

1 Ancient admixture from an extinct ape lineage into bonobos

2
3 Martin Kuhlwilm^{1*}, Sojung Han¹, Vitor C. Sousa^{2,3}, Laurent Excoffier^{3,4}, Tomas Marques-
4 Bonet^{1,5,6,7*}

5
6 ¹ Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Doctor Aiguader 88, Barcelona,
7 08003, Spain.

8 ² cE3c - Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências, Universidade de
9 Lisboa, 1749-016 Lisboa, Portugal

10 ³ Institute of Ecology and Evolution, University of Berne, 3012 Berne, Switzerland

11 ⁴ Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

12 ⁵ CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology
13 (BIST), Baldri i Reixac 4, 08028 Barcelona, Spain

14 ⁶ Institutio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia 08010, Spain.

15 ⁷ Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Columnes s/n,
16 08193 Cerdanyola del Vallès, Spain

17 *Correspondence to: martin.kuhlwilm@upf.edu; tomas.marques@upf.edu

18
19
20
21 Admixture emerges to be a recurrent phenomenon in humans and other great ape populations.
22 Genetic information from extinct hominins allows to study historical interactions with modern
23 humans and discover adaptive functions of gene flow. Here, we investigate whole genomes from
24 bonobo and chimpanzee populations for signatures of gene flow from unknown archaic populations,
25 finding evidence for an ancient admixture event between bonobos and a divergent lineage. This
26 result exposes a complex population history in our closest living relatives, most likely several
27 hundred thousand years ago. We excavate ~3% of the genome of this “ghost” ape, which represents
28 the first genomic data of an extinct great ape population. Genes contained in archaic fragments
29 might confer functional consequences for immunity, behavior and physiology of bonobos. Finally,
30 comparing the landscape of introgressed regions in humans and bonobos shows that a recurrent
31 depletion for introgression is rare, suggesting that genomic incompatibilities arose seldom in these
32 lineages.

33
34
35 A picture of complex and recurrent interactions in humans and their extinct relatives emerged after
36 the initial discovery of gene flow from Neandertals¹, notably from other hominins into modern
37 humans^{2–8}, between Neandertals, Denisovans and other lineages⁹, 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, possibly represented by fossils close to the common ancestor like *Sahelanthropus*¹⁶, only chimpanzee fossils of an age of ~0.5 Mya (million years ago) have been described¹⁷.

However, signatures of admixture have been found in genomic data between different great ape populations^{18,19}, and might be common in other primate taxa²⁰. Ancient gene flow from bonobos into chimpanzees most likely more than 200,000 years ago has been described previously²¹, but it is possible that these species of the *Pan* clade might have experienced further historical events of gene flow, hidden to us so far. The 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 which 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 at ~1.7 Mya, rather than by the formation of the river itself^{23,24}, which may date back up to 4 Mya²⁵. 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 modeling or an excess of private variation (Fig. S1), to the whole genomes of 69 chimpanzee and bonobo individuals, in order to explore archaic gene flow by using present-day variation. Western and central chimpanzees (*Pan troglodytes verus* and *Pan troglodytes troglodytes*) 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 is focusing on these two groups, together with their sister species bonobos (*Pan paniscus*).

Results

Gene flow between *Pan* populations

In order to detect introgressed genomic regions between species, we first computed the S^* statistic, which reflects the amount and physical proximity (linkage disequilibrium, LD) of private variation when compared to 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 elsewhere⁷, 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 detect outlier regions which we consider to be due to past introgression. In central chimpanzees, we find an unexpected sharing of private variation with bonobos (Fig. S3), in agreement with gene flow from bonobos into non-western chimpanzees²¹.

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 segments³², and 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 to the same number of randomly sampled windows. In agreement 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 (Fig. S4). Moreover, in these regions chimpanzees show a closer affinity to bonobos in a principal component analysis (PCA; Fig. S7 & Extended Data), and in a phylogenetic tree (Fig. 1a-b & S5-6).

