nature genetics

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

Mayukh Mondal[1,9], Ferran Casals[2,9], Tina Xu[3,9], Giovanni M Dall'Olio[4], Marc Pybus[1], Mihai G Netea[5], David Comas[1], Hafid Laayouni[1,6], Qibin Li[3,10], Partha P Majumder[7,10] & Jaume Bertranpetit[1,8,10]

To shed light on the peopling of South Asia and the origins of the morphological adaptations found there, we analyzed whole-genome sequences from 10 Andamanese individuals and compared them with sequences for 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 of migration[1–4]. We also show that populations from South and Southeast Asia harbor a small proportion of ancestry from an unknown extinct hominin, and this ancestry is absent from Europeans and East Asians. The footprints of adaptive selection in the genomes of the Andamanese show that the characteristic distinctive phenotypes of this population (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.

The origin of the Andamanese people (Andaman Islands, Bay of Bengal, India) has been considered to be different from that of other Asian populations because of the very distinctive so-called 'Negrito' morphology in Andamanese 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 who used the southern exit route and did not subsequently mix with other populations[1,2] (there have been multiple OOA events in human evolution, but 'OOA' here refers only to Out-of-Africa event(s) involving fully modern humans). A common origin for Andaman (and other) Negrito populations, Melanesians and Aboriginal Australians was initially proposed on the basis of 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 the ancestries of Indian populations[8]. However, a 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 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. These features match known adaptations due to insularity in many groups of large animals, which may explain the 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 ~15× coverage (**Supplementary Note**), including 60 individuals from mainland India and 10 individuals from the Jarawa (JAR) and Onge (ONG) populations in the Andaman Islands (**Supplementary Fig. 1** and **Supplementary Table 1**). The demographically small and historically isolated Andamanese population showed higher relatedness among individuals as well as higher inbreeding coefficients and longer runs of homozygosity than all the continental Indian populations examined (**Supplementary Figs. 2–4**). In agreement with previous studies[8,11], principal-component analysis (PCA) showed that the Andamanese constituted a genetically distinct cluster in comparison with the mainland Indian populations (**Supplementary Fig. 5**). Interestingly, the Jarawa and Onge clustered tightly together, indicative of their genomic homogeneity, and showed a lack of recent admixture (**Fig. 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 are descendants from 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, *D*-statistic analysis[13] (**Supplementary Fig. 6**) showed that Andamanese share more alleles with each of the OOA populations than with sub-Saharan Africans, suggesting that the Andamanese have a common ancestry with all other OOA populations. Second, TreeMix analysis[14] also supported Africans as an outgroup to all OOA populations (**Fig. 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

[1]Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, Spain. [2]Servei de Genòmica, Universitat Pompeu Fabra, Barcelona, Spain. [3]BGI Shenzhen, Shenzhen, China. [4]Computational Biology, Target Sciences, GSK R&D, GlaxoSmithKline, Stevenage, UK. [5]Department of Internal Medicine, Radboud University Medical Center, Nijmegen, the Netherlands. [6]Departament de Genètica i de Microbiologia, Universitat Autonòma de Barcelona, Bellaterra, Spain. [7]National Institute of BioMedical Genomics, Kalyani, India. [8]Leverhulme Centre for Human Evolutionary Studies, Department of Archaeology and Anthropology, University of Cambridge, Cambridge, UK. [9]These authors contributed equally to this work. [10]These authors jointly directed this work. Correspondence should be addressed to J.B. (jaume.bertranpetit@upf.edu) or P.P.M. (ppm1@nibmg.ac.in).

