNG and all other chunks by PNG.EUR. We obtain confidence intervals by bootstrap resampling of haplotypes for each donor/recipient pair.

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 kya,[45,46]). 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 Section 2.2.12) that, if the xOoA entered the Papuan genome >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[15].

References

1 **Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52-58, doi:10.1038/nature09298 (2010).**

2 **The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65, doi:10.1038/nature11632 (2012).**

3 **Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327, 78-81, doi:10.1126/science.1181498 (2010).**

4 **Lachance, J. *et al.* Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* 150, 457-469, doi:10.1016/j.cell.2012.07.009 (2012).**

5 **Pagani, L. *et al.* Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. *American journal of human genetics* 96, 986-991, doi:10.1016/j.ajhg.2015.04.019 (2015).**

6 **Clemente, F. J. *et al.* A Selective Sweep on a Deleterious Mutation in CPT1A in Arctic Populations. *American journal of human genetics* 95, 584-589, doi:10.1016/j.ajhg.2014.09.016 (2014).**

7 **Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* 47, 435-444, doi:10.1038/ng.3247 (2015).**

8 **Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45, 1113-1120, doi:10.1038/ng.2764 (2013).**

9 **Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* 475, 493-496, doi:10.1038/nature10231 (2011).**

10 **Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* 46, 919-925, doi:10.1038/ng.3015 (2014).**

11 **Harris, K. & Nielsen, R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS genetics* 9, e1003521, doi:10.1371/journal.pgen.1003521 (2013).**

12 **Sheehan, S., Harris, K. & Song, Y. S. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. *Genetics* 194, 647-662, doi:10.1534/genetics.112.149096 (2013).**

13 **Patterson, N. *et al.* Ancient admixture in human history. *Genetics* 192, 1065-1093, doi:10.1534/genetics.112.145037 (2012).**

14 **Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet* 13, 745-753, doi:10.1038/nrg3295 (2012).**

15 **Groucutt, H. S. *et al.* Rethinking the dispersal of Homo sapiens out of Africa. *Evol Anthropol* 24, 149-164, doi:10.1002/evan.21455 (2015).**

16 Grove, M. *et al.* Climatic variability, plasticity, and dispersal: A case study from Lake Tana, Ethiopia. *Journal of human evolution* 87, 32-47, doi:10.1016/j.jhevol.2015.07.007 (2015).
17 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. *Proceedings of the National Academy of Sciences of the United States of America* 110, 10699-10704, doi:Doi 10.1073/Pnas.1306043110 (2013).
18 Reyes-Centeno, H. *et al.* Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proceedings of the National Academy of Sciences of the United States of America* 111, 7248-7253, doi:Doi 10.1073/Pnas.1323666111 (2014).
19 Liu, W. *et al.* The earliest unequivocally modern humans in southern China. *Nature* 526, 696-699, doi:10.1038/nature15696 (2015).
20 Prugnolle, F., Manica, A. & Balloux, F. Geography predicts neutral genetic diversity of human populations. *Current Biology* 15, R159-R160 (2005).
21 Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100-1104, doi:DOI 10.1126/science.1153717 (2008).
22 Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America* 102, 15942-15947 (2005).
23 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* 328, 710-722, doi:10.1126/science.1188021 (2010).
24 Reich, D. *et al.* Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American journal of human genetics* 89, 516-528, doi:10.1016/j.ajhg.2011.09.005 (2011).
25 Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514, 445-449, doi:10.1038/nature13810 (2014).
26 Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology* 23, 553-559, doi:10.1016/j.cub.2013.02.044 (2013).
27 Seguin-Orlando, A. *et al.* Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. *Science* 346, 1113-1118, doi:10.1126/science.aaa0114 (2014).
28 Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America* 108, 11983-11988, doi:10.1073/pnas.1019276108 (2011).
29 Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* 338, 222-226, doi:Doi 10.1126/Science.1224344 (2012).
30 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* 19, 1655-1664, doi:DOI 10.1101/gr.094052.109 (2009).

31 Cardona, A. *et al.* Genome-wide analysis of cold adaptation in indigenous siberian populations. *PloS one* 9, e98076, doi:10.1371/journal.pone.0098076 (2014).

32 Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS genetics* 8, e1002453, doi:10.1371/journal.pgen.1002453 (2012).

33 Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet* 48, 94-100, doi:10.1038/ng.3464 (2016).

34 Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* 505, 87-91, doi:10.1038/nature12736 (2014).

