nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-019-0881-7

Ancient admixture from an extinct ape lineage into bonobos

Martin Kuhlwilm[1*], Sojung Han[1], Vitor C. Sousa[2,3], Laurent Excoffier[3,4] and Tomas Marques-Bonet[1,5,6,7*]

Admixture is a recurrent phenomenon in humans and other great ape populations. Genetic information from extinct hominins allows us to study historical interactions with modern humans and discover adaptive functions of gene flow. Here, we investigate whole genomes from bonobo and chimpanzee populations for signatures of gene flow from unknown archaic populations, finding evidence for an ancient admixture event between bonobos and a divergent lineage. This result reveals a complex population history in our closest living relatives, probably several hundred thousand years ago. We reconstruct up to 4.8% of the genome of this 'ghost' ape, which represents genomic data of an extinct great ape population. Genes contained in archaic fragments might confer functional consequences for the immunity, behaviour and physiology of bonobos. Finally, comparing the landscapes of introgressed regions in humans and bonobos, we find that a recurrent depletion of introgression is rare, suggesting that genomic incompatibilities arose seldom in these lineages.

A picture of complex and recurrent interactions in humans and their extinct relatives emerged after the initial discovery of gene flow from Neandertals[1]—notably, from other hominins into modern humans[2–8], between Neandertals, Denisovans and other lineages[9], and from humans into Neandertals[10,11]. Although introgressed haplotypes are often deleterious on the human background[12,13], admixture seems to have been beneficial in some cases[14,15]. Unlike for the human lineage, fossils are rare for great apes. Since the split from hominins, which is possibly represented by fossils close to the common ancestor such as *Sahelanthropus*[16], only chimpanzee fossils of an age of ~0.5 million years ago (Ma) have been described[17].

However, signatures of admixture have been found in genomic data between different great ape populations[18,19], and might be common in other primate taxa[20]. Ancient gene flow from bonobos into chimpanzees, probably more than 200,000 years ago, has been described previously[21], but it is possible that these species of the *Pan* clade might have experienced further historical events of gene flow that have remained hidden from us so far. Knowledge about the divergence of chimpanzees and bonobos, and the range and habitat of proto-*Pan* populations, is not conclusive, particularly since it is unclear when and to what extent the Congo River has been a natural barrier[22,23]. It seems likely that the ancestors of bonobos separated from the ancestors of chimpanzees by crossing a reduced Congo River during a dry glacial period ~1.7 Ma, rather than by the formation of the river itself[23,24], which may date back to 4 Ma[25]. Episodes of migration and gene flow might have happened during different glaciation periods, when river levels were low enough to provide windows of opportunity for crossing.

Here, we apply methods developed to identify introgression in the absence of ancient genomes[7,26]—either based on demographic modelling or an excess of private variation (Supplementary Fig. 1)—to the whole genomes of 69 chimpanzee and bonobo individuals, to explore archaic gene flow using present-day variation. Western and central chimpanzees (*Pan troglodytes verus* and *Pan troglodytes troglodytes*, respectively) are the two chimpanzee populations that differ the most from each other, both regarding the amount of gene flow with bonobos and their effective population sizes[18,21,27]. Hence, our main analysis focuses on these two groups, together with their sister species, bonobos (*Pan paniscus*).

Results

Gene flow between *Pan* populations. To detect introgressed genomic regions between species, we first computed the S^* statistic, which reflects the amount and physical proximity (linkage disequilibrium) of private variation compared with a divergent reference panel, and has been used to infer signatures of gene flow in humans[3,28–31] and to identify introgressed genomic segments[5,7]. We performed these calculations as implemented elsewhere[7], but in a pairwise manner, testing each individual of the test population independently, with one of the two other populations as reference panels (Methods). Based on the results from a given reference, we could predict the expected S^* for the other population using a generalized linear model and also detect outlier regions that we consider to be due to past introgression. In central chimpanzees, we find an unexpected sharing of private variation with bonobos (Supplementary Fig. 3), in agreement with gene flow from bonobos into non-western chimpanzees[21].

To verify that S^* outlier regions correctly detect introgression, we confirmed that they overlap more than expected with a previous screen for introgressed bonobo-like segments[32]. Both methods identify only a small proportion of the genome as introgressed (0.16 and 0.24%, respectively). We further compared the number of pairwise differences of single-nucleotide variants (SNVs)[8] between all individuals across all putatively introgressed windows, compared with the same number of randomly sampled windows. In agreement

[1]Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), Parc de Recerca Biomèdica de Barcelona, Barcelona, Spain. [2]Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal. [3]Institute of Ecology and Evolution, University of Berne, Berne, Switzerland. [4]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [5]National Centre for Genomic Analysis–Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. [6]Institucio Catalana de Recerca i Estudis Avançats, Barcelona, Spain. [7]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain. *e-mail: martin.kuhlwilm@upf.edu; tomas.marques@upf.edu

with gene flow between species, we find that bonobo-like windows in central chimpanzees carry, on average, 1.75-fold more such differences to other chimpanzee individuals than random regions (Supplementary Fig. 4). Moreover, in these regions, chimpanzees show a closer affinity to bonobos in a principal component analysis (PCA; Supplementary Fig. 7 and Supplementary Data) and in a phylogenetic tree (Fig. 1a,b and Supplementary Figs. 5 and 6).

To quantify the historical levels of gene flow and compare the likelihood of models with and without migration between chimpanzees and bonobos, we used a site frequency spectrum (SFS)-based composite likelihood method[33], as described in detail previously[21]. We find support for gene flow between chimpanzees and bonobos, as those models fit the SFS data better (Supplementary Table 2), coherent with previous, more complex models[21] (Methods). These models, as well as the S^* analysis (Supplementary Fig. 3), might also support ancestral bidirectional gene flow (that is, from chimpanzees into bonobos), although it remains difficult to discern the relationship of the introgressing population with the extant chimpanzees (Supplementary Information). Indeed, segregating sites across putative chimpanzee-like windows in bonobos do not show a different topology, suggesting that this analysis might be confounded by other factors; for example, high-frequency bonobo-like fragments in chimpanzees (Supplementary Information). Furthermore, we find 3.5–5.0% of windows to be unexpectedly similar between the central and western chimpanzee populations (Supplementary Fig. 3), which might be the result of genetic exchange between these subspecies, in agreement with previous results[18,19,21,27,34].

Archaic admixture in bonobos. We then tested these populations for a signature of archaic introgression from an unknown source outside the known tree. Following the methodology developed to identify archaic fragments in human genomes[5,7], we determined outlier windows with unexpectedly high S^*. We used a simplification of the SFS-based demographic model with single pulses of migration between chimpanzees and bonobos as the null model for the extant *Pan* history (Methods). This model was used to simulate the expected distribution of S^* (Supplementary Information) and to detect windows in which S^* deviates from expectation when analysing the data with each of the two reference populations, given the respective numbers of segregating sites. We find that ~1% of windows in the bonobo genomes behave as outliers in S^* (Supplementary Fig. 3), but not in any of the chimpanzee populations, indicating a signature of putative archaic admixture.

