# Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation

**Authors:** Mayukh Mondal[1+], Ferran Casals[2+], Tina Xu[3+], Giovanni M. Dall'Olio[4], Marc Pybus[1], Mihai G. Netea[5], David Comas[1], Hafid Laayouni[1, 6], Qibin Li[3*], Partha P. Majumder[7*], Jaume Bertranpetit[1,8*]

**Affiliations:**

[1] Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

[2] Servei de Genòmica, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

[3] BGI Shenzhen, Yantian District, Shenzhen, 518083, China

[4] Computational Biology, Target Sciences, GSK R&D, GlaxoSmithKline , Stevenage, Hertfordshire , United Kingdom

[5]Department of Internal Medicine, Radboud University Medical Center, Nijmegen, The Netherlands

[6] Departament de Genètica i de Microbiologia, Universitat Autonòma de Barcelona, Bellaterra, Catalonia, Spain

[7] National Institute of BioMedical Genomics, Kalyani, West Bengal 741251, India

[8] Leverhulme Centre for Human Evolutionary Studies, Department of Archaeology and Anthropology, University of Cambridge, Cambridge, United Kingdom.

+ Co-first authorship

* Co-senior authorship

Correspondence: jaume.bertranpetit@upf.edu or ppm1@nibmg.ac.in

**Abstract:** To shed light on the peopling of South Asia and the origins of the morphological adaptations found there, we analyzed whole-genome sequences from ten Andamanese individuals and compared them with 60 individuals from mainland Indian populations with different ethnic histories, and with publicly-available data from other populations. We show that all Asian and Pacific populations share a single origin and expansion out of Africa, contradicting an earlier proposal of two independent waves[1–4]. We also show that populations from South and Southeast Asia harbor a small proportion of ancestry from an unknown extinct hominin, which is absent from Europeans and East Asians. The footprints of adaptive selection in the genomes of the Andamanese show that their characteristic distinctive phenotypes (including very short stature) do not reflect an ancient African origin, but instead result from strong natural selection on genes related to human body size.

**Main Text:**

The origin of the Andamanese people (Andaman Islands, Bay of Bengal, India) has been considered to be different from other Asian populations, because of their very distinctive so-called 'Negrito' morphology, and the unclassifiable language that they speak[5–7]. It has been suggested that they are a living relic of a first Out-of-Africa (OOA) wave of modern humans using the southern exit route, who did not subsequently mix with other populations[1,2] (since there have been multiple OOA events in human evolution, here 'OOA' refers to the Out-of-Africa event(s) for fully modern humans only). A common origin of the Andaman (and other) 'Negrito' populations, Melanesians and Australians, was initially proposed based on morphological characteristics[1,2] and subsequently supported by some genetic studies[4]. Previous analysis of genome-wide genotyping data from several Indian populations showed that the Andamanese are one of two main reference populations for estimating ancestries of Indian populations[8]. However, the lack of whole-genome sequence data from the Andamanese has limited understanding of both their ancestry and the specificity of the adaptations that may have resulted in their distinctive morphological features. Whether their distinctive 'Negrito' morphological features (small body size, dark skin, curly hair, etc.) are ancestral or derived may potentially be inferred by analyzing footprints of selection in their genomes. It matches known adaptations due to insularity in many groups of large animals, which may explain their fast evolution in body size, a feature that is shared by some extinct hominin populations[9] as well as present-day humans[10].

Seventy individuals from India were sequenced at ~15x coverage (Supplementary Note), including 60 individuals from mainland India and 10 from the Jarawa (JAR) and Onge (ONG) populations in the

Andaman islands (Supplementary Table 1, Supplementary Figure 1). The demographically small and historically isolated Andamanese population show higher relatedness among individuals as well as higher inbreeding coefficients and longer runs of homozygosity than all continental Indian populations examined (Supplementary Figure 2, 3 and 4). In agreement with previous studies[8,11], Principal Component Analysis (PCA) showed that the Andamanese constitute a genetically distinct cluster compared with the mainland Indian populations (Supplementary Figure 5). Interestingly, the Jarawa and the Onge cluster tightly together, indicative of their genomic homogeneity, and show a lack of recent admixture (Figure 1a), which is known to have taken place in Andaman during the last century[12], but did not affect the individuals sampled.

Using several approaches, we investigated whether the Andamanese were descendants of the same OOA event that resulted in the peopling of mainland India, or whether some part of their origins can be traced to an earlier and independent OOA wave, as has been proposed for Aboriginal Australians[4]. First, the D-statistic (Dstat) analysis[13] (Supplementary Figure 6) showed that Andamanese share more alleles with all OOA populations than with sub-Saharan Africans, suggesting that Andamanese shared a common and similar ancestry with all other OOA populations. Second, TreeMix analysis[14] also supports Africans as an outgroup to all OOA populations (Figure 1b), with a closer relationship of Andamanese with Asians and continental Indians than with Pacific populations. Third, relative cross coalescent analysis by MSMC[15] displayed a much earlier split for Andamanese and Africans than for Andamanese and any other OOA population, which are themselves very similar (Figure 1c). Estimation of historical effective population sizes by MSMC suggests a similar bottleneck event for Andamanese and all other OOA populations at around 50,000 years ago (Supplementary Figure 7). All of these results suggest that the Andamanese shared a common ancestry with all the other OOA populations, indicative of a commonality of all Asian and Pacific populations and consistent with a single main OOA migration.