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 method³³, as described in detail previously²¹. We find support for gene flow between chimpanzees and bonobos, as those models fit better the SFS data (Table S2), coherent with previous, more complex models²¹ (Methods). These models as well as the S^* analysis (Fig. S3) might also support ancestral bi-directional gene flow, *i.e.* 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% of windows to be unexpectedly similar between the central and western chimpanzee populations (Fig. S3), 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 detect windows in which S^* deviates from expectation when analyzing 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^* (Fig. S3), 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, *i.e.* outlier S^* regions in bonobos. These should correlate across all individual comparisons across all populations if systematic features (*e.g.* 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 the other bonobo individuals are elevated by 37% when compared to random regions. As expected, segregating sites in these windows form a longer branch in a phylogenetic tree (Fig. 1c & S13-14) and explain ~60% more of the variance in a PCA (Fig. 2b-d). Furthermore, bonobos start to separate from each other in PC7 (1.63% of the variance), which is not observed for random regions up to PC20 (Fig. S15). Even though these windows seem to be strongly deviating from the overall species divergence, the difference between bonobo individuals is not as pronounced, consistent with genetic drift after an ancient gene flow event. Haplotype networks of these windows typically show a large distance between bonobos and chimpanzees, often similar to the distance of both to modern humans (Fig. S16), 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: (i) SFS-based modeling, and (ii) 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 LD information from all high-coverage genome sequences, 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 (Table S5). 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 (Table S4). 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 (Table S4, Fig. S17-18).

We then used ABC-based modeling to infer the demographic parameters of a model with archaic gene flow. We first inferred the population parameters of all populations, in a second step 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 Mya. Finally, we also inferred the divergence of the archaic population (Fig. S1, Supplementary Information). The resulting finetuned 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 (Fig. S19). An ABC-based model selection test shows the largest support for the finetuned model with archaic gene flow (Fig. 4A; posterior probability = 0.98, Bayes factor > 60), and low levels of misclassification (<0.001%, Fig. S20). 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 either equally distinct from all chimpanzees or from 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* clade²²⁻²⁴. The SFS-based modeling of archaic gene flow (Table S2) 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 like differences in transitions or transversions, or copy number variants (Supplementary Information).

Alternative inference of gene flow

Since S^* relies on the demographic model, prior 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 history²⁶. 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 basepairs (bp) in a given individual (Methods, Supplementary Information). When applying this method in a setting without gene flow, this results in significantly lower posterior probabilities 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 decoding²⁶ and filtering archaic regions for posterior probabilities > 0.9 , we identify 74.2-107.1 Mbp of archaic fragments for the individual genomes (2.6-3.7% per individual, covering in total 4.8% of the genome) (Table S9, Fig. S25). We call 30% more archaic fragments when using only western chimpanzees as reference panel, possibly because gene flow between non-western chimpanzees and bonobos²¹ 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 (Table S12). This is only 15% lower than in modern humans²⁶, 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 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% (80.4-91.2%, 95% CI) 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 precision and sensitivity when compared to Neandertal introgression in modern humans.

An old event from an early diverging lineage

We estimate a migration pulse of 0.9-4.2% at a time of 377-637 kya (95% Credible Interval (CrI); Fig. 3, Table S4) in the finetuned ABC-based model using S^* (Supplementary Information), which agrees well with an introgression time at 367-407 kya 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 modeling suggests wide confidence intervals, with an admixture event of 0.03-6.87% (95% CI) occurring at 466-1,627 kya (95% CI), hence that the above admixture times might be a lower bound estimate. The split time of the archaic population is inferred at 3.3 Mya (2.89-3.75 Mya, 95% CrI) using ABC modeling and 2.45-3.7 Mya (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 Mya (95% CI; Table S8), as expected older than the actual population divergence time²⁶. When applying the Skov HMM to data simulated under the ABC-based demographic model with 3.3 Mya simulated divergence time, we obtain a raw emission value of 4.98 Mya. This tendency of higher time estimates is consistent with observations in humans, where the Skov HMM yields estimates of 853-984 kya for the coalescence with Neandertals, compared to 484-640 kya divergence times^{10,35}. When correcting the coalescence time for the Skov HMM by a factor of 1.509, the divergence time of ~3.32-3.55 Mya (95% CI) is well contained within the ABC- and SFS-based inferences.