Received 31 December 2015; accepted 17 June 2016; published online 25 July 2016; doi:10.1038/ng.3621

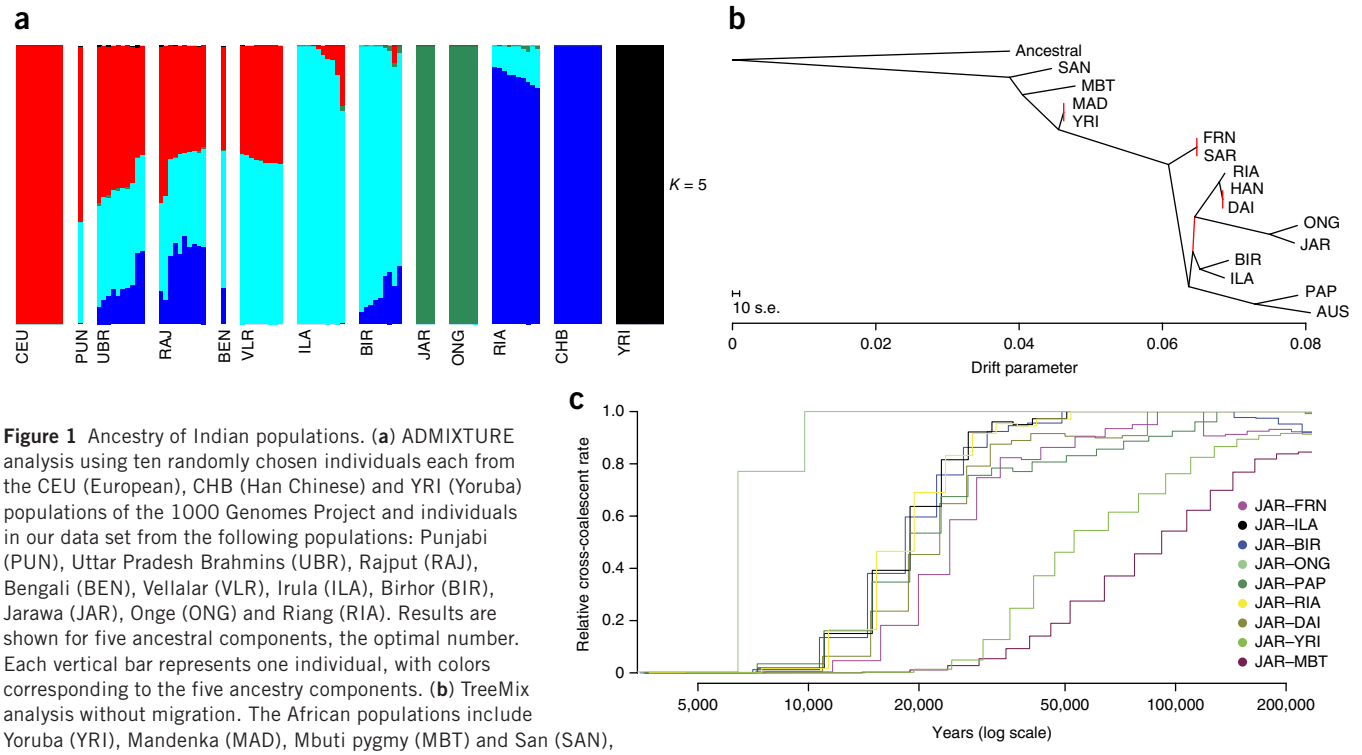


Figure 1 Ancestry of Indian populations. (**a**) ADMIXTURE analysis using ten randomly chosen individuals each from the CEU (European), CHB (Han Chinese) and YRI (Yoruba) populations of the 1000 Genomes Project and individuals in our data set from the following populations: Punjabi (PUN), Uttar Pradesh 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, with colors corresponding to the five ancestry components. (**b**) TreeMix analysis without migration. The African populations include Yoruba (YRI), Mandenka (MAD), Mbuti pygmy (MBT) and San (SAN), the European populations include French (FRN) and Sardinians (SAR), the East Asian populations include Dai (DAI) and Han Chinese (HAN), the Pacific populations include Papuans (PAP) and Aboriginal Australians (AUS), Indians include Birhor (BIR), Irula (ILA) and Riang (RIA), and Andamanese include Jarawa (JAR) and Onge (ONG). Inferred ancestral genome information from the 1000 Genomes Project was used as the outgroup. The scale bar shows ten units of standard error (s.e.), and the amount of drift is plotted along the *x* axis. Drift that is considered to be non-significant is indicated by a red line, resulting in three main branches (RIA, HAN and DAI; ONG and JAR; and BIR and ILA) that form a trichotomy. (**c**) MSMC relative cross-coalescent rate analysis showing the genetic separation between pairs of populations. In each curve, one individual was from the Jarawa population and the other was from a tribal population of India (ILA, BIR or RIA), the Onge population or a population outside India (FRN, DAI, PAP or YRI). The *x* axis shows time, and the *y* axis shows a measure of similarity for each pair of populations compared.