Dstat analysis (Supplementary Figure 8) revealed that the Andamanese shared more alleles with East Asian, Papuan, and mainland Indian tribal populations than with Europeans, indicating that Europeans are an out-group for all Asian populations. Both TreeMix (Figure 1b) and Dstat outgroup analysis (Supplementary Table 2) supported this inference. Relative cross-coalescent analysis (Figure 1c) also showed a similar result: the separation between Andamanese and Europeans predates the separation of Andamanese from Asians. Analysis using available ancient European genome sequences from La Braña, Loschbour, and Stuttgart[16–18] supported our results (Supplementary Figure 8-10 and Supplementary Table 3), showing Europeans as the most distinct branch of all Eurasian and Pacific populations, even when considering the extinct Basal

Eurasian component of Europeans[18,19]. Mitochondrial DNA analysis also supports a single origin for Asian populations (Supplementary Table 4).

The analysis of the contribution of extinct hominin populations to the current genetic pool also suggests a single origin for modern Asians, including Andamanese. Andamanese genomes have a similar amount of Neanderthal[13,20] introgression to other OOA populations (~2-4%), suggesting that the Neanderthal admixture took place at a very early stage, before the OOA populations separated from each other (Supplementary Figure 12). On the other hand, Papuans harbor a much higher proportion of Denisovan[21] ancestry than any other OOA population examined here (Supplementary Figure 13); all other Asian populations examined (including the Andamanese) have only slightly more Denisovan ancestry than Europeans (Supplementary Figure 14), as previously suggested[20]. Besides that, no other difference in ancient contributions was observed between the Andamanese and other Southern or Eastern Asian or Pacific populations.

We found that Andamanese, mainland Indian and Papuan populations carry ~2-3% fewer African alleles than Europeans (Figure 2a) or East Asians (Figure 2b), as do Australians (similar yet higher value, see below), a very intriguing result. We performed extensive simulations to show that this deficiency of African alleles in the Andamanese cannot be explained by the Andamanese having low effective population size; thus is not caused by private variants produced by specific mutations in their genome (no Admixture model, Supplementary Table 5), or by later admixture between Europe or Asia and Africa (i.e. it cannot be due to a "back to Africa" event; Supplementary Note and Supplementary Table 5), or by admixing with the initial OOA modern humans settling in Eurasia. In contrast, it could be caused by mixture with a population that diverged at least 300 kya (Supplementary Figure 15). In fact, an introgression from any hominin population that can cause a bias in the Dstat calculations (Supplementary Note) would generate a false two-wave of OOA (for modern humans) signal for the South Asian and Pacific populations, which is not observed. This reduction in African ancestry for South Asian populations likewise cannot have originated from Neanderthals or Denisovans, as these two populations have similar amounts of well-recognized ancestry in Andamanese and East Asians. An alternative hypothesis is that this 2-3% reduction of African ancestry originated from admixture with other hominin population(s) in Southeast Asia, such as *Homo erectus*[22] or an unknown extinct archaic population. A three-population model[23] confirms it (Supplementary Note and Supplementary Figure 16). By calculating Dstat values for 50kb regions with a sliding window, we infer that this unknown population diverged from Neanderthals and Denisova before they diverged from each other, as seen initially by TreeMix (Supplementary Figure 17). To further identify specific DNA regions derived from

this hominin population, we implemented Sstar[24] on these putative fragments, and detected ~15Mb per individual (average region length 65kb) from this hominin population that behaves either as a sister group to Neanderthal and Denisova or even diverged earlier (Supplementary Figures 18 and 19). For Aboriginal Australians, the deficit of African alleles is even higher (~6-7%; Figure 2), suggesting that this reduction might be caused by admixture with some unknown ancient hominin population; this result needs to be confirmed with additional Australian data. Rasmussen et al.[4] suggested that Aboriginal Australians are the descendants of admixture of the first OOA with later OOA populations. We failed to detect this first OOA event either by Dstat (Supplementary Tables 6 and 7) or relative cross-coalescent analysis by MSMC (Supplementary Figure 20). Our simulations suggest that the bias in Dstat calculation, which was interpreted as the product of the first OOA population admixture with Aboriginal Australians, can instead be explained by ancient hominin admixture with Aboriginal Australians.

To explain the genetic structure of mainland India, it has been suggested[8] that all populations have arisen from admixture between two components: (1) Ancestral North Indian (ANI) and (2) Ancestral South Indian (ASI), which is genetically related to Andamanese. However, although ADMIXTURE analysis (Figure 1a) showed that the Irula (ILA) and Birhor (BIR) tribal populations have high amounts of this ASI component, also present in all the other non-tribal populations of Southern India examined (shown also in[11,25]), TreeMix analysis (Figure 1b) suggested that Andamanese are not directly related to this South Indian component. Rather, the Andamanese are slightly closer to East Asians than to these two tribal Indian populations. Also, the Andamanese do not share direct ancestry with the Australian and Papuan sequences tested (Figure 1b), as has been traditionally assumed because of morphological similarities between these populations[1].

Since we have shown that the Andamanese and other modern Asian populations have a common origin, we hypothesized that the distinct phenotype of the Andamanese should have originated by recent adaptation to their environment. To detect positive selection we used the Hierarchical Boosting (HB) method, a machine-learning classification framework that exploits the combined ability of some selection tests to uncover features expected under the hard sweep model, while controlling for population-specific demography, achieving higher power than single tests and a low rate of false positive results[26]. We found some 1,000 genomic regions to have significant footprints of positive selection among the Andamanese (212 regions, encompassing 107 genes, under the complete hard sweep model; and 805 regions, encompassing 509 genes, under the incomplete hard sweep model). Among them, we found a significant excess of genes related to body morphology, with signals in 11 of the 107 genes related to height (according to the Genetics

Association Database, GAD[27]) for complete selective sweeps (Yates Chi Square=5.70, P=0.02) and 48 out of 509 for incomplete sweeps (Yates Chi Square=22.59, P<0.0001). Other regions under positive selection included genes related to obesity or body shape and composition. It is interesting to note that these results point to selective pressure on body size, likely related to low stature (in fact, the very low stature of Andamanese can be recognized by the individual genotypes at height-related SNPs; see Supplementary Figure 21); it could therefore represent insular dwarfism, a well-known adaptation of large animals to a restricted environment that predicts a derived state for the morphology of the Andamanese. These results thus provide insights into the biological bases of such adaptations, also described recently in Sardinia[9].