Furthermore, the estimated age³⁶ of S^* SNVs in the significant windows shows an increase between 2.0 and 3.5 million years (Fig. 2f), which is unusual in comparison 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 Mya seems well supported, with a population split time between bonobos and chimpanzees of likely not more than 2 Mya^{21,37,38} (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 lineage³⁷. 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 regions³⁷.

The 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 population and the ancestral population of bonobos must have been able to produce fertile offspring, local incompatibilities may lead to regions of depleted introgression³⁹. When applying *S** and the Skov HMM to the X chromosome (Supplementary Information), we find an 8-fold reduction of archaic ancestry (Fig. 5). In humans, this chromosome shows a 5-fold reduction for Neandertal introgression¹³, suggesting a barrier to gene flow between populations both within the clades of *Homo*^{10,13} and *Pan*²¹, possibly due to recurrent selective sweeps⁴⁰. We screened the autosomes for regions of reduced archaic ancestry (Table S13), finding the largest proportion 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 humans⁴¹. One of the largest depleted regions (chr1:109-125 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 fetal rejection⁴², we speculate that a derived non-synonymous change in this gene on the bonobo lineage⁴³ 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^{9,44} (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 GWAS traits related to behavioral and sleep phenotypes (Table S16), suggesting a potential role of introgression for unique behavioral features of bonobos⁴⁵, as well as “iron biomarker measurement” in blood. Interestingly, a protein-coding change⁴³ in the gene encoding for Erythrocyte Membrane Protein 42 (*EPB42*) falls within a known signature of positive selection in bonobos⁴⁶. This gene appears to be downregulated in bonobos in brain, cerebellum and kidney (adjusted P-value < 0.05)⁴⁷, the only putatively introgressed gene we find differentially expressed in as many as three tissues (Table S20). This position is conserved across other mammals, and only three amino acids downstream of a missense mutation in humans causing hemolytic anemia⁴⁸. However, it is unclear how this mutation relates to past adaptations, considering that haematology values of captive present-day bonobos appear unremarkable⁴⁹. Immune adaptation might be a possible explanation, similar to the well-described malaria-protective mutation in human hemoglobin, which causes sickle cell anemia⁵⁰.

It is known that the retention of introgression in immunity-related genes conferred benefits^{32,51}, and we find that within the longest regions, *SERPINA11* and *SERPINA9* play a role in adaptive immunity⁵² and carry protein-coding changes in bonobos⁴³. Among other genes possibly involved in the immune response (Supplementary Information), the gene *VNN2*, encoding for a protein with a role in neutrophil migration⁵³, carries four protein-coding changes older than 2 Mya in bonobos (Table S21). 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, two out of the regions larger than 100 kbp (Fig. 5) overlap with genome-wide outliers (top 0.5%) of the F_{ST} statistic and might have been under selection.

Discussion

A bonobo founder population diverged from chimpanzees most likely less than 2 Mya by crossing the Congo river, followed by population retractions and expansions likely due to climatic changes²⁴. It has been suggested that the deepest mitochondrial split dates to ~0.95 Mya²⁴, 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 Mya or during later dry periods²³ 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 suggests that more genomic diversity may be found in the wild than represented here²⁴. 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 for to explore these long-gone populations. By increasing the sample size for bonobos and other great apes using non-invasive samples⁵⁴, 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 study²¹, mapped to the human reference genome (*hg19*), using the 22 autosomes and the X chromosome. The data consists of genotype calls for 10 bonobo, 18 central chimpanzee, 20 eastern chimpanzee, 10 Nigeria-Cameroon chimpanzee and 11 western chimpanzee individuals (Table S1). In order to avoid biases from the use of the chimpanzee reference genome in the ancestral allele inference provided by Ensembl⁵⁵, 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*⁵⁶ and *Rtracklayer*⁵⁷ in the *R* environment⁵⁸. Finally, we modified scripts from the package *freezing-archer*⁵⁹ 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 humans⁷. We used the *R* environment and the packages *GenomicRanges*⁶⁰ and *bedr*⁶¹ for further data processing.