35 Jones, E. R. *et al.* Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature communications* 6, 8912, doi:10.1038/ncomms9912 (2015).

36 Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* 343, 747-751, doi:10.1126/science.1243518 (2014).

37 Rasmussen, M. *et al.* An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. *Science* 333, 94-98, doi:Doi 10.1126/Science.1211177 (2011).

38 Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468, 1053-1060, doi:10.1038/nature09710 (2010).

39 Wall, J. D. *et al.* Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics* 194, 199-209, doi:10.1534/genetics.112.148213 (2013).

40 Hughes, A. L. *et al.* Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proceedings of the National Academy of Sciences of the United States of America* 100, 15754-15757, doi:10.1073/pnas.2536718100 (2003).

41 Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451, 994-997 (2008).

42 Khrameeva, E. E. *et al.* Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans. *Nature communications* 5, 3584, doi:10.1038/ncomms4584 (2014).

43 Chapman, N. H. & Thompson, E. A. A model for the length of tracts of identity by descent in finite random mating populations. *Theoretical population biology* 64, 141-150 (2003).

44 Migliano, A. B., Vinicius, L. & Lahr, M. M. Life history trade-offs explain the evolution of human pygmies. *Proceedings of the National Academy of Sciences of the United States of America* 104, 20216-20219, doi:10.1073/pnas.0708024105 (2007).

45 Soares, P. *et al.* The Archaeogenetics of Europe. *Current Biology* 20, R174-R183 (2010).

46 Behar, D. M. *et al.* A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root. *American journal of human genetics* 90, 675-684, doi:Doi 10.1016/J.Ajhg.2012.03.002 (2012).

47 Karmin, M. *et al.* A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome research* 25, 459-466, doi:10.1101/gr.186684.114 (2015).

48 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, doi:10.1038/ng.937 (2011).

49 Posth, C. *et al.* Pleistocene Mitochondrial Genomes Suggest a Single Major Dispersal of Non-Africans and a Late Glacial Population Turnover in Europe. *Current biology : CB*, doi:10.1016/j.cub.2016.01.037 (2016).

50 Kuhlwilm, M. *et al.* Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* 530, 429-433, doi:10.1038/nature16544 (2016).

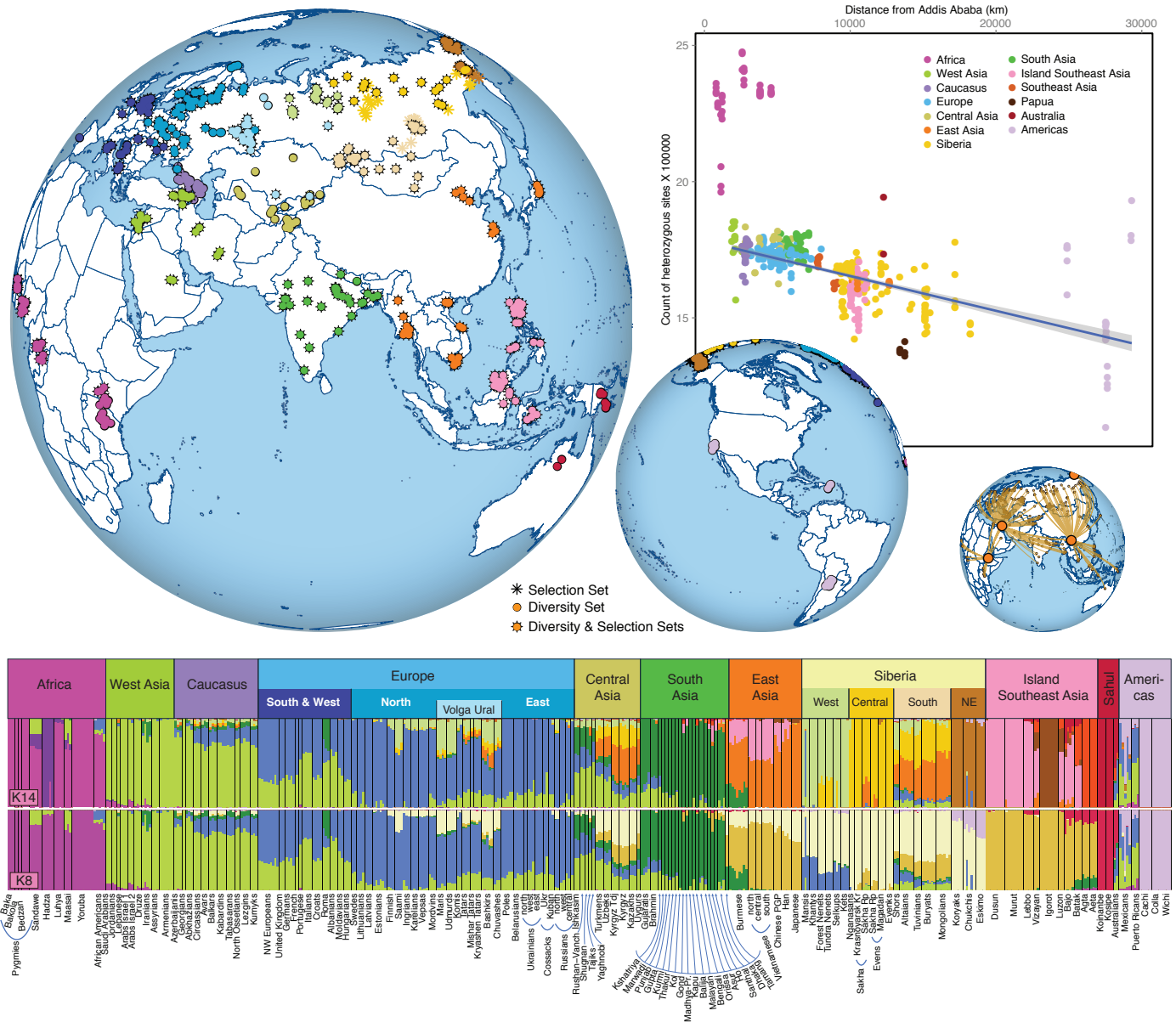
51 Clarkson, C. *et al.* The archaeology, chronology and stratigraphy of Madjedbebe (Malakunanja II): A site in northern Australia with early occupation. *Journal of human evolution* 83, 46-64, doi:10.1016/j.jhevol.2015.03.014 (2015).