Our analysis supports a distinct model for the human settlement of Asia and Pacific, with two novel insights (Figure 3): (i) Asian populations, including ones from the Pacific, have a single origin and OOA expansion, sharing a more recent common ancestor between themselves than with Europeans; our analyses do not support the hypothesis of two independent OOA events, postulated a long time ago based on physical appearance[1] and apparently confirmed by genetics[4]; and, (ii) Indian mainland populations, Andamanese, Papuans and Aboriginal Australians (but not East Asians) carry genomic contributions from an extinct hominin population, with admixture ranging between 2-3% (higher in Australians, but this estimate needs to be confirmed with new data). Our results do not indicate whether or not the introgression is derived from the same hominin in all populations, but in the case of the Andamanese (Supplementary Figure 22) we have shown that it comes from a new unknown hominin population, that likely separated very early in the hominin tree. Also, we have shown that the hominin admixture in these populations can cause a bias in Dstat calculation that can be erroneously interpreted as a first OOA migration of modern. Finally, the distinctive morphology of the Andamanese (and probably of other 'Negrito' populations) has probably originated from strong adaptive selection as shown by the excess of genes under selection related to height and body mass, and it is not an ancestral character, but derived, leading to the possibilities of understanding the basic biology of a complex adaptation in an island environment.

# References


1. Coon, C. S. & Hunt, E. E. *The living races of man*. (Knopf, 1966).
2. Molnar, S. *Human Variation: Races, Types and Ethnic Groups*. (Routledge, 2015).
3. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The history and geography of human genes*. (Princeton University Press, 1994).
4. Rasmussen, M. *et al.* An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. *Science* **334,** 94–98 (2011).
5. Huxley, T. H. On the Geographical Distribution of the Chief Modifications of Mankind. *J. Ethnol. Soc. London* **2,** 404–412 (1870).
6. Brown, A. R. *The Andaman Islanders: A Study in Social Anthropology*. (Cambridge University Press, 1922).
7. Abbi, A. Is Great Andamanese genealogically and typologically distinct from Onge and Jarawa? *Lang. Sci.* **31,** 791–812 (2009).
8. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461,** 489–494 (2009).
9. Montgomery, S. H. Primate brains, the 'island rule' and the evolution of Homo floresiensis. *J. Hum. Evol.* **65,** 750–760 (2013).
10. Zoledziewska, M. *et al.* Height-reducing variants and selection for short stature in Sardinia. *Nat. Genet.* **47,** 1352–1356 (2015).
11. Basu, A., Sarkar-Roy, N. & Majumder, P. P. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proc. Natl. Acad. Sci.* **113,** 201513197 (2016).
12. Dass, F. *The Andaman Islands.PDF*. (The Good Shepherd Convent Press, 1937).
13. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
14. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8,** e1002967 (2012).
15. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46,** 919–925 (2014).
16. Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505,** 87–91 (2014).
17. Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507,** 225–228 (2014).
18. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).
19. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522,** 207–211 (2015).

20. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).

21. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338,** 222–226 (2012).

22. Swisher, C. C. *et al.* Latest Homo erectus of Java: potential contemporaneity with Homo sapiens in southeast Asia. *Science* **274,** 1870–1874 (1996).

23. Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C. & Wall, J. D. Genetic evidence for archaic admixture in Africa. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 15123–15128 (2011).

24. Vernot, B., Akey, J. M. & Vernot B., A. J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343,** 1017–1021 (2014).

25. Juyal, G. *et al.* Population and genomic lessons from genetic analysis of two Indian populations. *Hum. Genet.* **133,** 1273–1287 (2014).

26. Pybus, M. *et al.* Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics* **31,** btv493 (2015).

27. Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The genetic association database. *Nat. Genet.* **36,** 431–432 (2004).

## URLs

European Nucleotide Archive (http://www.ebi.ac.uk/ena), Picard tools (http://picard.sourceforge.net/), Broad ftp server (ftp.broadinstitute.org), 1000 Genome ancestral file (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/).

## Accession codes

The whole-genome sequences (Andamanese vcf files) have been deposited in the European Nucleotide Archive, Accession ID: PRJEB11455.

## Acknowledgements

Our main funding was provided by the joint Spain-India bilateral grant PRI-PIBIN-2011-0942 from the Ministerio de Economía y Competitividad (Spain). Complementary funding was from grant BFU2013-43726-P from the Ministerio de Economia y Competitividad (Spain), with the support of Secretaria d'Universitats i Recerca, Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2014 SGR866). Jessica Nye (Universitat Pompeu Fabra) and Chris Tyler-Smith kindly corrected in depth the manuscript. Thanks to R.A. Foley for discussion and inspiring input for Figure 3.

## Author Contributions

MM, FC, PPM and JB conceived and designed the project. PPM provided the samples. PPM, ZH and QL sequenced samples and carried out initial analyses. MM performed the remaining genetic data analyses. FC, GMDO, MP, MGN, DC, HL, PPM and JM participated in and discussed analyses. MM, FC, PPM and JB wrote the manuscript.

## Author Information

The authors declare no competing financial interests. Correspondence should be addressed to JB (jaume.bertranpetit@upf.edu) or PPM (ppm1@nibmg.ac.in).