Implementation of S^* . We used the package *freezing-archer*, 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 kilo-basepairs (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” (*i.e.* genotypes were retrieved in all individuals, as described in de Manuel *et al.*, 2016²¹), 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*⁶². 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 previously⁵ to obtain a generalized linear model given the number of segregating sites. Briefly, we simulated⁶³ 20,000 windows of 40 kbp for predefined numbers of segregating sites from 25 to 700 in steps of 5, and obtained a generalized linear model, analogous to previous work⁷. Windows in which the empirical S^* was outside the 99% CI for two different reference populations were considered putatively introgressed from a source population unrelated to the reference populations (Supplementary Information). Longest regions were defined as consecutive overlapping windows in at least one individual. Regions of at least 5 mega-basepairs (Mbp) in which at most one significant window in at most one 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 10 bonobo individuals is a limitation of our dataset.

Statistical modeling. We performed demographic modeling and inference using two approaches: (i) site frequency spectrum (SFS)-based composite likelihoods, and (ii) approximate Bayesian computation (ABC) based on S^* statistics. These approaches are complementary given that in the SFS all sites are assumed to be independent and LD information is discarded, while the ABC-based analysis is able to use LD information captured by the S^* statistics to infer introgression. All demographic estimates were done assuming a mutation rate of 1.2×10^{-8} ⁶⁴ and were re-scaled into time in years assuming a generation time of 25 years ⁶⁵.

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 (mega-basepairs) 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 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 modeling based on Approximate Bayesian Computation (ABC) ⁶⁶ with neural networks. The initial Null model for S^* was adjusted *ad hoc* in order 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 as well as the proportions of outliers as summary statistics (Table S5). 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 (Table S4) without archaic gene flow, which was then defined as 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, and finally fine-tuned inferences for archaic divergence time and migration rates were obtained with the same parameters (Fig. S1). 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 (Table S4). ABC modeling 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 Mya, 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, the SNV density, and the numbers of private variants in each individual, for the 22 autosomal chromosomes and the X

chromosome. We applied settings²⁶ without gene flow, one and 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 matrix was $\begin{bmatrix} 0.999 & 0.001 \\ 0.01 & 0.99 \end{bmatrix}$ and $\begin{bmatrix} 0.998 & 0.001 & 0.0001 \\ 0.0195 & 0.98 & 0.0005 \\ 0.012 & 0.012 & 0.975 \end{bmatrix}$, the emission matrix [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 reference panel, and bonobos compared to western and central chimpanzees separately. The decoding was performed as provided by the package, at a probability cutoff at 0.9 and with a minimum number of five private sites to call introgressed fragments. For time estimates we used a mutation rate of 1.2×10^{-8} mutations per generation per bp, and a constant recombination rate of 0.7×10^{-8} per generation per bp, considering lower recombination rates in *Pan* species than humans⁶⁷. Example conversions are shown in Table S10. Simulations were performed using *msprime*⁶⁸, under the finetuned ABC-based model using S^* (see above). The coalescence time of the archaic fraction to all chimpanzees is inferred at 5.01-5.36 Mya. Since this coalescence time is older than the split time and depending on the effective population size, it may serve as a proxy for the divergence time, but is not identical to the split time. When applying the Skov HMM to simulated data with a divergence time of 3.3 Mya between species, the estimate from the emission probability is 4.98 Mya. We suggest that this coalescence time can be converted to divergence time through a factor of 1.509.