We compared the pairwise SNV differences between individuals in random regions and putative archaic regions (that is, outlier S^* regions in bonobos). These should correlate across all individual comparisons across all populations if systematic features (for example, higher mutation rates) caused the signal. However, we found that the differences between any bonobo and any chimpanzee are elevated by 1.94-fold in putative archaic introgressed windows in bonobos, while the numbers of pairwise SNV differences between chimpanzees are similar between these same test and random regions (Fig. 2a). We conclude that these regions show random variation within chimpanzees, but an increased difference between chimpanzees and bonobos. The pairwise SNV differences between the putative introgressed windows in the test bonobo and other bonobo individuals are elevated by 37% when compared with random regions. As expected, segregating sites in these windows form a longer branch in a phylogenetic tree (Fig. 1c and Supplementary Figs. 13 and 14) and explain ~60% more of the variance in a PCA (Fig. 2b–d). Furthermore, bonobos start to separate from each other in principal component 7 (1.63% of the variance), which is not observed for random regions up to principal component 20 (Supplementary Fig. 15). Even though these windows seem to strongly deviate from the overall species divergence, the difference between bonobo individuals is not as pronounced, consistent with

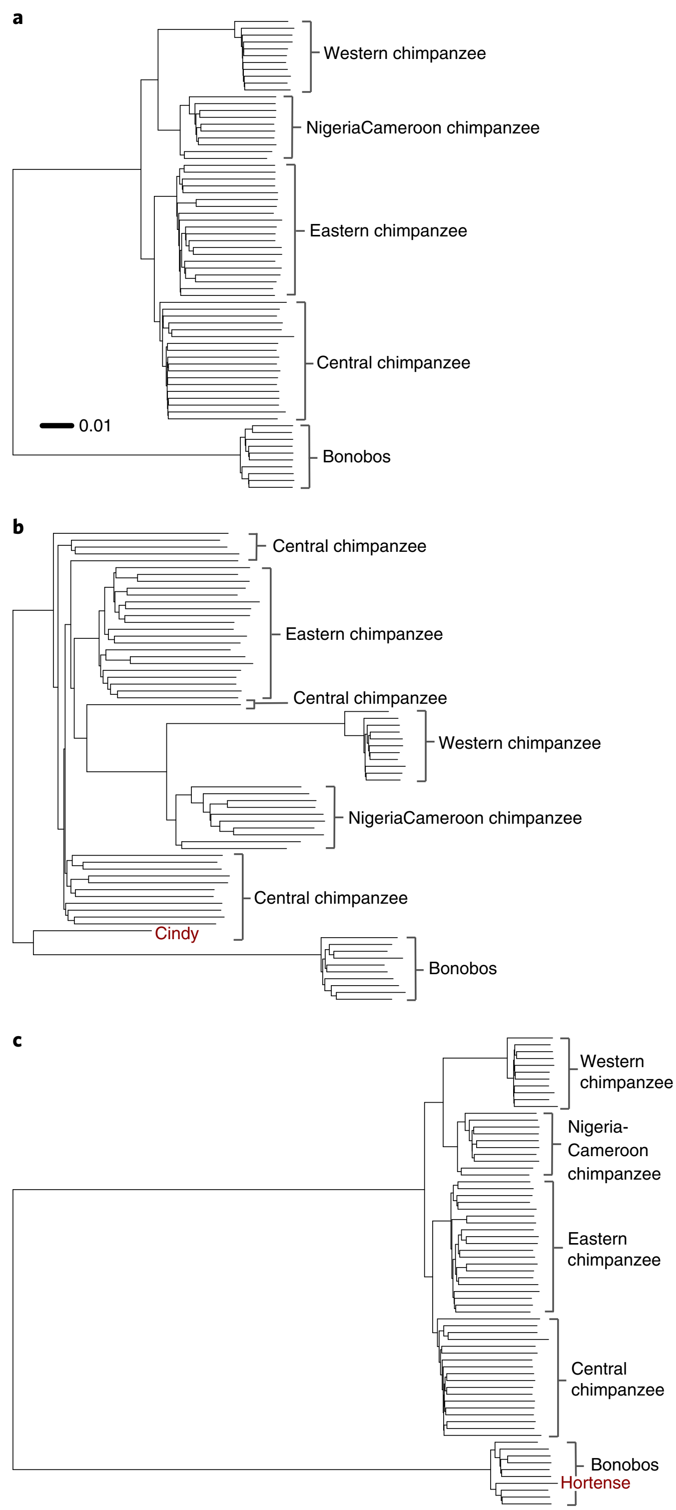

Fig. 1 | Trees of putatively introgressed fragments. a–c, Neighbour-joining trees drawn to the same scale. **a**, Random fragments across the genome, representing the average phylogeny. **b**, Windows with bonobo-like introgression in a specific central chimpanzee (Cindy). **c**, Windows with putative archaic introgression in a specific bonobo individual (Hortense).

genetic drift after an ancient gene flow event. Haplotype networks of these windows typically show a large distance between bonobos and chimpanzees that is often similar to the distance between both and modern humans (Supplementary Fig. 16), but we also find segregating haplotypes where most bonobos form a cluster, while few individuals show a distance larger than that of the bonobo cluster to chimpanzees (Fig. 2e).

To compare demographic models and infer parameters, we used two approaches: (1) SFS-based modelling; and (2) approximate Bayesian computation (ABC) with neural networks based on genome-wide statistics (Methods). The ABC approach aims to use the underlying window-based data and linkage disequilibrium information from all high-coverage genome sequences, and hence complements SFS-based analyses. As summary statistics, the mean values and standard deviations of the number of segregating sites, the pairwise S^* statistic and the percentage of outlier windows were used (Supplementary Table 5). The topology of the tree was inferred with the SFS-based model, and parameters for past and current population sizes as well as migration rates were randomly sampled, while divergence times were fixed (Methods). The ABC-based demographic inference without archaic gene flow provided estimates very similar to the SFS-based model, notably including support for gene flow between the extant *Pan* populations (Supplementary Table 4). We used this demographic model as a refined null model to recalculate the generalized linear model of expected S^* distributions. Again, simulations under this model could not recover the excess of archaic outliers found in the bonobo genomes (Supplementary Table 4 and Supplementary Figs. 17 and 18).

We then used ABC-based modelling to infer the demographic parameters of a model with archaic gene flow. First, we inferred the population parameters of all populations. In a second step, we refined the inference for bonobo-specific parameters, together with the amount and time of archaic gene flow, while fixing the other parameters and assuming a fixed archaic population divergence at 3.5 Ma. Finally, we also inferred the divergence of the archaic population (Supplementary Fig. 1 and Supplementary Information). The resulting fine-tuned estimates indicate that bonobos received 0.9–4.2% from an unknown archaic population (Fig. 3). Simulations performed under this model can replicate the excess of outlier windows observed in the real data, while simulations without this gene flow cannot replicate this pattern (Supplementary Fig. 19). An ABC-based model selection test shows the largest support for the fine-tuned model with archaic gene flow (Fig. 4a; posterior probability = 0.98; Bayes factor > 60) and low levels of misclassification (<0.001%; Supplementary Fig. 20). Applying this ABC-based approach to the other chimpanzee populations (eastern and Nigeria–Cameroon chimpanzees) generally confirms these observations, without evidence for additional gene flow events (Supplementary Information). However, we note that the methods applied here might not be sensitive enough to discover gene flow events to a much smaller extent.