**Fig. 1: Ancestry of Indian Populations.**

a. ADMIXTURE analysis using 10 randomly-chosen individuals from CEU, CHB and YRI taken from the 1000 Genomes Project and individuals from our data set: Punjabi (PUN), Uttar Pradesh Brahmins (UBR), Rajput (RAJ), Bengali (BEN), Vellalar (VLR), Irula (ILA), Birhor (BIR), Jarawa (JAR), Onge (ONG) and Riang (RIA). Results are shown for five ancestral components, the optimal number. Each vertical bar represents one individual, colored according to the proportion of the five ancestral components.

b. TreeMix analysis without migration. Africans are Yoruba (YRI), Mandenka (MAD), Mbuti pygmy (MBT) and San (SAN), Europeans are French (FRN) and Sardinian (SAR), East Asians are Dai (DAI) and Han Chinese (HAN), Pacific population are Papuans (PAP) and Aboriginal Australians (AUS), and Indians are (BIR, ILA and RIA) and Andamanese (JAR and ONG). Inferred ancestral genome information from the 1000 Genomes Project was used as outgroup. The scale bar shows 10 times the standard error and the x axis shows the amount of drift. Drift considered non-significant is indicated by a red line, so the three branches (RIA, HAN, DAI), (ONG, JAR) and (BIR, ILA) form a trichotomy.

MSMC Relative Cross Coalescent Rate showing genetic separation between two populations. In each curve one individual was from Jarawa (JAR) and the other from either a Tribal population of India (ILA, BIR or RIA), ONG, or from outside India (FRN, DAI, PAP and YRI). The x-axis shows time and the y-axis shows a measure of the similarity between the two populations.

**Fig. 2: Fewer African derived alleles in Indian, Andamanese, Papuan and Aboriginal Australians than Europeans or East Asians.** Each horizontal line shows the result of D-statistics [Dstat(W,X;Y,Z)] where the W population is either a) French (FRN) or b) East Asian Dai (DAI). The X population is either from India: Punjabi (PUN), Uttar Pradesh Brahmins (UBR), Rajput (RAJ), Bengali (BEN), Vellalar (VLR), Irula (ILA), Birhor (BIR) or Riang (RIA); Andamanese: Jarawa (JAR) or Onge (ONG); and FRN, Sardinia (SAR), DAI, Han Chinese (HAN), Papuans (PAP) or Aboriginal Australians (AUS); names are shown to the right of the two figures. The Y population is African (Yoruba (YRI), Mandenka (MAD), Mbuti pygmy (MBT) or San (SAN). Ancestral allele information from the 1000 Genomes Project is used as outgroup (Z population). Colour coding of the populations: Europeans (Pink), East Asians (Deep Yellow), African (Brown), Indo-Europeans (Red), Dravidians (Black), Austro Asiatics (Blue), Andamanese (Light Green), Tibeto Burman (Yellow), Pacific Islanders and Australian Aboriginals (Deep Green). A positive value means that the W and Y populations share more derived alleles with each other compared with X and Y, while a negative value means X and Y populations share more derived allele with each other as compared with W and Y. The statistically significant results (in this case defined by a Z score more or less than ±3) are marked with a star.

a. Dstat results of D(FRN(W),X;AFR(Y),Ancestral(Z)).
b. Dstat results of D(DAI(W),X;AFR(Y),Ancestral(Z)).

**Fig. 3: Model of gene flow in Asia.** Red boxes indicate extinct non-African hominins who introgressed into modern humans; these introgressions are marked with dotted lines. The green box indicates populations that may have admixed with the new unknown hominin; Andamanese and Indian are fully analyzed here; the others will have to be further studied in the future. To properly solve the question mark trichotomy would require more data.

# Methods

## Samples

In total, 70 samples were collected from 10 Indian populations from different geographical regions, linguistic affiliations and social categories (Supplementary Table 1). The 10 populations were: Punjabi (PUN), Uttar Pradesh Upper caste Brahmins (UBR), Rajput (RAJ), Bengali (BEN), Vellalar (VLR), Irula (ILA), Birhor (BIR), Jarawa (JAR), Onge (ONG) and Riang (RIA). The blood and saliva samples were collected with voluntary informed consent from the participants. More information on the populations is found in Basu et al[11].

Additional samples were also used to understand Indian populations from a global perspective. We used the 1000 Genomes Phase 1 data[28], the Great Ape Genome Project (GAGP) data[29], high-coverage data from three Aboriginal Australians[30], nine Yoruba (YRI) high-coverage data and five Utah residents with Northern and Western European Ancestry (CEU)[31]. We used some Ancient genome sequences: Malta[16], La Braña[17], Loschbour and Stuttgart[18]. Neanderthal[20] and Denisova[21] data were used to calculate the admixture level of these subspecies in Indian populations. We have used the 1000 Genomes Project ancestral file[32] to identify the ancestral allele.

## Sequencing

The whole-genome sequencing was done in two different places (BGI, NIBMG) using Illumina technology. 50 of the 70 samples were sequenced in BGI, whereas 20 were sequenced in NIBMG (Supplementary Tables 1 and 8). Sequencing libraries with an insert size of ~500 bp were constructed and paired-end reads were generated by HiSeq 2000. The raw sequencing reads were mapped to hg19 using BWA[33]. Duplicates were removed by Picard tools. We followed best practice recommendations from GATK 2.8-1[34] using IndelRealigner and BaseRecalibrator with their default values. For IndelRealigner we used 1000 Genomes Phase 1 indel interval files, and for BaseRecalibrator we used dbSNP 137. Variants were called by HaplotypeCaller from GATK. After creation of the raw vcf files, we used VariantRecalibrator from GATK on the autosomes using dbSNP 137, HapMap 3.3, 1000 Genomes Project Omni 2.5 and 1000 Genomes Project Phase 1 SNPs with high confidence, Mills and 1000 Genomes Project gold standard indels to assign a well-calibrated probability to each variants; all these files were downloaded from the Broad Institute ftp site (date 11/05/2013) as described in the website of GATK. The average coverage for autosomes was ~15x and the accessible genome was close to 100% (Supplementary Table 8). Though the sequencing was done in

two different institutes, Principal Component (PC) and ADMIXTURE analysis (Supplementary Note) demonstrated a very tight clustering for samples from the same population, suggesting that influences from the two sequencing centers were not detectable.