Other analyses. Pairwise differences of single-nucleotide variants (SNVs) were calculated in a similar approach as in a previous study⁸, between all individuals in a pairwise fashion across all significant windows, and for the same number of randomly sampled regions. The analysis of SNV differences, phylogenetic trees⁶⁹, principle component analyses⁷⁰ and significance tests were performed in the *R* environment⁵⁸ (Supplementary Information). Haplotype networks from all SNPs in the archaic fragments were built using the package *pegas*⁷¹. Results from the program *ARGweaver*³⁶ as applied and described previously²¹ were re-analyzed, and allele age estimated with “arg-summarize -A”. Information on functional changes was retrieved from previous studies on public data^{43,44} (Supplementary Information), and an enrichment test for GWAS traits was performed as described elsewhere^{43,72} (Supplementary Information). We mapped and quantified chimpanzee and bonobo transcriptome data⁴⁷ using the reference genome *hg19*^{73,74}, and tested for differential gene expression between the two species using *DESeq2*⁷⁵ (Table S20). We calculated the genome-wide distribution of F_{ST} between bonobos and chimpanzees in windows of 40 kbp, with 10 kbp steps, using *PopGenome*⁷⁶. Phylogenetic trees were drawn using *phangorn*⁶⁹, with Kimura’s distance⁷⁷. More details and additional analyses are described in the Supplementary Information.

Figure legends

Figure 1. Trees of putatively introgressed fragments. Neighbor 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).

Figure 2. Analysis of putatively introgressed windows. a) Number of basepair differences (Δbp) between all pairs of individuals used in this study, for putative archaic introgressed windows in bonobos using S^* (x-axis), and in the same number of random windows (y-axis). Violet: Chimpanzee vs. chimpanzee Δbp . Blue: Chimpanzee vs. bonobo Δbp . Turquoise: Bonobo vs. bonobo Δbp . Green: Δbp between the individual for which the S^* test was performed, and all other individuals. Δbp between bonobos and chimpanzees is larger in these windows than in random windows (topright, green for test individual, blue for other individuals in the same regions), suggesting elevated genetic distance. b) Principal component analysis (PCA) of SNPs⁷⁰. 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 b); note that the y-axis is flipped since this is calculated from different SNPs. e) Haplotype network⁷¹ of one archaic fragment in bonobos (chr8:30,599,999-30,670,000), 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 Fig. S16. f) Inferred age distribution³⁶ of SNVs falling in putative archaic windows in bonobos, and bonobo-like windows in central chimpanzees, compared to random windows. Archaic windows carry an excess of SNVs older than 2 Mya in archaic windows.

Figure 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 (*Pan paniscus*), and an unknown “ghost” population. Grey arrows: Previously described gene flow events between chimpanzees and bonobos²¹. Violet arrow: Archaic gene flow into bonobos. 95% CrI for introgression and archaic divergence times as well as introgression amounts are shown as inferred using S^* with ABC modeling (Methods), the divergence times of extant *Pan* populations were inferred using SFS-based modeling (Methods).

Figure 4. Posterior values of the models used. a) Posterior probabilities of 100 replicate tests for the ABC model selection test^{62,66} for the simplified SFS-based demographic model, the ABC-based model without archaic gene flow into bonobos, the ABC-based model with archaic gene flow into bonobos, the adjusted ABC-based model with archaic gene flow into bonobos. b) Δ Likelihoods (\log_{10}) of 100 replicates for the SFS-based model^{21,33} with and without archaic gene flow, and with ancient substructure in bonobos (Methods). c) Median log likelihoods for the Skov HMM²⁶ for 10 bonobo individuals, assuming no gene flow, one and two gene flow events (0, 1 and 2), using either all chimpanzees, or only central (CC) or western chimpanzees (WC) as reference panels.

Figure 5. Distribution of introgression across the genome. Karyogram of human chromosomes with 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 humans⁷ in orange. Inset table: Longest cumulative introgressed regions larger than 100 kbp, and overlapping genes. Inset figure: Chromosomes ordered by proportion covered by depleted regions (> 5 Mbp).