Historical population structure in bonobos after the split from chimpanzees is unlikely to cause signatures as observed here. In such a scenario, some bonobo individuals would appear more closely related to chimpanzees. Here, we observe haplotypes where all bonobos appear equally distinct from either all chimpanzees or all chimpanzees and other bonobos. The scenario of gene flow suggested here might resemble population structure before the split of chimpanzees and bonobos, with subsequent isolation of only the chimpanzee lineage. This is not supported by the models of population history inferred here, and seems unlikely in the biogeographical context of the separation of the *Pan* clade[22–24]. The SFS-based modelling of archaic gene flow (Supplementary Table 2) also suggests that a model with archaic gene flow of 0.03–6.87% (95% confidence interval (CI)) has a higher likelihood; hence, it provides a better fit to the data than models without such gene flow, or with ancient substructure of the ancestral bonobo population (Fig. 4b). Finally, the signature is not driven by possible confounding factors, such as differences in transitions or transversions, or copy number variants (Supplementary Information).

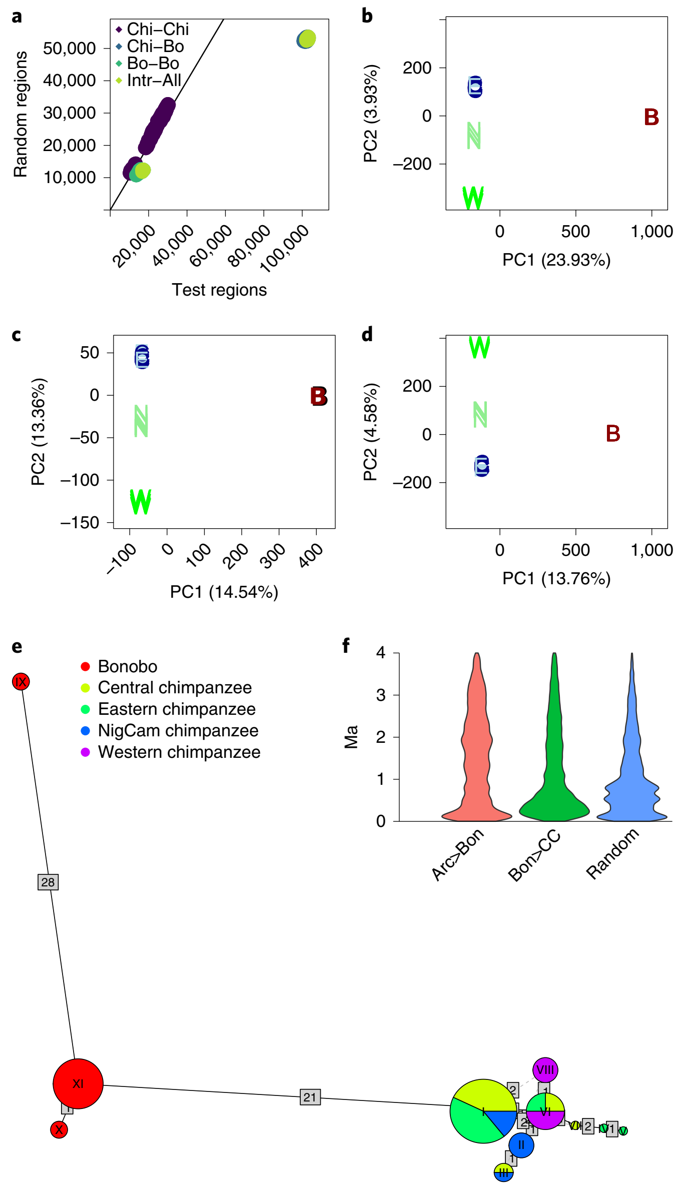

Fig. 2 | Analysis of putatively introgressed windows. a, Number of base pair differences (Δbp) between all pairs of individuals in this study, for putative archaic introgressed windows in bonobos using S^* (*x* axis), and in the same number of random windows (*y* axis). Bo, bonobo; Chi, chimpanzee; Intr–All, comparison between the individual for whom the S^* test was performed and all other individuals. Δbp between bonobos and chimpanzees is larger in these windows than in random windows (top right, green for the test individual and blue for all other individuals in the same regions), suggesting elevated genetic distance. **b**, PCA[71] for SNPs in windows with putative archaic introgression in any bonobo. **c**, PCA for SNPs in windows with putative archaic introgression in a specific bonobo individual (Hortense). **d**, PCA for SNPs in random windows, drawn on the same scale as in **b**). Note that the *y* axis is flipped since this is calculated from different SNPs. **e**, Haplotype network[72] of one archaic fragment in bonobos (chromosome 8, region 30,599,999–30,670,000 bp), representative of haplotypes still segregating in the population. Haplotypes in chimpanzees form one cluster, most bonobos form a distinct cluster, and one haplotype in bonobos (IX) falls outside their distribution. For fixed haplotypes, see Supplementary Fig. 16. NigCam, Nigeria–Cameroon. **f**, Inferred age distribution[37] of SNVs falling in putative archaic windows in bonobos (Arc>Bon) and bonobo-like windows in central chimpanzees (Bon>CC) compared to random windows. Archaic windows carry an excess of SNVs older than 2 Ma in archaic windows.

Alternative inference of gene flow. Since S^* relies on the demographic model, previous assumptions on the population history might influence the results. To confirm our observations, we used a recently developed method for detecting introgression without assumptions about the demographic history[26]. This method works in the absence of ancient genomes, although in humans the available ancient genomes were used to confirm the robustness of this method. This hidden Markov model (henceforth termed 'Skov HMM') detects unexpected densities of private sites in small segments of 1,000 base pairs (bp) in a given individual (Methods and Supplementary Information). When applying this method in a setting without gene flow, this results in significantly lower likelihood than in a setting with one gene flow event (Fig. 4c; $P = 0.9 \times 10^{-5}$, Wilcoxon rank test). This supports the existence of two distinct classes of genomic regions in bonobos, one of which represents a *Pan*-like state, and a smaller fraction of the genome being more divergent. After decoding[26] and filtering archaic regions for posterior probabilities >0.9, we identify 74.2–107.1 megabase pairs (Mbp) of archaic fragments for the individual genomes (2.6–3.7% per individual, covering 4.8% of the genome in total) (Supplementary Table 9 and Supplementary Fig. 25). We call 30% more archaic fragments when using only western chimpanzees as a reference panel, possibly because gene flow between non-western chimpanzees and bonobos[21] interferes with this signal (Fig. 3).

Interestingly, we find that on average 60% of the significant regions in bonobos inferred using the S^* method overlap with the decoded Skov HMM regions (Supplementary Table 12). This is only 15% lower than in modern humans[26], where archaic genomes were available and used for validation. Thus, we conclude that this overlap reflects similar signatures of archaic gene flow in our data for bonobos, detected by both methods. The introgressed segments are short (mean: 12 kilobase pairs (kbp)), in agreement with an old gene flow event. Simulations suggest that the majority of short segments might not be detected here (Supplementary Information). Indeed, the mean length of correctly detected simulated fragments is ~17 kbp, but the mean length of missed archaic fragments is only ~9 kbp. Still, 85.8% (95% CI: 80.4–91.2%) of the detected segments are correctly inferred, and for simulations under a model without gene flow we do not detect false archaic segments with posterior probabilities >0.9. Thus, our observations are only replicated by simulations under a model with archaic gene flow, although a smaller difference of the divergence times, together with an older introgression age, will decrease both the precision and sensitivity compared with Neandertal introgression in modern humans.