## Relatedness, Inbreeding and Homozygosity Run

Relatedness was calculated using KING[35] software with 13,679,600 autosomal bi-allelic SNPs. Inbreeding was calculated by vcftools[36] using the same SNPs and the default parameters. Homozygosity runs were done by PLINK v1.07[37] software using 4,475,795 autosomal bi-allelic unlinked SNPs with the default parameters. SNPs were unlinked according to the variance inflation factor (VIF) method implemented in PLINK with a window size of 50 SNPs, a step size of 5, and a variance inflation factor of 2.

## PCA

SmartPCA from the EIGENSOFT package[38] was used for PCA. We kept only autosomal, bi-allelic SNPs that have Minor Allele Frequency (MAF) of at least 0.05. We also removed SNPs which had missing information for any individual. Only 10 individuals per population from the 1000 Genomes Project data were kept to avoid sample size bias.

## Admixture

ADMIXTURE[39] was used to calculate admixture per individual with the same filters as the PCA analysis. To explore the optimal number of ancestral populations (k), we used k= 2–6, performing ten iterations for each. The best k value was estimated using the cross-validation error method implemented in ADMIXTURE.

## MSMC

Effective population size and population separation over time were calculated using MSMC[15]. Only autosomes were used. MSMC recommendations were followed to create input files from BAM files. We phased genomes using 1000 Genomes Project Phase 3 data as the reference using Shapeit[40].

## Dstat

ADMIXTOOLS were used[41] for Dstat analysis. To reduce biases (especially ascertainment bias), we called variants from India and the Great Ape Genome Project (only humans) together as described above. SNP information from Aboriginal Australians, Neanderthal, Denisova and other ancient samples were extracted as described in Supplementary Information 5. Ancestral information was extracted from the fasta file given on the 1000 Genomes Project website.

## TreeMix

TreeMix[14] was used to analyse the divergence of the populations from each other, using the data described above. We used migration values from 0 to 20. The inferred ancestral genome was used to root the tree. To allow for linkage disequilibrium (LD) we used the -k flag. The LD blocks were defined as 1 Mb in length, which in our case corresponds to about 5,000 SNPs.

## Simulations

For simulations, we used ms[42] following published parameters[43]. We added Andamanese parameters determined from our inferences about Andamanese ancestry (See Supplementary Note).

## Dadi and a three-population model for Archaic Admixture

We first built a null model without introgression of archaic hominins into the Andamanese using dadi-1.7.0[44] following parameters from Gravel et al[43]. Then a three-population model for archaic admixture was implemented to estimate the divergence of this unknown population from humans and the time of admixture with Andamanese by simulating 2% of hominin genome introgression into Andamanese at different time points.

## Selection

This analysis, used Andamanese genomes from our data and YRI sequences from Complete Genomics[31] and merged them. After removing any SNP which has missing information for any individual, we phased the Andamanese with Shapeit[45] using 1000 Genomes Project phase 1 samples as a reference[40]. Then, the following selection tests were performed on the data:

1. Tajima's D[46].
2. CLR[47].
3. Fay and Wu's H[48].
4. Fu & Li's D[49].
5. XP-EHH[50].
6. ΔiHH[51].
7. iHS[51].
8. EHH average[52].

After calculating all tests, we ran the boosting algorithm[26] using parameters both from the East Asian and the European hierarchical boosting strategy (simulated under neutrality and under selection using cosi with demographic models from Schaffner et al[53] for both East Asian and European demography and then calculating the best strategy to detect selection). In fact, results for the hierarchical boosting strategy for non-African populations are very similar (Supplementary Note). Information about body size genes was obtained from the Genetics Association Database[27] and their functional annotation from ANNOVAR[54]

**Dstat with sliding windows and Sstar**

To identify candidate introgressed regions from an unknown hominin, we calculated Dstat per individual for 50 kb regions with sliding windows of 5kb and retained regions where Andamanese have fewer African derived alleles than Europeans or East Asians:

$$D_{stat} = \frac{\sum(F_w - F_x)(F_y - F_z)}{\sum(F_w + F_x - 2F_wF_x)(F_y + F_z - 2F_yF_z)}$$

F is the allele frequency in w, x, y or z populations.

We ran TreeMix on the putative introgressed regions (Supplementary Note) and Sstar[24] to refine the identification of the introgressed hominin haplotypes thus only taking regions which is positive for both Dstat by sliding windows and Sstar (Supplementary Note).

# Methods Reference


28. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

29. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499,** 471–475 (2013).

30. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349,** aab3761 (2015).

31. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2010).

32. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

34. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

35. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).

36. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

37. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

38. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).

39. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).

40. Delaneau, O. *et al.* Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat. Commun.* **5,** 3934 (2014).

41. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192,** 1065–1093 (2012).

42. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338 (2002).

43. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 11983–11988 (2011).

44. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5,** (2009).

45. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9,** 179–181 (2011).

46. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).

47. Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res.* **15,** 1566–1575 (2005).

48. Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155,** 1405–1413 (2000).

49. Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133,** 693–709 (1993).

50. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449,** 913–918 (2007).