472 References

- 473 1. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* (80-.). **328**, 710–
474 722 (2010).
- 475 2. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia.
476 *Nature* **468**, 1053–1060 (2010).
- 477 3. Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C. & Wall, J. D. Genetic evidence
478 for archaic admixture in Africa. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 15123–8 (2011).
- 479 4. Meyer, M. *et al.* a High Coverage Genome Sequence From an Archaic Denisovan Individual.
480 *Science* **338**, 222–226 (2012).
- 481 5. Vernot, B. & Akey, J. M. Resurrecting Surviving Neandertal Lineages from Modern Human
482 Genomes. *Science* (80-.). **343**, 1017–1021 (2014).
- 483 6. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor.
484 *Nature* **524**, 216–219 (2015).
- 485 7. Vernot, B. *et al.* Excavating Neandertal and Denisovan DNA from the genomes of
486 Melanesian individuals. *Science* (80-.). **352**, 235–239 (2016).
- 487 8. Xu, D. *et al.* Archaic Hominin Introgression in Africa Contributes to Functional Salivary
488 MUC7 Genetic Variation. *Mol. Biol. Evol.* **34**, 2704–2715 (2017).
- 489 9. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains.
490 *Nature* **505**, 43–9 (2014).
- 491 10. Kuhlwilm, M. *et al.* Ancient gene flow from early modern humans into Eastern Neanderthals.
492 *Nature* **530**, 429–433 (2016).
- 493 11. Posth, C. *et al.* Deeply divergent archaic mitochondrial genome provides lower time
494 boundary for African gene flow into Neanderthals. **8**, 16046 (2017).
- 495 12. Juric, I., Aeschbacher, S. & Coop, G. The Strength of Selection against Neanderthal
496 Introgression. *PLOS Genet.* **12**, e1006340 (2016).
- 497 13. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The Combined Landscape of
498 Denisovan and Neanderthal Ancestry in Present-Day Humans. *Curr. Biol.* **26**, 1241–1247
499 (2016).
- 500 14. Huerta-Sanchez, E. *et al.* altitude adaptation in Tibetans caused by introgression of
501 Denisovan-like DNA. *Nature* **512**, 194–7 (2014).
- 502 15. Racimo, F., Marnetto, D. & Huerta-Sánchez, E. Signatures of Archaic Adaptive Introgression
503 in Present-Day Human Populations. *Mol. Biol. Evol.* **34**, 296–317 (2017).
- 504 16. Brunet, M. *et al.* A new hominid from the Upper Miocene of Chad, Central Africa. *Nature*
505 **418**, 145 (2002).

- 506 17. McBrearty, S. & Jablonski, N. G. First fossil chimpanzee. *Nature* **437**, 105–108 (2005).
- 507 18. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, 471–
508 5 (2013).
- 509 19. Hey, J. The divergence of chimpanzee species and subspecies as revealed in multipopulation
510 isolation-with-migration analyses. *Mol. Biol. Evol.* **27**, 921–33 (2010).
- 511 20. Tung, J. & Barreiro, L. B. The contribution of admixture to primate evolution. *Curr. Opin.*
512 *Genet. Dev.* **47**, 61–68 (2017).
- 513 21. De Manuel, M. *et al.* Chimpanzee genomic diversity reveals ancient admixture with bonobos.
514 *Science* (80-.). **354**, (2016).
- 515 22. Kawamoto, Y. *et al.* Genetic structure of wild bonobo populations: diversity of mitochondrial
516 DNA and geographical distribution. *PLoS One* **8**, e59660 (2013).
- 517 23. Takemoto, H., Kawamoto, Y. & Furuichi, T. How did bonobos come to range south of the
518 congo river? Reconsideration of the divergence of *Pan paniscus* from other *Pan* populations.
519 *Evol. Anthropol. Issues, News, Rev.* **24**, 170–184 (2015).
- 520 24. Takemoto, H. *et al.* The mitochondrial ancestor of bonobos and the origin of their major
521 haplogroups. *PLoS One* **12**, e0174851 (2017).
- 522 25. Myers Thompson, J. A. A model of the biogeographical journey from Proto-pan to Pan
523 paniscus. *Primates* **44**, 191–197 (2003).
- 524 26. Skov, L. *et al.* Detecting archaic introgression using an unadmixed outgroup. *PLOS Genet.*
525 **14**, e1007641 (2018).
- 526 27. Won, Y.-J. J. & Hey, J. Divergence population genetics of chimpanzees. *Mol. Biol. Evol.* **22**,
527 297–307 (2005).
- 528 28. Plagnol, V. & Wall, J. D. Possible Ancestral Structure in Human Populations. *PLoS Genet.* **2**,
529 e105 (2006).
- 530 29. Wall, J. D., Lohmueller, K. E. & Plagnol, V. Detecting ancient admixture and estimating
531 demographic parameters in multiple human populations. *Mol. Biol. Evol.* **26**, 1823–7 (2009).
- 532 30. Hsieh, P. *et al.* Model-based analyses of whole-genome data reveal a complex evolutionary
533 history involving archaic introgression in Central African Pygmies. *Genome Res.* (2016).
534 doi:10.1101/gr.196634.115
- 535 31. Lachance, J. *et al.* Evolutionary History and Adaptation from High-Coverage Whole-
536 Genome Sequences of Diverse African Hunter-Gatherers. *Cell* **150**, 457–469 (2012).
- 537 32. Nye, J. *et al.* Selection in the Introgressed Regions of the Chimpanzee Genome. *Genome*
538 *Biol. Evol.* evy077-evy077 (2018).