for Andamanese and any other OOA population, with these other populations showing similar split times among themselves (**Fig. 1c**). Estimation of historical effective population sizes by MSMC suggested a similar bottleneck event for the Andamanese and all other OOA populations at around 50,000 years ago (**Supplementary Fig. 7**). All of these results suggest that the Andamanese share their ancestry with all other OOA populations, indicating a commonality for all Asian and Pacific populations and consistent with a single, main OOA migration.

D-statistic analysis (**Supplementary Fig. 8**) showed that the Andamanese shared more alleles with East Asian, Papuan and mainland Indian tribal populations than they did with Europeans, indicating that Europeans are an outgroup for all Asian populations. Both TreeMix (**Fig. 1b**) and *D*-statistic outgroup analysis (**Supplementary Table 2**) supported this inference. Relative cross-coalescent analysis (**Fig. 1c**) also gave a similar result: in this analysis, the separation of Andamanese and Europeans predated the separation of Andamanese and Asians. Analysis using available ancient European genome sequences from La Braña, Loschbour and Stuttgart[16–18] supported our results (**Supplementary Figs. 9–11** and **Supplementary Table 3**), showing Europeans as the most distinct branch among all Eurasian and Pacific populations, even when considering the extinct Basal Eurasian component of Europeans[18,19]. Mitochondrial DNA analysis also supported a single origin for Asian populations (**Supplementary Table 4**).

Analysis of the contribution of extinct hominin populations to the current genetic pool also suggests a single origin for modern Asians, including the Andamanese. Andamanese genomes had a similar amount of Neanderthal[13,20] introgression as other OOA populations (~2–4%), suggesting that Neanderthal admixture took place at a very early stage, before the OOA populations separated from each other (**Supplementary Fig. 12**). In contrast, Papuans harbored a much higher proportion of Denisovan[21] ancestry than any other OOA population examined here (**Supplementary Fig. 13**); all other Asian populations examined (including the Andamanese) had only slightly more Denisovan ancestry than Europeans (**Supplementary Fig. 14**), as previously suggested[20]. Besides this, no other difference in ancient genomic contributions was observed between the Andamanese and the other South and East Asian and Pacific populations.

We found that Andamanese, mainland Indian and Papuan populations carried ~2–3% fewer African alleles than Europeans (**Fig. 2a**) or East Asians (**Fig. 2b**), as was also the case for Aboriginal Australians (similar yet higher proportion of unshared alleles). We performed extensive simulations to show that this reduction in the proportion of African alleles in Andamanese could not be explained by the low effective population size of Andamanese; thus, the reduction is not caused by private variants resulting from specific mutations in the Andamanese genome (no-admixture model; **Supplementary Table 5**), by later admixture between European or Asian and African populations (that is, the reduction cannot be due to a 'back-to-Africa' event; **Supplementary Table 5** and **Supplementary Note**) or by admixture with the modern humans from the initial OOA event settling in Eurasia. In contrast, the reduction in the proportion of African alleles in Andamanese could be caused by admixture with a population that diverged from modern humans at least 300,000 years ago (**Supplementary Fig. 15**). In fact, an