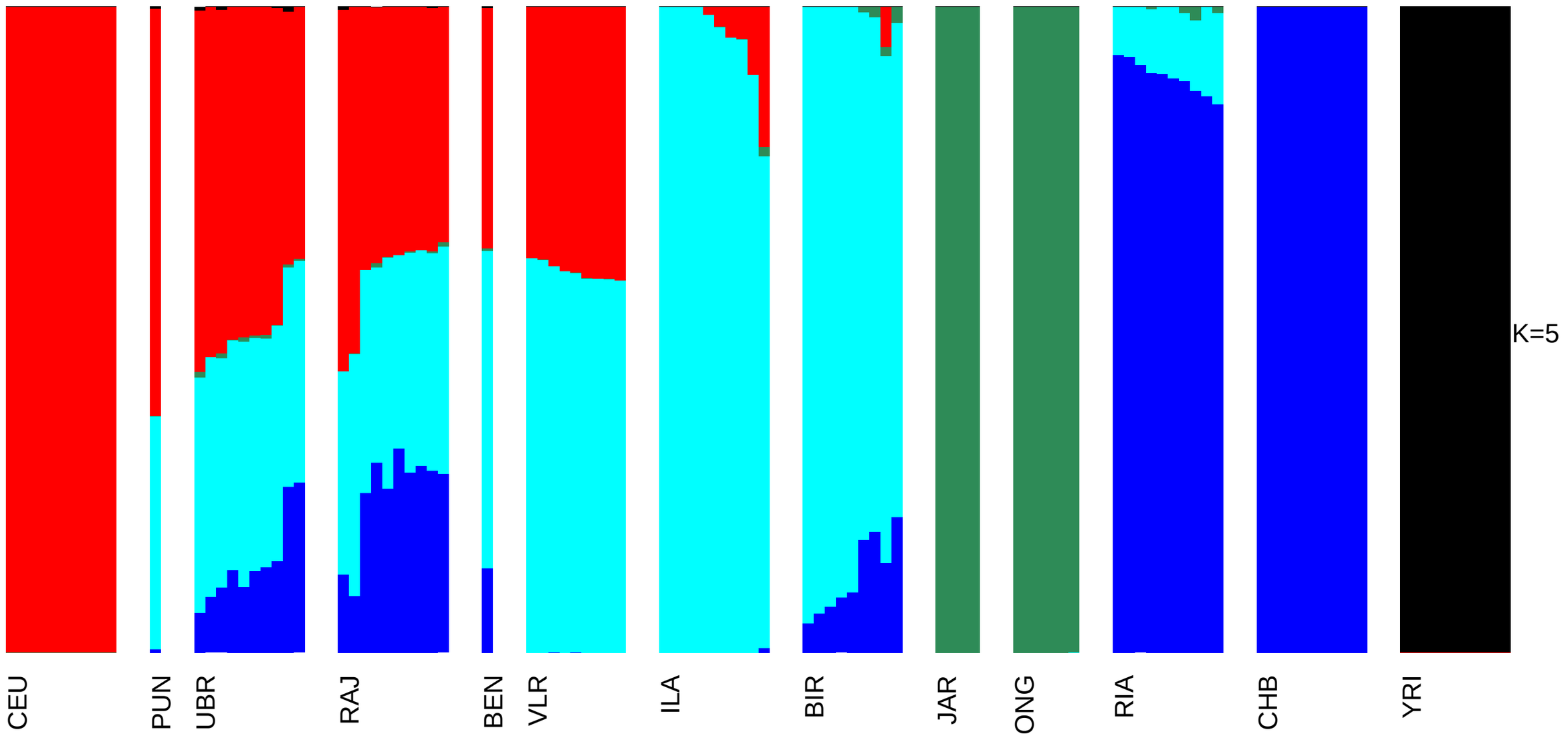
51. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4,** 0446–0458 (2006).

52. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).

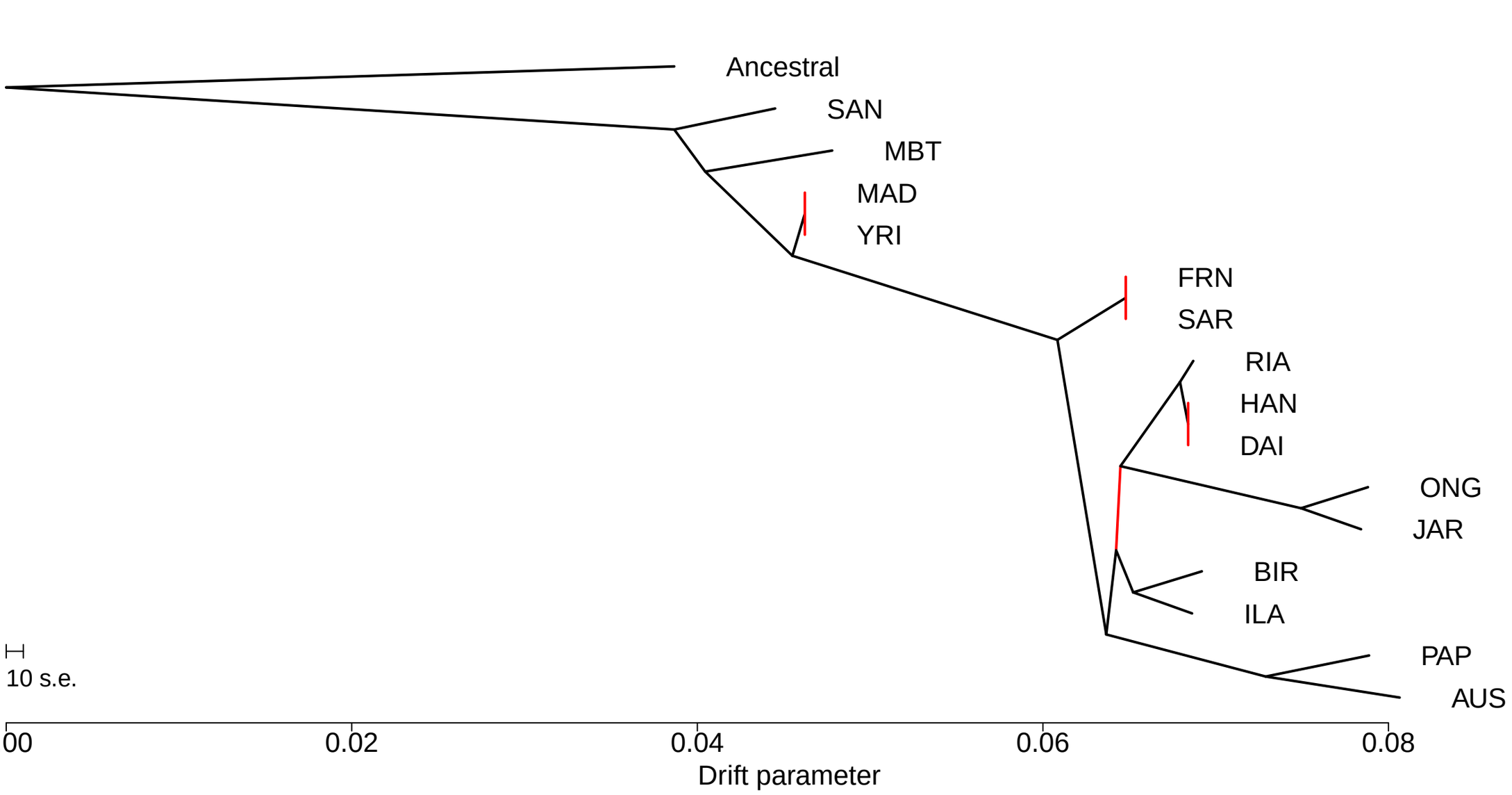
53. Schaffner, S. F. *et al.* Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15,** 1576–1583 (2005).

54. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

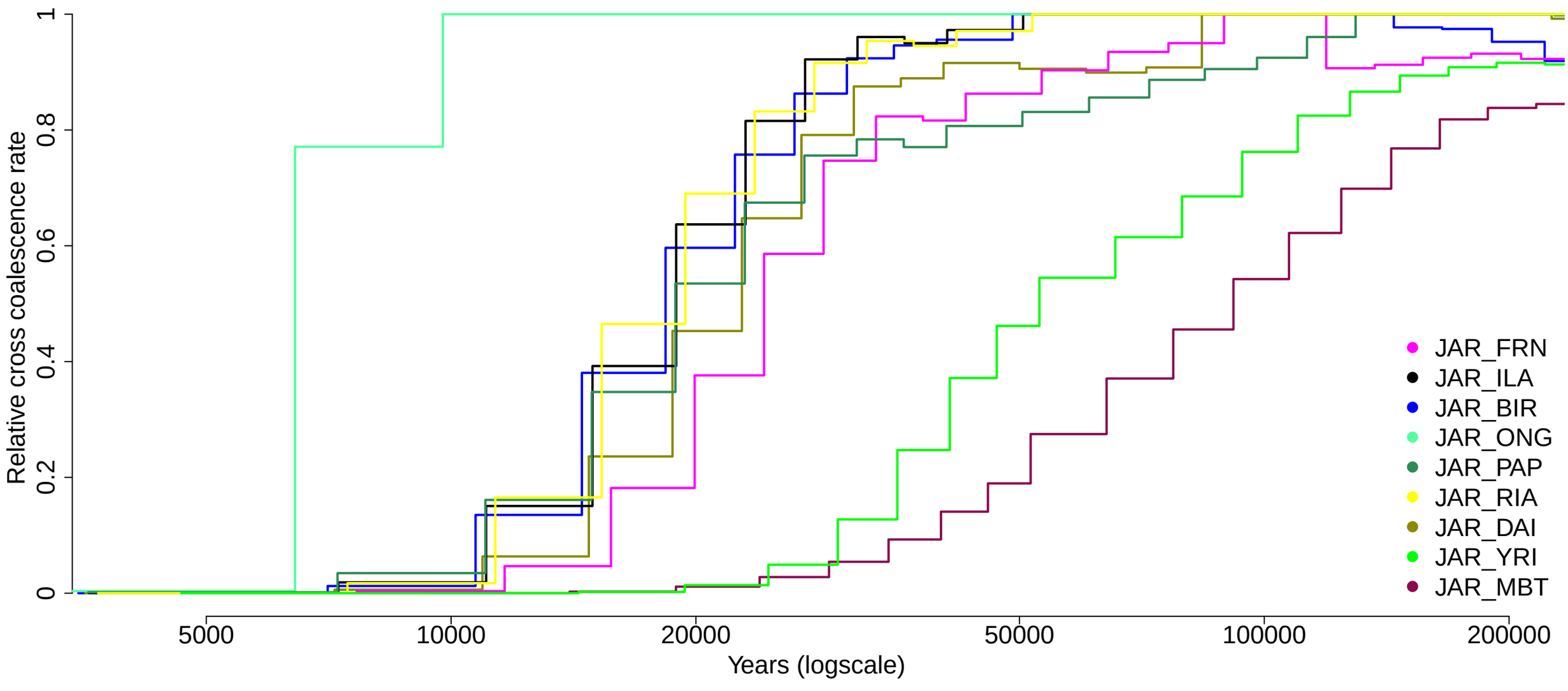
**a**


K=5
CEU
PUN
UBR
RAJ
BEN
VLR
ILA
BIR
JAR
ONG
RIA
CHB
YRI

**b**


Ancestral
SAN
MBT
MAD
YRI
FRN
SAR
RIA
HAN
DAI
ONG
JAR
BIR
ILA
PAP
AUS
10 s.e.
0.00
0.02
0.04
0.06
0.08
Drift parameter