- 539 33. Excoffier, L., Dupanloup, I., Huerta-S?nchez, E., Sousa, V. C. & Foll, M. Robust
540 Demographic Inference from Genomic and SNP Data. *PLoS Genet.* **9**, (2013).
- 541 34. Hey, J. *et al.* Phylogeny Estimation by Integration over Isolation with Migration Models.
542 *Mol. Biol. Evol.* msy162-msy162 (2018).
- 543 35. Prüfer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*
544 (80-.). (2017).
- 545 36. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-wide inference of
546 ancestral recombination graphs. *PLoS Genet.* **10**, e1004342 (2014).
- 547 37. Prüfer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes.
548 *Nature* **486**, 527 (2012).
- 549 38. Kuhlwilm, M. *et al.* Evolution and demography of the great apes. *Curr. Opin. Genet. Dev.*
550 (2016). doi:10.1016/j.gde.2016.09.005
- 551 39. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day
552 humans. *Nature* **507**, 354–357 (2014).
- 553 40. Nam, K. *et al.* Extreme selective sweeps independently targeted the X chromosomes of the
554 great apes. *Proc. Natl. Acad. Sci.* **112**, 6413–6418 (2015).
- 555 41. Steinrücken, M., Spence, J. P., Kamm, J. A., Wiecek, E. & Song, Y. S. Model-based
556 detection and analysis of introgressed Neanderthal ancestry in modern humans. *Mol. Ecol.* **0**,
557 (2018).
- 558 42. Piccinni, M.-P. T cells in normal pregnancy and recurrent pregnancy loss. *Reprod. Biomed.*
559 *Online* **13**, 840–844 (2006).
- 560 43. Han, S., Andres, A. M., Marques-Bonet, T. & Kuhlwilm, M. Genetic variation in Pan species
561 is shaped by demographic history and underlies lineage-specific functions. *bioRxiv* (2018).
- 562 44. Kuhlwilm, M. & Boeckx, C. Genetic differences between humans and other hominins
563 contribute to the “human condition”; *bioRxiv* (2018).
- 564 45. Furuichi, T. Social interactions and the life history of femalePan paniscus in Wamba, Zaire.
565 *Int. J. Primatol.* **10**, 173–197 (1989).
- 566 46. Cagan, A. *et al.* Natural Selection in the Great Apes. *Mol. Biol. Evol.* **33**, 3268–3283 (2016).
- 567 47. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature*
568 **478**, 343–8 (2011).
- 569 48. Bouhassira, E. E. *et al.* An alanine-to-threonine substitution in protein 4.2 cDNA is
570 associated with a Japanese form of hereditary hemolytic anemia (protein 4.2NIPPON). *Blood*
571 **79**, 1846–1854 (1992).