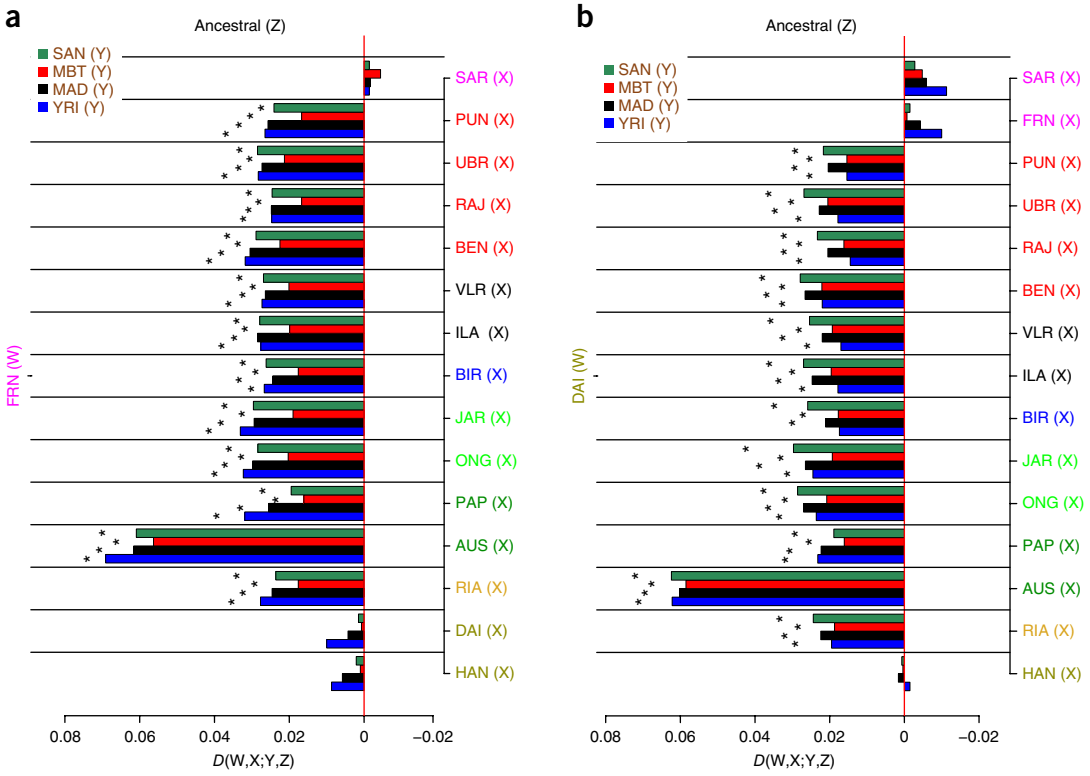


Figure 2 Fewer African-derived alleles in Indians, Andamanese, Papuans and Aboriginal Australians than in Europeans and East Asians. Each horizontal bar shows the result of *D*-statistic analysis of the form *D*(W,X;Y,Z), where the W population is either French or East Asian Dai. The X population is from India (Punjabi, Uttar Pradesh Brahmin, Rajput, Bengali, Vellalar, Irula, Birhor or Riang); Andamanese (Jarawa or Onge); or French, Sardinian, Dai, Han Chinese, Papuan or Aboriginal Australian. The Y population is African (Yoruba, Mandenka, Mbuti pygmy or San). Ancestral allele information from the 1000 Genomes Project is used as an outgroup (Z population). The populations are color-coded: pink, European; dark yellow, East Asian; brown, African; red, Indo-European; black, Dravidian; blue, Austroasiatic; light green, Andamanese; yellow, Tibeto-Burman; dark green, Pacific Islander and Australian Aboriginal. A positive value means that the W and Y populations share more derived alleles with each other than with the X and Y populations, whereas a negative value means that the X and Y populations share more derived allele with each other than with the W and Y populations. Statistically significant results (in this case defined by a *z* score greater than 3 or less than −3) are marked with an asterisk. (**a**) *D*-statistic results for *D*(French (W), X; African (Y), ancestral (Z)). (**b**) *D*-statistic results of *D*(Dai (W), X; African (Y), ancestral (Z)).

introgression from any hominin population could cause a bias in the *D*-statistic calculations (**Supplementary Note**), which would generate a false two-wave OOA signal (for modern humans) corresponding to the South Asian and Pacific populations. The reduction in African ancestry for South Asian populations likewise cannot have originated from Neanderthal or Denisovan introgression, as Andamanese and East Asian populations have similar amounts of well-recognized ancestry for these two populations. An alternative hypothesis is that the 2–3% reduction in 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] confirmed this hypothesis (**Supplementary Fig. 16** and **Supplementary Note**). By calculating *D* statistics for 50-kb regions with a sliding window, we inferred that this unknown population diverged from Neanderthals and Denisova before they diverged from each other, as seen initially with TreeMix (**Supplementary Fig. 17**). To further identify specific DNA regions derived from this hominin population, we implemented Sstar[24] on these putative fragments and detected ~15 Mb of sequence for each individual (average region length of 65 kb) derived from this hominin population, which behaves either as a sister group to Neanderthals and Denisova or even diverged earlier (**Supplementary Figs. 18** and **19**). For Aboriginal Australians, the deficit in African alleles was even higher (~6–7%; **Fig. 2**), suggesting that the reduction might be caused by admixture with some unknown ancient hominin population; this possibility needs to be confirmed with additional Australian data. Rasmussen *et al.*[4] suggested that Aboriginal Australians are the descendants of admixture between the first OOA population and later OOA populations. We were unable to detect this first OOA event either by *D*-statistic analysis (**Supplementary Tables 6** and **7**) or MSMC relative cross-coalescent analysis (**Supplementary Fig. 20**). Our simulations suggest that the bias in *D*-statistic calculation, which was interpreted as the product of admixture between the first OOA population and Aboriginal Australians, can instead be explained by admixture of an ancient hominin population with Aboriginal Australians.

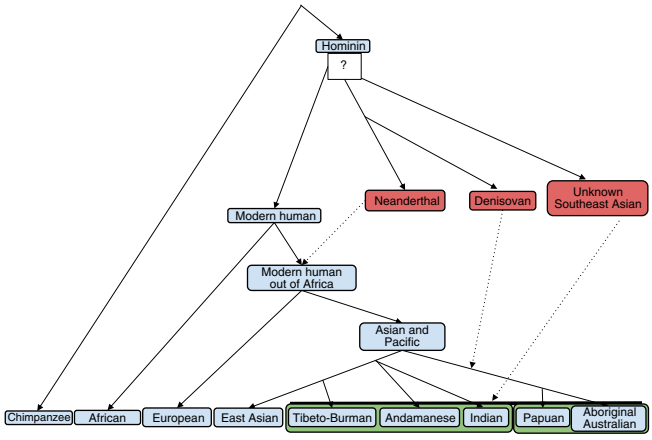
To explain the genetic structure of mainland India, it has been suggested[8] that all populations have arisen from admixture between two components: (i) ancestral North Indian (ANI) and (ii) ancestral South Indian (ASI), which is genetically related to Andamanese. However, although ADMIXTURE analysis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

Figure 3 Model of gene flow in Asia. Red boxes represent extinct non-African hominins who introgressed into modern humans; these introgressions are marked with dotted lines. Green boxes represent populations that may have admixed with the new unknown hominin. Andamanese and Indian populations are analyzed fully here; others will have to be further studied in the future. Properly solving the trichotomy (question mark) will require more data.

in contrast to what has traditionally been assumed because of morphological similarities between these populations[1].

As 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 that had 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 these, we found a significant excess of genes related to body morphology, with signals in 11 of the 107 genes for complete selective sweeps (Yates χ^2 = 5.70, P = 0.02) and 40 of the 509 genes for incomplete sweeps (Yates χ^2 = 9.495, P = 0.0021) related to height (according to the Genetics Association Database, GAD[27]). Other regions under positive selection included genes related to obesity or body shape and composition. 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 from individual genotypes at height-related SNPs; **Supplementary Fig. 21**); this selection 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 the Pacific, with two new insights (**Fig. 3**): (i) Asian populations, including ones from the Pacific, correspond to a single origin and OOA expansion, sharing a more recent common ancestor among themselves than with Europeans (our analyses do not support the hypothesis of two independent OOA events, postulated a long time ago on the basis of physical appearance[1] and seemingly 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% (admixture is higher in Australians, but this estimate needs to be confirmed with new data). Our results do not indicate whether the introgression is derived from the same hominin in all populations, but in the case of the Andamanese (**Supplementary Fig. 22**) we have shown that it comes from a new unknown hominin population, which likely separated very early in the hominin tree. Also, we have shown that the hominin admixture in these populations can cause a bias in D-statistic calculation that can be erroneously interpreted as a first OOA migration. Finally, the distinctive morphology of the Andamanese probably has originated from strong adaptive selection, as demonstrated by the excess of genes under selection related to height and body mass, leading to possibilities of understanding the basic biology of a complex adaptation to an island environment.



URLs. Picard tools, http://picard.sourceforge.net/, Broad Institute ftp server, ftp://ftp.broadinstitute.org/, 1000 Genome Project ancestral alignments file, http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The whole-genome sequences (Andamanese vcf files) have been deposited in the European Nucleotide Archive under accession PRJEB11455.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

J. Nye and C. Tyler-Smith kindly corrected the manuscript in depth. Thanks are given to R.A. Foley for discussion and inspiring input for **Figure 3**. 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 provided by grant BFU2013-43726-P from the Ministerio de Economía 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).

AUTHOR CONTRIBUTIONS

M.M., F.C., P.P.M. and J.B. conceived and designed the project. P.P.M. provided the samples. P.P.M., T.X. and Q.L. sequenced samples and carried out initial analyses. M.M. performed the remaining genetic data analyses. F.C., G.M.D., M.P., M.G.N., D.C., H.L., P.P.M. and J.B. participated in and discussed analyses. M.M., F.C., P.P.M. and J.B. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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. USA* **113**, 1594–1599 (2016).
12. Dass, F. *The Andaman Islands* (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. III *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. USA* **108**, 15123–15128 (2011).
24. Vernot, B. & Akey, 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**, 3946–3952 (2015).
27. Becker, K.G., Barnes, K.C., Bright, T.J. & Wang, S.A. The genetic association database. *Nat. Genet.* **36**, 431–432 (2004).

ONLINE METHODS

Samples. In total, 70 samples were collected from ten Indian populations from different geographical regions, linguistic affiliations and social categories (**Supplementary Table 1**). The ten populations were Punjabi (PUN), Uttar Pradesh upper caste Brahmin (UBR), Rajput (RAJ), Bengali (BEN), Vellalar (VLR), Irula (ILA), Birhor (BIR), Jarawa (JAR), Onge (ONG) and Riang (RIA). Blood and saliva samples were collected with voluntary informed consent from the participants. More information on the populations can be found in Basu *et al.*[11]. The study protocol was approved by the institutional ethics committees of the Indian Statistical Institute (the primary institution of P.P.M. when the samples were collected and anonymized), the National Institute of Biomedical Genomics and, for the Universitat Pompeu Fabra, the Comitè Ètic d'Investigació Clínica del Pac de Salut MAR.

Additional samples were also used to understand Indian populations from a global perspective. We used 1000 Genomes Project Phase 1 data[28], Great Ape Genome Project (GAGP) data[29], high-coverage data from three Aboriginal Australians[30], and high-coverage data for nine Yoruba (YRI) 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 used the 1000 Genomes Project ancestral alignment file[32] to identify ancestral alleles.