**c**


Relative cross coalescence rate
0
0.2
0.4
0.6
0.8
1
5000
10000
20000
50000
100000
200000
Years (logscale)
JAR_FRN
JAR_ILA
JAR_BIR
JAR_ONG
JAR_PAP
JAR_RIA
JAR_DAI
JAR_YRI
JAR_MBT

a

Ancestral(Z)

SAN (Y)
MBT (Y)
MAD (Y)
YRI (Y)

FRN(W)

SAR(X)
PUN(X)
UBR(X)
RAJ(X)
BEN(X)
VLR(X)
ILA(X)
BIR(X)
JAR(X)
ONG(X)
PAP(X)
AUS(X)
RIA(X)
DAI(X)
HAN(X)

0.08 0.06 0.04 0.02 0.00 −0.02

Dstat (W,X;Y,Z)

b

Ancestral(Z)

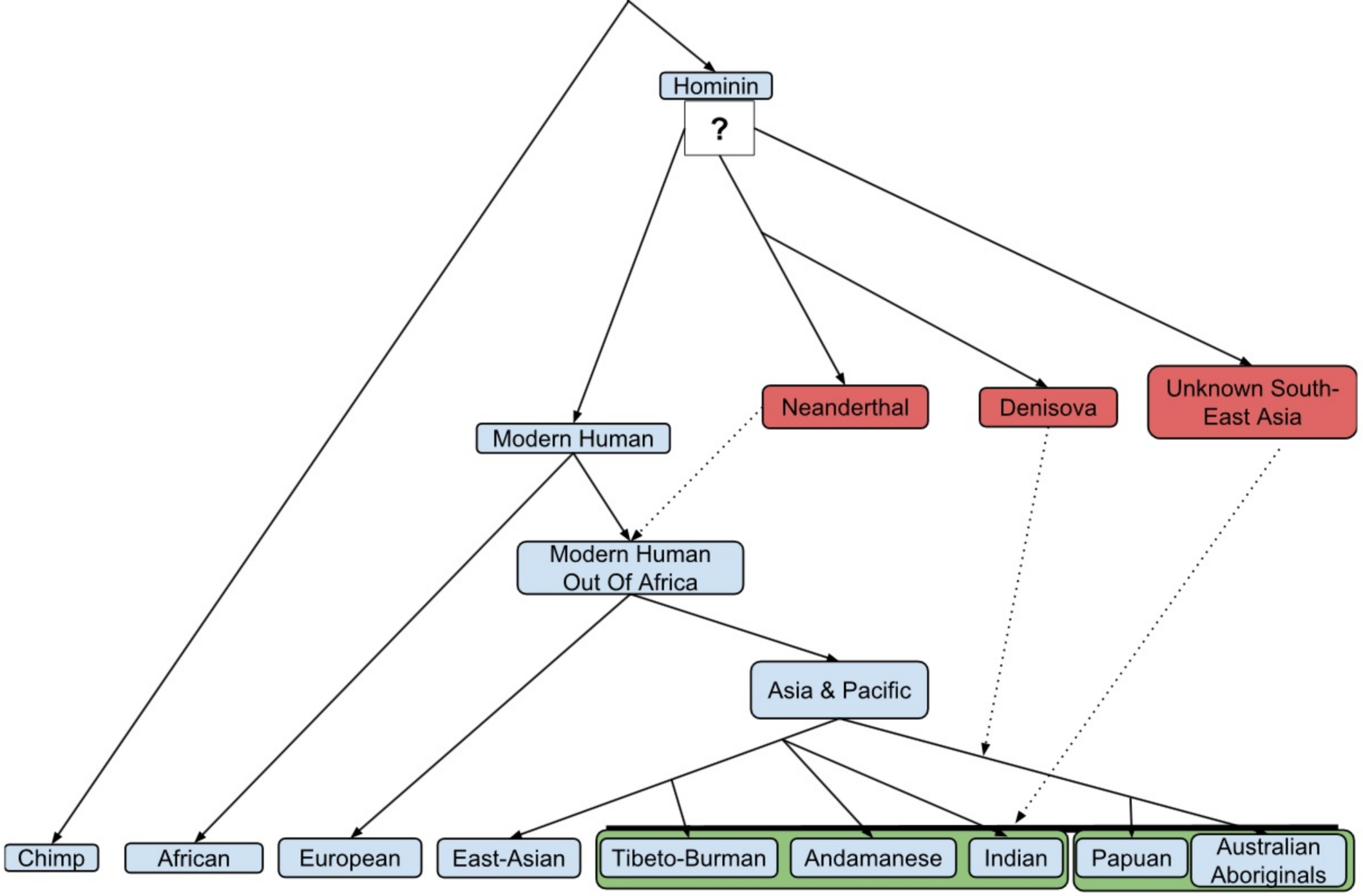
SAN (Y)
MBT (Y)
MAD (Y)
YRI (Y)

DAI(W)

SAR(X)
FRN(X)
PUN(X)
UBR(X)
RAJ(X)
BEN(X)
VLR(X)
ILA(X)
BIR(X)
JAR(X)
ONG(X)
PAP(X)
AUS(X)
RIA(X)
HAN(X)

0.08 0.06 0.04 0.02 0.00 −0.02

Dstat (W,X;Y,Z)


Hominin
?
Modern Human
Neanderthal
Denisova
Unknown South-East Asia
Modern Human Out Of Africa
Asia & Pacific
Chimp
African
European
East-Asian
Tibeto-Burman
Andamanese
Indian
Papuan
Australian Aboriginals