- 572 49. Loomis, M. R. Great Apes. in *Zoo and Wild Animal Medicine 5th Edition* (eds. Fowler, M. E.
573 & Miller, R. E.) 381–397 (Saunders (Elsevier Science), 2003).
- 574 50. Cyrklaff, M. *et al.* Hemoglobins S and C interfere with actin remodeling in *Plasmodium*
575 *falciparum*-infected erythrocytes. *Science* **334**, 1283–6 (2011).
- 576 51. Dannemann, M., Andrés, A. M. & Kelso, J. Introgression of Neandertal- and Denisovan-like
577 haplotypes contributes to adaptive variation in human toll-like receptors. *Am J Hum Genet*
578 **98**, 22–33 (2016).
- 579 52. Frazer, J. K. *et al.* Identification of centerin: a novel human germinal center B cell-restricted
580 serpin. *Eur. J. Immunol.* **30**, 3039–3048 (2000).
- 581 53. Suzuki, K. *et al.* A novel glycosylphosphatidyl inositol-anchored protein on human
582 leukocytes: a possible role for regulation of neutrophil adherence and migration. *J. Immunol.*
583 **162**, 4277–84 (1999).
- 584 54. Hernandez-Rodriguez, J. *et al.* The impact of endogenous content, replicates and pooling on
585 genome capture from faecal samples. *Mol. Ecol. Resour.* (2017). doi:10.1111/1755-
586 0998.12728

587

588 **Methods references**

- 589 55. Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome*
590 *Res.* **18**, 1829–1843 (2008).
- 591 56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
592 features. *Bioinformatics* **26**, 841–842 (2010).
- 593 57. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with
594 genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
- 595 58. R Core Team. R: A Language and Environment for Statistical Computing. (2015).
- 596 59. Vernot, B. freezing-archer. (2016).
- 597 60. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS*
598 *Comput. Biol.* **9**, e1003118 (2013).
- 599 61. Haider, S. *et al.* A bedr way of genomic interval processing. *Source Code Biol. Med.* **11**, 14
600 (2016).
- 601 62. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of
602 semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **73**, 3–
603 36 (2011).
- 604 63. Hudson, R. R., Slatkin, M. & Maddison, W. Estimation of levels of gene flow from DNA
605 sequence data. *Genetics* **132**, 583–589 (1992).

64. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471 (2012).
65. Langergraber, K. E. *et al.* Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. ...* **109**, 15716–15721 (2012).
66. Csilléry, K., François, O. & Blum, M. G. B. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3**, 475–479 (2012).
67. Stevison, L. S. *et al.* The Time-Scale of Recombination Rate Evolution in Great Apes. *bioRxiv* 013755 (2015). doi:10.1101/013755
68. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Comput. Biol.* **12**, e1004842 (2016).
69. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–3 (2011).
70. Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).
71. Paradis, E. pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* **26**, 419–420 (2010).
72. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
73. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
74. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166 (2015).
75. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–34 (2014).
76. 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).
77. 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).

Supplementary Information is linked to the online version of the paper at xxx.

Acknowledgements

We thank A. M. Andres and B. Vernot 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. by Fundação para a Ciência e a Tecnologia (project UID/BIA/00329/2013), and by EU H2020 programme (Marie Skłodowska-Curie grant 799729), L.E. by Swiss NSF No. 310030B-166605, and T.M.-B. by MINECO BFU2014-55090-P (FEDER), U01 MH106874 grant, Howard Hughes International Early Career, 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. analyzed data. M.K. and T.M.-B. wrote the manuscript. The authors declare no competing interests. Correspondence and requests for materials should be addressed to: martin.kuhlwilm@upf.edu; tomas.marques@upf.edu

Data and materials availability: All data is publicly available or in the Supplementary Information.