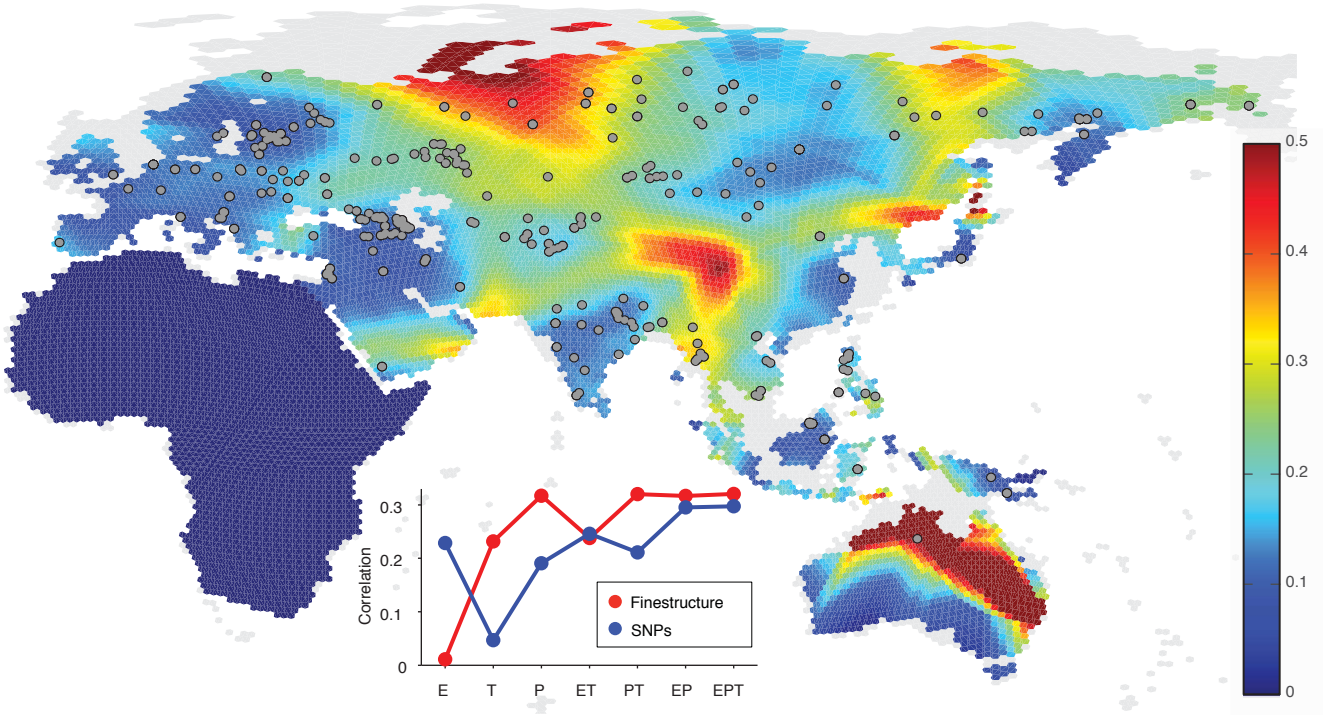
52 O'Connell, J. F. & Allen, J. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. *Journal of Archaeological Science* 56, 73 - 84, doi:http://dx.doi.org/10.1016/j.jas.2015.02.020 (2015).