An old event from an early-diverging lineage. We estimate a migration pulse at a time of 377–637 thousand years ago (ka) (95% credible interval; Fig. 3 and Supplementary Table 4) in the fine-tuned ABC-based model using S^* (Supplementary Information), which agrees well with an introgression time at 367–407 ka using the length distribution of introgressed fragments with the Skov HMM. We note that this model infers a single migration pulse to summarize the observations, while a longer migration period or several admixture pulses are possible scenarios as well. Additionally, SFS-based modelling suggests wide CIs, with an admixture event of 0.03–6.87% (95% CI) occurring at 466–1,627 ka (95% CI); hence, the above admixture times might be a lower-bound estimate. The split time of the archaic population is inferred at 3.3 Ma (95% credible interval: 2.89–3.75 Ma) using ABC modelling and 2.45–3.7 Ma (95% CI) using the SFS-based method. The coalescence time of the archaic fraction using the Skov HMM is inferred at 5.01–5.36 Ma (95% CI; Supplementary Table 8), which, as expected, is older than the actual population divergence time[26]. When applying the Skov HMM to data simulated under the ABC-based demographic model with a 3.3 Ma simulated divergence time, we obtain a raw emission value of 4.98 Ma. When correcting the coalescence time for

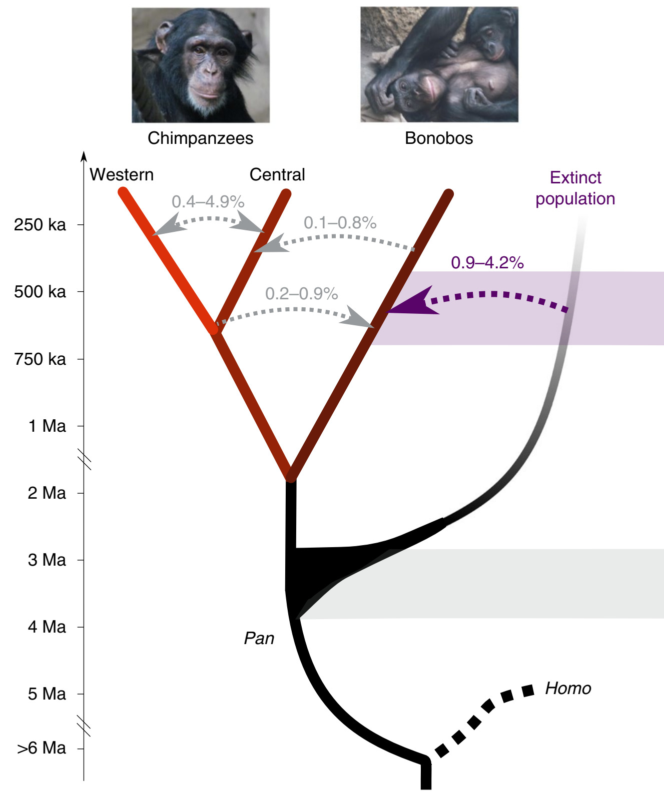

Fig. 3 | Model of population history in *Pan* species with archaic gene flow into bonobos. Simplified phylogenetic tree of central (*P. t. troglodytes*) and western chimpanzees (*P. t. verus*), bonobos (*P. paniscus*) and an unknown 'ghost' population. Grey arrows represent previously described gene flow events between chimpanzees and bonobos[21]. The violet arrow represents archaic gene flow into bonobos. The 95% credible intervals for introgression and archaic divergence times as well as introgression amounts are shown, as inferred using S^* with ABC modelling (Methods). The divergence times of extant *Pan* populations were inferred using SFS-based modelling (Methods).

the Skov HMM by a factor of 1.509, based on these simulations, the divergence time of ~3.32–3.55 Ma (95% CI) is well contained within the ABC- and SFS-based inferences. This tendency of higher time estimates is consistent with observations in humans, where the Skov HMM yields estimates of 853–984 ka for the coalescence with Neandertals, compared with 484–640 ka divergence times[10,35]. We note that these divergence times would be scaled to lower values under the assumption of a faster mutation rate in chimpanzees, as has been suggested recently[36].

Furthermore, the estimated age[37] of S^* SNVs in the significant windows shows an increase between 2.0 and 3.5 Myr (Fig. 2f), which is unusual compared with random regions of the genome ($P < 2.2 \times 10^{-16}$, Wilcoxon rank test). In conclusion, a divergence of the archaic population beyond 3 Ma seems well supported, with a population split time between bonobos and chimpanzees of probably not more than 2 Ma[21,38,39] (Fig. 3). We note that this divergence time might be slightly overestimated due to archaic gene flow. Interestingly, fragments inferred using both methods overlap with regions where bonobos fall outside the chimpanzee variation in a previous test for external regions on the chimpanzee lineage[38]. Since some of these regions might be the result of archaic admixture in bonobos rather than selection in chimpanzees, this might explain the unexpected absence of protein-coding genes in many of these regions[38].