Sequencing. Whole-genome sequencing was performed in two different places (at BGI Shenzhen and the National Institute of BioMedical Genomics (NIBMG)) using Illumina technology. Fifty of the 70 samples were sequenced at BGI Shenzhen, and 20 were sequenced at NIBMG (**Supplementary Tables 1** and **8**). Sequencing libraries with an insert size of ~500 bp were constructed, and paired-end reads were generated on the HiSeq 2000 platform. Raw sequencing reads were mapped to the hg19 reference genome using BWA[33]. Duplicates were removed with Picard tools. We followed best-practice recommendations from GATK 2.8-1 (ref. 34), using IndelRealigner and BaseRecalibrator with their default values. For IndelRealigner, we used 1000 Genomes Project 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 and Mills and 1000 Genomes Project gold-standard indels to assign a well-calibrated probability to each variant; all these files were downloaded from the Broad Institute ftp site (5 November 2013) as described on the GATK website. The average coverage for autosomes was ~15×, and the proportion of the genome that was accessible was close to 100% (**Supplementary Table 8**). Although sequencing was performed at two different institutes, PCA and ADMIXTURE analysis (**Supplementary Note**) demonstrated very tight clustering for samples from the same population, suggesting that influences from the two sequencing centers were not detectable. To check the quality of the data and to detect the power of our inference, we performed various tests (**Supplementary Figs. 23–38** and **Supplementary Tables 9–17**); some of them are described briefly here, but for further details refer to the **Supplementary Note**.

Relatedness, inbreeding and homozygosity runs. Relatedness was calculated using KING[35] software with 13,679,600 autosomal biallelic SNPs. Inbreeding was calculated by vcftools[36] using the same SNPs and default parameters. Runs of homozygosity were detected by PLINK v1.07 (ref. 37) using 4,475,795 autosomal biallelic unlinked SNPs with 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 SNPs and a variance inflation factor of 2.

Principal-component analysis. SmartPCA from the EIGENSOFT package[38] was used for PCA. We kept only autosomal biallelic SNPs that had a minor allele frequency (MAF) of at least 0.05. We also removed SNPs that had missing information for any individual. Only ten individuals per population from 1000 Genomes Project data were kept to avoid sample size bias.

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

MSMC analysis. 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 with SHAPEIT[40].

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

TreeMix. TreeMix[14] was used to analyze 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 LD, we used the -k flag. LD blocks were defined as 1 Mb in length, which in our case corresponded 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 (**Supplementary Note**).

Dadi and the 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 (ref. 44) with the 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% 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 that was 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: (i) Tajima's D (ref. 46), (ii) CLR[47], (iii) Fay and Wu's H (ref. 48), (iv) Fu and Li's D (ref. 49), (v) XP-EHH[50], (vi) ΔiHH[51], (vi) iHS[51] and (viii) EHH average[52]. After calculating all tests, we ran the boosting algorithm[26] using parameters from both the East Asian and European HB strategies (simulated under neutrality and under selection using cosi with the 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 were very similar for the HB strategy for any non-African population (**Supplementary Note**). Information about genes related to body size was obtained from GAD[27].

***D* statistics with sliding windows and Sstar.** To identify candidate introgressed regions from an unknown hominin, we calculated D statistics for each individual in 50-kb regions with a sliding window of 5 kb and retained regions where Andamanese had fewer African-derived alleles than Europeans or East Asians

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

where F is the allele frequency in population W, X, Y or Z.

We ran TreeMix on the putative introgressed regions (**Supplementary Note**) and Sstar[24] to refine the identification of the introgressed hominin haplotypes, thus only choosing regions that were positive for both D statistics by sliding windows and Sstar (**Supplementary Note**).

doi:10.1038/ng.3621

28. 1000 Genomes Project Consortium. 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. 1000 Genomes Project Consortium. 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., Marchini, J. & 1000 Genomes Project Consortium 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. USA* **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**, e1000695 (2009).
45. Delaneau, O., Marchini, J. & Zagury, J.-F. Alinear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
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**, e72 (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).