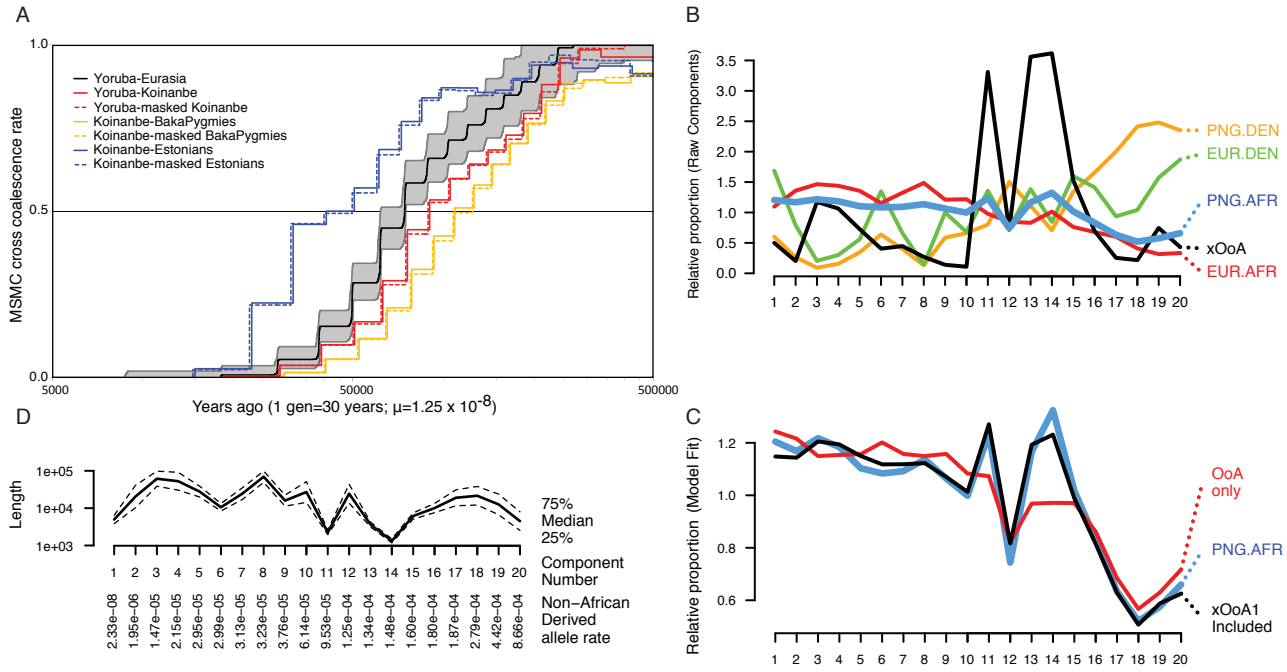
53 Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* 524, 216-219, doi:10.1038/nature14558 (2015).

54 Stewart, J. R. & Stringer, C. B. Human evolution out of Africa: the role of refugia and climate change. *Science* 335, 1317-1321, doi:10.1126/science.1215627 (2012).