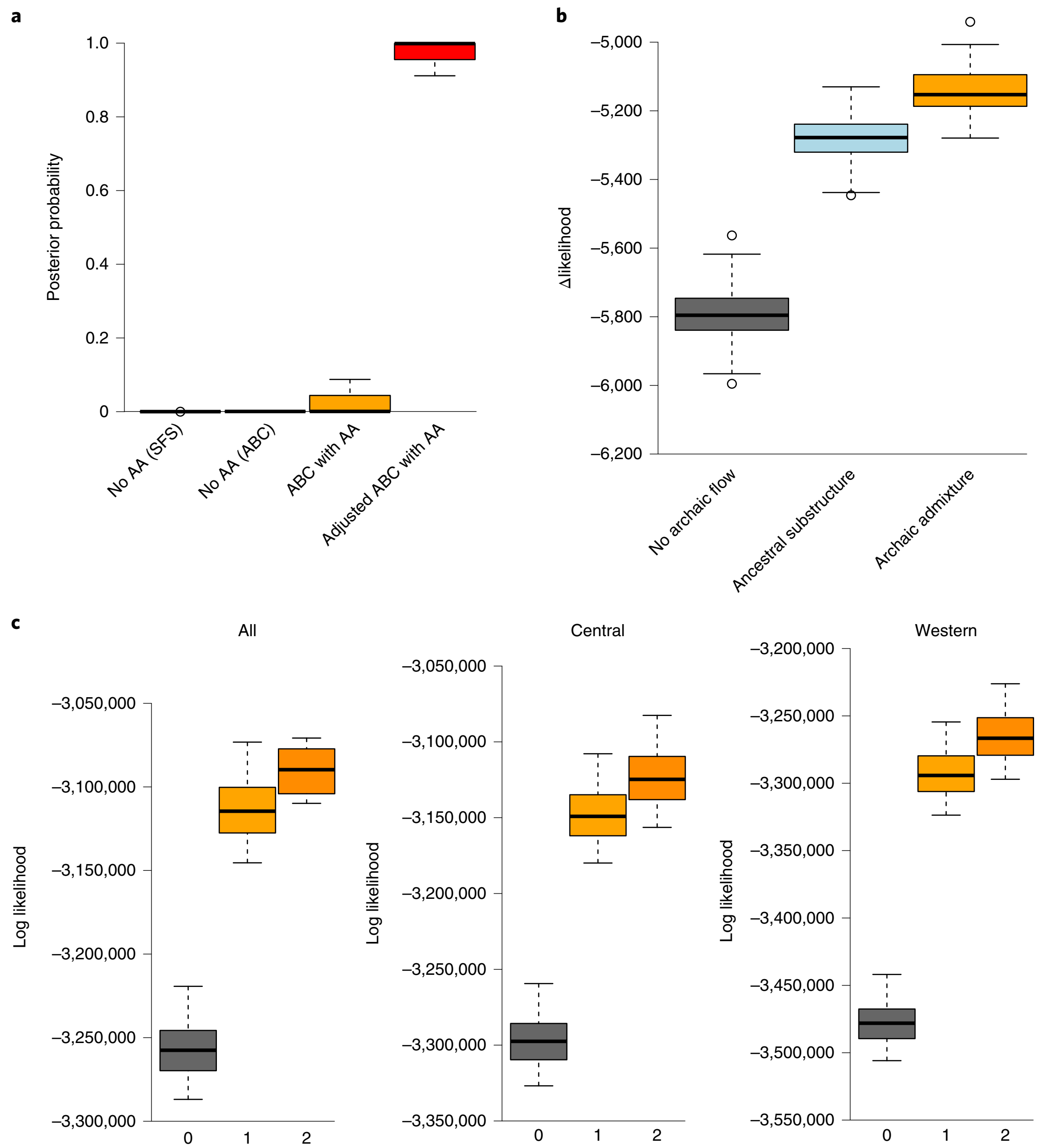

Fig. 4 | Posterior values of the models used. a, Posterior probabilities of 100 replicate tests for the ABC model selection test[63,67] for the simplified SFS-based demographic model, the ABC-based model without archaic admixture (AA) in bonobos, the ABC-based model with archaic admixture in bonobos and the adjusted ABC-based model with archaic admixture in bonobos. **b**, Differences between the likelihood of a constrained tree and the maximum-likelihood tree (Δlikelihood ($\log_{10}$)) of 100 replicates for the SFS-based model[21,33] with and without archaic admixture, and with ancient substructure in bonobos (Methods). **c**, log likelihoods for the Skov HMM[26] for ten bonobo individuals, assuming no gene flow (0), or one (1) or two gene flow events (2), using either all chimpanzees (left) or only central (middle) or western chimpanzees (right) as reference panels. In **a**–**c**, the central black lines are median values, the box edges represent upper and lower quartiles, and the whiskers represent the most extreme data point within 1.5 times the interquartile range from the box.

Landscape of introgression across the genome. In total, only ~3% of the autosomes shows a signature of archaic introgression. This partial archaic *Pan* genome is not evenly distributed across the chromosomes, with many regions carrying introgressed haplotypes in several or all individuals, while other regions are depleted (Fig. 5). Even though the archaic ghost population and the ancestral population of bonobos must have been able to produce fertile offspring, local incompatibilities may have led to regions of depleted introgression[40]. When applying S^* and the Skov HMM to the X chromosome (Supplementary Information), we find an eightfold reduction of archaic ancestry (Fig. 5). In humans, this chromosome shows a fivefold reduction for Neandertal introgression[13], suggesting a barrier to gene flow between populations both within the clades of *Homo*[10,13] and *Pan*[21], possibly due to recurrent selective sweeps[41].

We screened the autosomes for regions of reduced archaic ancestry (Supplementary Table 13), finding the largest proportions of putative introgression deserts in chromosomes 1, 17 and 19 (Fig. 5), among which chromosome 17 is known to carry the smallest proportion of introgression from archaic hominins into modern humans[42]. One of the largest depleted regions (chromosome 1; 109–125 Mbp) overlaps with a large archaic introgression desert in modern humans[7,13] (Fig. 5). Since in this region deficiencies in the gene *CSF1* lead to pregnancy loss in humans, possibly by foetal rejection[43], we speculate that a derived non-synonymous change in this gene on the bonobo lineage[44] might have had functional consequences leading to a rejection of archaic introgression. We find no protein-coding changes, but regulatory variants at high frequency on both the modern human and archaic lineages, respectively[9,45]

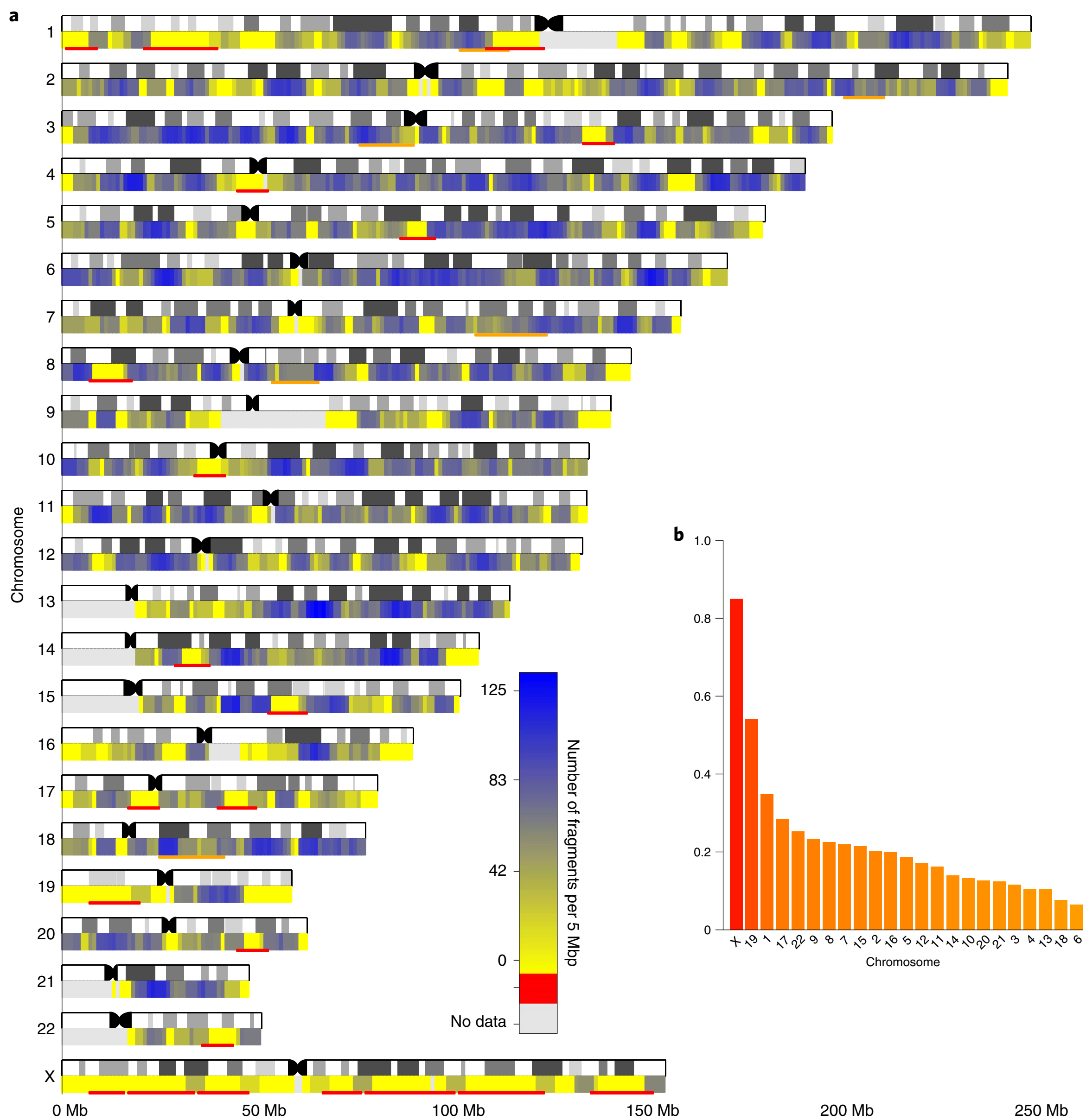

Fig. 5 | Distribution of introgression across the genome. a, Karyogram of human chromosomes showing the density of archaic fragments, calculated for the number of significant S^* fragments of 40 kbp across 10 bonobo individuals in sliding windows of 5 Mbp in 1-Mbp steps (Methods). Putative introgression deserts (>8 Mbp) are marked in red. Known introgression deserts in humans[7] are shown in orange. The colour bar shows number of fragments per 5 megabase pairs. **b**, Chromosomes ordered by the proportion covered by depleted regions (>5 Mbp).