55 Mikkelsen, T. *et al.* Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69-87 (2005).


Distance from Addis Ababa (km)
0
10000
20000
30000
25
20
15
Count of heterozygous sites X 100000
Africa
West Asia
Caucasus
Europe
Central Asia
East Asia
Siberia
South Asia
Island Southeast Asia
Southeast Asia
Papua
Australia
Americas
Selection Set
Diversity Set
Diversity & Selection Sets
Africa
West Asia
Caucasus
Europe
South & West
North
Volga Ural
East
Central Asia
South Asia
East Asia
Siberia
West
Central
South
NE
Island Southeast Asia
Sahul
Ameri-cas
K14
K8


0.5
0.4
0.3
0.2
0.1
0
Correlation
0.3
0.2
0.1
0
E
T
P
ET
PT
EP
EPT
Finestructure
SNPs


A
B
C
D
MSMC cross coalescence rate
Yoruba-Eurasia
Yoruba-Koinanbe
Yoruba-masked Koinanbe
Koinanbe-BakaPygmies
Koinanbe-masked BakaPygmies
Koinanbe-Estonians
Koinanbe-masked Estonians
Years ago (1 gen=30 years; μ=1.25 x 10^-8)
Relative proportion (Raw Components)
PNG.DEN
EUR.DEN
PNG.AFR
xOoA
EUR.AFR
Relative proportion (Model Fit)
OoA only
PNG.AFR
xOoA1 Included
Length
75%
Median
25%
Component Number
Non−African Derived allele rate

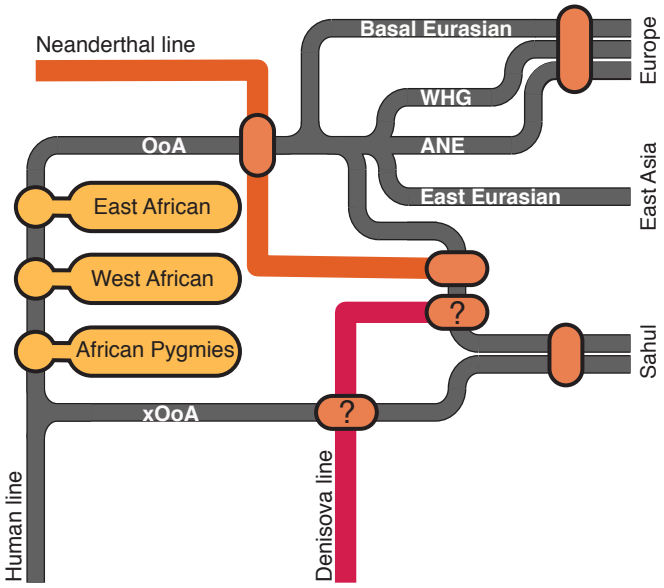

Neanderthal line
Basal Eurasian
Europe
WHG
OoA
ANE
East Asia
East Eurasian
East African
West African
?
African Pygmies
Sahul
xOoA
?
Human line
Denisova line

Table 1 Eurasian subset of variants highlighted by positive selection tests

Gene	SNP	Variant Type	Test	Population	Phenotype
FADS2	rs2524296	intronic	di	Wsi	Fatty acid desaturation
ZNF646	rs749670	missense	dDAF,DIND	CSi	Lipid metabolism, bile synthesis
PPARA	rs6008197	missense	iHS,nSL,TD,DIND	SoA	Lipid metabolism
GANC	rs8024732	missense	iHS,DIND	SoA	Carbohydrate metabolism
PKDREJ	rs6519993	missense	iHS,nSL,TD,DIND	SoA	Sperm-Receptor, kidney disease
CSMD1	rs7816731	non-coding	di	Wsi	Blood pressure
LYPD3	rs117823872	non-coding	di	Wsi	Wound healing
POU2F3	rs882856	missense	dDAF	WEu	Wound healing
B9D1	rs4924987	missense	dDAF	EEu	Ciliogenesis
PCDH15	rs4935502	missense	dDAF	CSi	Ciliogenesis
TMEM216	rs10897158	missense	dDAF	Wsi	Ciliogenesis
PLCB2	rs936212	missense	dDAF	NSi	Ciliogenesis
MYO18B	rs2236005	missense	dDAF	SeI	Motor activity
FLNB	rs12632456	missense	dDAF	SeI	Motor activity
TTN	rs10497520	missense	dDAF	MiE	Motor activity

Note the abbreviations of the population group names are according to Table S2.2

iHS,nSL, or TD, indicates that the variant is a from a top 1% window by that test for the indicated population. DIND indicates that the variant is significantly (>5SD) above the neutral background by the DIND test (See Supplementary Section 3)

di indicates that the variant was in the top 12 of the most highly divergent SNVs by the di score in each of the twelve population groups (See Supplementary Section 3)

dDAF indicates that the variant was in the top 20 most highly differentiated SNPs in its class in a given comparison (See Supplementary Section 3)