(Supplementary Information). However, recurrent hybrid incompatibilities between populations arose rarely in these lineages.

Archaic fragments might be functionally relevant (Supplementary Information). We find an enrichment for genome-wide association study traits related to behavioural and sleep phenotypes (Supplementary Table 16), suggesting a potential role of introgression for unique behavioural features of bonobos[46], as well as 'iron biomarker measurement' in blood. Interestingly, a protein-coding change[44] in the gene encoding for erythrocyte membrane protein 42 (*EPB42*) falls within a known signature of positive selection in bonobos[47]. This gene appears to be downregulated in bonobos in the brain, cerebellum and kidney (adjusted $P < 0.05$)[48], and is the only putatively introgressed gene we find differentially expressed in as many as three tissues (Supplementary Table 20). This position is conserved across other mammals and located only three amino acids downstream of a missense mutation in humans causing haemolytic anaemia[49]. However, it is unclear how this mutation relates to past adaptations, considering that haematology values of captive present-day bonobos appear unremarkable[50]. Immune adaptation might be a possible explanation, similar to the well-described malaria-protective mutation in human haemoglobin, which causes sickle cell anaemia[51].

It is known that the retention of introgression in immunity-related genes conferred benefits[32,52], and we find that within the longest regions (Supplementary Table 7), *SERPINA11* and *SERPINA9* play a role in adaptive immunity[53] and carry protein-coding changes in bonobos[44]. Among other genes possibly involved in the immune response (Supplementary Information), the gene *VNN2*, encoding for a protein with a role in neutrophil migration[54], carries four protein-coding changes older than 2 Ma in bonobos (Supplementary Table 21). Introgression might have also played a role in ancient adaptation to food resources; for example, through protein-altering changes in the alcohol dehydrogenase-encoding gene *ADH4* (Supplementary Information). The functional consequences of these differences and their biological relevance need to be explored in future studies. Finally, 2 of the regions larger than 100 kbp (chromosome 10 (76,140,000–76,300,000 bp) with the *ADK* gene and chromosome 3 (144,450,000–144,580,000) without protein-coding genes) overlap with genome-wide outliers (top 0.5%) of the population differentiation statistic F_{ST} (Methods) and might have been under selection.

Discussion

A bonobo founder population probably diverged from chimpanzees <2 Ma by crossing the Congo River, followed by population retractions and expansions probably due to climatic changes[24]. It has been suggested that the deepest mitochondrial split dates to ~0.95 Ma[24], and bonobos spread westwards afterwards. It seems possible that bonobos encountered a distinct branch of the *Pan* clade during their expansion, with hybridization leaving the genomic traces discussed here. A separation of ancestral populations with the Congo River formation ~3.5 Ma or during later dry periods[23] may provide the context for an early population split from the *Pan* clade, which our results suggest has hybridized with the ancestral bonobo population (Fig. 3). It remains unclear how well the genetic diversity of bonobos is reflected by the available genomes, but mitochondrial data suggest that more genomic diversity may be found in the wild than is represented here[24]. Since it might well be that no ape fossils with preserved ancient DNA are to be found in the Congo Basin, excavating parts of extinct ape genomes from present-day variation could be the only strategy with which to explore these long-gone populations. By increasing the sample size for bonobos and other great apes using non-invasive samples[55], larger fractions of 'genomic fossils' may be uncovered, potentially providing more insights into the biology of extinct apes, as well as adaptation and incompatibilities in hominins.

Methods

Data and ancestral alleles. We used the genotypes of the individuals from a previous study[21], mapped to the human reference genome (hg19), using the 22 autosomes and the X chromosome. The data consist of genotype calls for 10 bonobo, 18 central chimpanzee, 20 eastern chimpanzee, 10 Nigeria–Cameroon chimpanzee and 11 western chimpanzee individuals (Supplementary Table 1). To avoid biases from the use of the chimpanzee reference genome in the ancestral allele inference provided by Ensembl[56], we used the macaque genome as an outgroup to infer the ancestral state. We lifted over the rhesus macaque reference genome (rheMac3) to the human genome coordinates using bedtools[57] and rtracklayer[58] in the R environment[59]. Finally, we modified scripts from the package freezing-archer[60] to create a custom ancestral binary genome file in which any site that is segregating in the dataset of the 69 individuals or different from hg19 is replaced by the macaque reference allele. This package contains scripts used in a previous study on archaic admixture in humans[7]. We used the R environment and the packages GenomicRanges[61] and bedr[62] for further data processing.

Implementation of *S**.** We used the package freezing-archer, which was also used for *S** implementation in previous studies on archaic introgression in modern humans[5,7]. We calculated *S** on a genome-wide scale with a window size of 40 kilobase pairs (kbp) and a window step of 30 kbp, for 11 western and 18 central chimpanzees and 10 bonobos, in windows where 3/4 of sites were considered 'callable' (that is, genotypes were retrieved in all individuals, as described by de Manuel et al.[21]), and at least 30 segregating sites were observed across all individuals considered. We calculated the statistic in a pairwise manner, testing each individual of the test population independently, with one population from each of the two other populations used as reference panels (Supplementary Information). The *S** for a given reference population was used to predict the *S** for the other reference population to detect outlier regions in a generalized linear model using the R package mgcv[63]. The normalized deviation from expectation for *S** in each window was used to detect windows in which an individual shows unusually large *S** for one reference panel but small *S** for the other reference panel (outside the 95% CI). We used null distributions of *S** from demographic models without gene flow (described below) and simulated data as described previously[5] to obtain a generalized linear model given the number of segregating sites. Briefly, we simulated[64] 20,000 windows of 40 kbp for predefined numbers of segregating sites from 25–700 in steps of 5, and obtained a generalized linear model, analogous to previous work[7]. Windows in which the empirical *S** was outside the 99% CI for 2 different reference populations were considered putatively introgressed from a source population unrelated to the reference populations (Supplementary Information). The longest regions were defined as consecutive overlapping windows in at least one individual. Regions of at least 5 Mbp in which at most 1 significant window in at most 1 individual was found, and where at least 1/3 of the windows contained data, were defined as putatively depleted regions. We note that the number of only ten bonobo individuals is a limitation of our dataset.

Statistical modelling. We performed demographic modelling and inference using two approaches: (1) SFS-based composite likelihoods; and (2) ABC based on *S** statistics. These approaches are complementary given that in the SFS all sites are assumed to be independent and linkage disequilibrium information is discarded, while the ABC-based analysis is able to use linkage disequilibrium information captured by the *S** statistics to infer introgression. All demographic estimates were done assuming a mutation rate of 1.2×10^{-8} (ref. [65]) and were rescaled into time (in years) assuming a generation time of 25 years[66].

We used the joint 3D-SFS of bonobo and western and central chimpanzees following the approach described in detail previously[21] to infer effective population sizes, split times and migration rates (Supplementary Information). The SFS was built based on 1,084 blocks of 1 Mbp on the autosomes[21], resulting in an SFS with a total of 763,965,527 sites without missing data, of which 4,839,432 were biallelic single-nucleotide polymorphisms (SNPs). The settings to run the fastsimcoal2[33] analyses were the same as described previously[21]. We further estimated the likelihood of models of increasing complexity (Supplementary Information) to test whether models with archaic gene flow between an unsampled ghost population and bonobo fitted the SFS data better than alternative models (without ghost population or ancestral population substructure in bonobos).

We performed modelling based on ABC[67] with neural networks. The initial null model for *S** was adjusted ad hoc to match the distribution of segregating sites in 40-kbp windows (Supplementary Information). For parameter estimates, we simulated 333 windows of 250 kbp for each random combination of effective population sizes and migration rates (Supplementary Information) as input, and used the numbers and standard deviations of segregating sites in 40-kbp windows, *S** values and proportions of outliers as summary statistics (Supplementary Table 5). Initial inferences were based on 45,000 simulations with a tolerance threshold of 0.01 to infer the best fit for effective population sizes and migration rates (Supplementary Table 4) without archaic gene flow, which was then defined as the new null model (ABC-based null model). The best fit for a model with archaic gene flow was also estimated from 90,000 simulations and a tolerance of 0.001. Finally, fine-tuned inferences for archaic divergence time and migration rates were obtained with the same parameters (Supplementary Fig. 1). When replicating the inference of demographic parameters using ABC for the model without archaic gene flow using the same procedure, we obtain very similar values for effective population sizes and migration rates (Supplementary Table 4). ABC modelling and *S** calculations were also applied to the genomes of 20 eastern and 10 Nigeria–Cameroon chimpanzees, with ~10,000 simulations for each (tolerance: 0.05). The ABC model selection test was performed on the adjusted SFS-based model, the best ABC-based model without gene flow, the best ABC-based model with archaic gene flow and a fixed archaic divergence time of 3.5 Ma, and the adjusted ABC-based model with archaic gene flow. We obtained ~6,200 simulations of 333 fragments of 250 kbp, and applied the neural networks method with a tolerance threshold of 0.05.

Implementation of the Skov HMM. We used the Skov HMM on private sites in a given individual[26] (Supplementary Information), implemented in the introgression-detection package. Briefly, we calculated the numbers of callable sites in 1-kbp windows, SNV density and numbers of private variants in each individual for the 22 autosomal chromosomes and the X chromosome. We applied settings[26] without gene flow, or with one or two gene flow events. Starting probabilities were set to (0.95, 0.05) and (0.95, 0.035, 0.015) for one and two gene flow events, respectively. The transition matrices were ((0.999, 0.001),(0.01, 0.99)) and ((0.998, 0.001, 0.0001),(0.0195, 0.98, 0.0005),(0.012, 0.012, 0.975)), and the emission matrices were (0.05, 1.0) and (0.1, 0.7, 1.5), respectively. We tested the chimpanzee and bonobo individuals with all individuals from the respective other species as a reference panel, and bonobos compared with western and central chimpanzees separately. The decoding was performed as provided by the package, at a probability cutoff of 0.9 and with a minimum number of 5 private sites to call introgressed fragments. For time estimates, we used a mutation rate of 1.2×10^{-8} mutations generation^{-1} bp^{-1}, and a constant recombination rate of 0.7×10^{-8} generation^{-1} bp^{-1}, considering lower recombination rates in *Pan* species than humans[68]. Example conversions are shown in Supplementary Table 10. Simulations were performed using msprime[69] under the fine-tuned ABC-based model using *S** (see above). The coalescence time of the archaic fraction to all chimpanzees is inferred at 5.01–5.36 Ma. Since this coalescence time is older than the split time and dependent on the effective population size, it may serve as a proxy for the divergence time, but it is not identical to the split time. When applying the Skov HMM to simulated data with a divergence time of 3.3 Ma between species, the estimate from the emission probability is 4.98 Ma. We suggest that this coalescence time can be converted to divergence time through a factor of 1.509.

Other analyses. Pairwise differences of SNVs were calculated with a similar approach as used in a previous study[8], between all individuals in a pairwise fashion across all significant windows, and for the same number of randomly sampled regions. Analyses of SNV differences, phylogenetic trees[70], PCAs[71] and significance tests were performed in the R environment[59] (Supplementary Information). Haplotype networks from all SNPs in the archaic fragments were built using the package pegas[72]. The results from the program ARGweaver[37] as applied and described previously[21] were re-analysed, and allele age was estimated with 'arg-summarize –A'. Information on functional changes was retrieved from previous studies on public data[44,45] (Supplementary Information), and an enrichment test for genome-wide association study traits was performed as described elsewhere[44,73] (Supplementary Information). We mapped and quantified chimpanzee and bonobo transcriptome data[48] using the reference genome hg19 (refs. [74,75]), and tested for differential gene expression between the two species using DESeq2 (ref. [76]) (Supplementary Table 20). We calculated the genome-wide distribution of F_{ST} between bonobos and chimpanzees in windows of 40 kbp, with 10-kbp steps, using PopGenome[77]. Phylogenetic trees were drawn using phangorn[70] with Kimura's distance[78]. More details and additional analyses are described in the Supplementary Information.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequence data from a previous study are publicly available under the accession code PRJEB15086 at the European Nucleotide Archive. Genotype data are available at http://biologiaevolutiva.org/tmarques/data/. Data pertaining to the results are in the Supplementary Information.

Received: 1 October 2018; Accepted: 21 March 2019;
Published online: 29 April 2019

References

1. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
2. Reich, D. et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
3. 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).
4. Meyer, M. et al. A high coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
5. Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017–1021 (2014).
6. Fu, Q. et al. An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
7. Vernot, B. et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239 (2016).
8. Xu, D. et al. Archaic hominin introgression in Africa contributes to functional salivary MUC7 genetic variation. *Mol. Biol. Evol.* **34**, 2704–2715 (2017).
9. Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
10. Kuhlwilm, M. et al. Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**, 429–433 (2016).
11. Posth, C. et al. Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nat Commun.* **8**, 16046 (2017).
12. Juric, I., Aeschbacher, S. & Coop, G. The strength of selection against Neanderthal introgression. *PLoS Genet.* **12**, e1006340 (2016).
13. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr. Biol.* **26**, 1241–1247 (2016).
14. Huerta-Sanchez, E. et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).
15. Racimo, F., Marnetto, D. & Huerta-Sánchez, E. Signatures of archaic adaptive introgression in present-day human populations. *Mol. Biol. Evol.* **34**, 296–317 (2017).
16. Brunet, M. et al. A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418**, 145–151 (2002).
17. McBrearty, S. & Jablonski, N. G. First fossil chimpanzee. *Nature* **437**, 105–108 (2005).
18. Prado-Martinez, J. et al. Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013).
19. Hey, J. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Mol. Biol. Evol.* **27**, 921–933 (2010).
20. Tung, J. & Barreiro, L. B. The contribution of admixture to primate evolution. *Curr. Opin. Genet. Dev.* **47**, 61–68 (2017).
21. De Manuel, M. et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477–481 (2016).
22. Kawamoto, Y. et al. Genetic structure of wild bonobo populations: diversity of mitochondrial DNA and geographical distribution. *PLoS One* **8**, e59660 (2013).
23. Takemoto, H., Kawamoto, Y. & Furuichi, T. How did bonobos come to range south of the Congo river? Reconsideration of the divergence of *Pan paniscus* from other *Pan* populations. *Evol. Anthropol.* **24**, 170–184 (2015).
24. Takemoto, H. et al. The mitochondrial ancestor of bonobos and the origin of their major haplogroups. *PLoS One* **12**, e0174851 (2017).
25. Myers Thompson, J. A. A model of the biogeographical journey from proto-*Pan* to *Pan paniscus*. *Primates* **44**, 191–197 (2003).
26. Skov, L. et al. Detecting archaic introgression using an unadmixed outgroup. *PLoS Genet.* **14**, e1007641 (2018).
27. Won, Y.-J. J. & Hey, J. Divergence population genetics of chimpanzees. *Mol. Biol. Evol.* **22**, 297–307 (2005).
28. Plagnol, V. & Wall, J. D. Possible ancestral structure in human populations. *PLoS Genet.* **2**, e105 (2006).
29. Wall, J. D., Lohmueller, K. E. & Plagnol, V. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.* **26**, 1823–1827 (2009).
30. Hsieh, P. et al. Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. *Genome Res.* **26**, 291–300 (2016).
31. Lachance, J. et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150**, 457–469 (2012).
32. Nye, J. et al. Selection in the introgressed regions of the chimpanzee genome. *Genome Biol. Evol.* **10**, 1132–1138 (2018).
33. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
34. Hey, J. et al. Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.* **35**, 2805–2818 (2018).
35. Prüfer, K. et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).
36. Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T. & Schierup, M. H. Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nat. Ecol. Evol.* **3**, 286–292 (2019).
37. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* **10**, e1004342 (2014).
38. Prüfer, K. et al. The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**, 527–531 (2012).
39. Kuhlwilm, M. et al. Evolution and demography of the great apes. *Curr. Opin. Genet. Dev.* **41**, 124–129 (2016).
40. Sankararaman, S. et al. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014).
41. Nam, K. et al. Extreme selective sweeps independently targeted the X chromosomes of the great apes. *Proc. Natl Acad. Sci. USA* **112**, 6413–6418 (2015).
42. Steinrücken, M., Spence, J. P., Kamm, J. A., Wieczorek, E. & Song, Y. S. Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. *Mol. Ecol.* **27**, 3873–3888 (2018).
43. Piccinni, M.-P. T cells in normal pregnancy and recurrent pregnancy loss. *Reprod. Biomed. Online* **13**, 840–844 (2006).
44. Han, S., Andrés, A. M., Marques-Bonet, T. & Kuhlwilm, M. Genetic variation in *Pan* species is shaped by demographic history and harbors lineage-specific functions. *Genome Biol. Evol.* https://doi.org/10.1093/gbe/evz047 (2019).
45. Kuhlwilm, M. & Boeckx, C. Genetic differences between humans and other hominins contribute to the "human condition". Preprint at https://www.biorxiv.org/content/10.1101/298950v1 (2018).
46. Furuichi, T. Social interactions and the life history of female *Pan paniscus* in Wamba, Zaire. *Int. J. Primatol.* **10**, 173–197 (1989).
47. Cagan, A. et al. Natural selection in the great apes. *Mol. Biol. Evol.* **33**, 3268–3283 (2016).
48. Brawand, D. et al. The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
49. Bouhassira, E. E. et al. An alanine-to-threonine substitution in protein 4.2 cDNA is associated with a Japanese form of hereditary hemolytic anemia (protein 4.2NIPPON). *Blood* **79**, 1846–1854 (1992).
50. Loomis, M. R. in *Zoo and Wild Animal Medicine* 5th edn (eds Fowler, M. E. & Miller, R. E.) 381–397 (Saunders (Elsevier Science), 2003).
51. Cyrklaff, M. et al. Hemoglobins S and C interfere with actin remodeling in *Plasmodium falciparum*-infected erythrocytes. *Science* **334**, 1283–1286 (2011).
52. Dannemann, M., Andrés, A. M. & Kelso, J. Introgression of Neandertal- and Denisovan-like haplotypes contributes to adaptive variation in human toll-like receptors. *Am. J. Hum. Genet.* **98**, 22–33 (2016).
53. Frazer, J. K. et al. Identification of centerin: a novel human germinal center B cell-restricted serpin. *Eur. J. Immunol.* **30**, 3039–3048 (2000).

54. Suzuki, K. et al. A novel glycosylphosphatidyl inositol-anchored protein on human leukocytes: a possible role for regulation of neutrophil adherence and migration. *J. Immunol.* **162**, 4277–4284 (1999).
55. Hernandez-Rodriguez, J. et al. The impact of endogenous content, replicates and pooling on genome capture from faecal samples. *Mol. Ecol. Resour.* **18**, 319–333 (2017).
56. Paten, B. et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843 (2008).
57. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
58. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
59. R Core Development Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2015).
60. Vernot, B. freezing-archer (2016), GitHub repository, https://github.com/bvernot/freezing-archer.
61. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
62. Haider, S. et al. A bedr way of genomic interval processing. *Source Code Biol. Med.* **11**, 14 (2016).
63. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B* **73**, 3–36 (2011).
64. Hudson, R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
65. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
66. Langergraber, K. E. et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl Acad. Sci. USA* **109**, 15716–15721 (2012).
67. Csilléry, K., François, O. & Blum, M. G. B. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3**, 475–479 (2012).
68. Stevison, L. S. et al. The time-scale of recombination rate evolution in great apes. *Mol. Biol. Evol.* **33**, 928–945 (2016).
69. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).
70. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
71. Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).
72. Paradis, E. pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* **26**, 419–420 (2010).
73. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
74. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
75. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
76. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
77. Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936 (2014).
78. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).

Acknowledgements

We thank A. M. Andres, B. Vernot and L. J. Kuhlwilm for comments and discussion, and M. de Manuel for help with the data. M.K. was supported by a DFG fellowship (KU 3467/1-1). V.C.S. was supported by the Fundação para a Ciência e a Tecnologia (project UID/BIA/00329/2013) and EU Horizon 2020 programme (Marie Skłodowska-Curie grant 799729). L.E. was supported by the Swiss National Science Foundation (number 310030B-166605). T.M.-B. was supported by MINECO BFU2014-55090-P (FEDER), a U01 MH106874 grant, the Howard Hughes International Early Career programme, Obra Social 'La Caixa' and Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya.

Author contributions

M.K., S.H., V.C.S. and L.E. analysed the data. M.K. and T.M.-B. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-019-0881-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.K. or T.M.

nature research

Corresponding author(s): Martin Kuhlwilm, Tomas Marques-Bonet

Last updated by author(s): Mar 7, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	Statistical analysis: R, introgression-detection, freezing-archer; Simulations: ms, fastsimcoal2, msprime

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence data from a previous study is publicly available under the accession code PRJEB15086 at ENA, and genotype data under http://biologiaevolutiva.org/tmarques/data/, results are in the Supplementary Information.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Published data on wild individuals was used
Data exclusions	No data was excluded.
Replication	Not applicable.
Randomization	Samples were grouped by species/subspecies.
Blinding	Not relevant, published data was used.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging