RESEARCH ARTICLE SUMMARY

PRIMATE GENOMES

Pervasive incomplete lineage sorting illuminates speciation and selection in primates

Iker Rivas-González†, Marjolaine Rousselle†, Fang Li†, Long Zhou, Julien Y. Dutheil, Kasper Munch, Yong Shao, Dongdong Wu, Mikkel H. Schierup*, Guojie Zhang*

INTRODUCTION: Incomplete lineage sorting generates gene trees that are incongruent with the species tree. Incomplete lineage sorting has been described in many phylogenetic clades, including birds, marsupials, and primates. For example, the level of incomplete lineage sorting in the human-chimp-gorilla branch adds up to ~30%, which means that, even though our closest primate relatives are chimps, 15% of our genome resembles more the gorilla than the chimp genome, and another 15% groups the chimp with the gorilla first.

RATIONALE: Although incomplete lineage sorting is usually regarded as an obstacle for phylogenetic reconstruction, it holds valuable information about the evolutionary history of the species because its extent depends on the ancestral effective population sizes and the time between speciation events. Additionally, recurrent ancestral selective processes are expected to influence how the proportion of incongruent trees varies along the genome, which makes incomplete lineage sorting a useful tool to study ancient evolutionary events. In this study, we estimate the incomplete lineage sorting landscape by running a coalescent hidden Markov model in species trios along a 50-way primate genome alignment. We then leverage the signal of incomplete lineage sorting to reconstruct ancestral effective population parameters and to analyze the genomic determinants that influence the sorting of lineages.

RESULTS: We find widespread incomplete lineage sorting across the primate tree in 29 nodes, some reaching as much as 64% of the genome. Combining CoalHMM with a machine learning pipeline, we reconstruct the speciation times of the primate phylogeny without the need for fossil calibrations. Our speciation time estimates are more recent than divergence times, and they are in agreement with previous estimates based on fossil evidence. Our reconstructed ancestral effective population sizes show that they increase toward the past.

We additionally detect regions that have low or high incomplete lineage sorting levels consistently across several nodes. We show that incomplete lineage sorting proportions increase with the recombination rate in the genomic region—a difference that translates into an up to fourfold variation in the inferred local effective population size. Moreover, we report low levels of incomplete lineage sorting on the X chromosome. This reduction is more pronounced than expected under neutral evolution, which suggests that selective forces affect the X chromosome more strongly than the autosomes, reducing the effective population size of the X chromosome and, subsequently, the levels of incomplete lineage sorting.

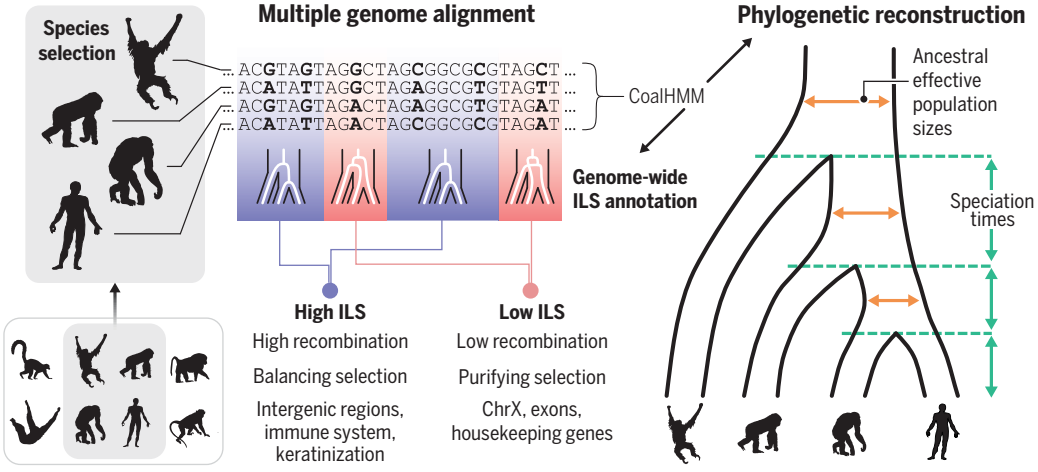
We further assess how selection affects the distribution of incomplete lineage sorting patterns by comparing the incomplete lineage sorting proportions of exons with those in intergenic regions. We find that there is an overall decrease in the levels of incomplete lineage sorting in exons that amounts to a reduction of 31% in the local effective population size as compared with intergenic regions.

Finally, we perform a gene ontology enrichment analysis on low– and high–incomplete lineage sorting genes. We find that immune system genes show large proportions of incomplete lineage sorting for many of the nodes, whereas housekeeping genes with basic cell functions show a lack of incomplete lineage sorting.

CONCLUSION: Most molecular-based methods that aim at timing a species tree provide estimates of divergence times, which are confounded by ancestral population sizes compared with the actual speciation times. We showed that using the coalescent theory and the signal of incomplete lineage sorting allows us to accurately estimate speciation times and ancestral population sizes in the primate tree, gaining key insights regarding some aspects of primate biology. Our study also emphasizes the prevalence of natural selection at linked sites that shapes the landscape of both genetic diversity and incomplete lineage sorting along the primate genome. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: mheide@birc.au.dk (M.H.S.); guojiezhang@zju.edu.cn (G.Z.)
†These authors contributed equally to this work.
Cite this article as I. Rivas-González *et al.*, *Science* **380**, eabn4409 (2023). DOI: 10.1126/science.abn4409

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn4409



Inference of the speciation history and the genomic landscape of natural selection in primates from patterns of incomplete lineage sorting. CoalHMM was used to capture the signal of incomplete lineage sorting (ILS) segments along the genomes of 50 primate species and to estimate coalescent parameters—i.e., the ancestral effective population sizes and speciation times. Moreover, the genome-wide variation in the levels of incomplete lineage sorting allowed for the inference of selective processes in primates. ChrX, X chromosome.

RESEARCH ARTICLE

PRIMATE GENOMES

Pervasive incomplete lineage sorting illuminates speciation and selection in primates

Iker Rivas-González[1]†, Marjolaine Rousselle[1]†, Fang Li[2,3,4]†, Long Zhou[5,6], Julien Y. Dutheil[7,8], Kasper Munch[1], Yong Shao[9], Dongdong Wu[9,10,11,12], Mikkel H. Schierup[1]*, Guojie Zhang[5,6,9,13,14]*

Incomplete lineage sorting (ILS) causes the phylogeny of some parts of the genome to differ from the species tree. In this work, we investigate the frequencies and determinants of ILS in 29 major ancestral nodes across the entire primate phylogeny. We find up to 64% of the genome affected by ILS at individual nodes. We exploit ILS to reconstruct speciation times and ancestral population sizes. Estimated speciation times are much more recent than genomic divergence times and are in good agreement with the fossil record. We show extensive variation of ILS along the genome, mainly driven by recombination but also by the distance to genes, highlighting a major impact of selection on variation along the genome. In many nodes, ILS is reduced more on the X chromosome compared with autosomes than expected under neutrality, which suggests higher impacts of natural selection on the X chromosome. Finally, we show an excess of ILS in genes with immune functions and a deficit of ILS in housekeeping genes. The extensive ILS in primates discovered in this study provides insights into the speciation times, ancestral population sizes, and patterns of natural selection that shape primate evolution.

Comparative genomics can offer insights into population processes deep in phylogenetic history. As a result of recombination, different parts of our genomes have different genealogical histories (*1*, *2*). Therefore, when speciation occurs, the genes of the resulting descendants can be traced back to different ancestors, each coalescing at different times that stochastically depend on both the species population size and natural selection acting on each gene. If the time between two consecutive speciation events is short and/or the effective population size (*Ne*) is large, then genes from the two most closely related species may coalesce deeper in the past than the time of the oldest speciation event. This can result in genealogical histories that are different from the species tree—a phenomenon called incomplete lineage sorting (ILS). ILS has affected the evolutionary history of the human genome as well as many other groups (*3*–*5*). Around 30% of the human genome does not follow the ((human, chimpanzee), gorilla) speciation tree (*2*, *6*–*8*), with 15% of nucleotide positions grouping human and gorilla, and 15% grouping gorilla and chimpanzee.

Although the phylogenetic incongruences produced by ILS can hamper gene tree reconstruction from single loci, they offer an opportunity to learn about the population history of species sitting in deep ancestral branches of the phylogeny (*6*, *9*–*11*). We can, for example, estimate the actual times when species split as opposed to the more ancient average time to the most recent common ancestor, and we can measure how natural selection, directly or indirectly, affected the genomic diversity of the ancestral species. For example, Dutheil *et al.* (*12*) have concluded that the lack of ILS on the X chromosome in the human-chimp ancestor first reported by Patterson *et al.* (*13*) was likely a result of several episodes of very strong positive selection.

The recent effort to de novo assemble a large number of primate genomes makes it possible to extend the study of ILS to many more nodes across the primate phylogeny, allowing estimation of the speciation times and the forces that shaped genetic diversity in the ancestral species. With many independent replicates of the ILS process, we can learn about common targets of natural selection during primate diversification. In this work, we apply an extended version of the CoalHMM model (*14*) to a whole-genome alignment of 50 primate species (10 prosimians, 7 New World monkeys, 23 Old World monkeys, and 10 great and lesser apes). We report high levels of ILS on 29 of the total number of internal branches, and we estimate dates of the speciation times independently of fossil calibration that are in concordance with available fossil evidence. Additionally, we report recombination rate, ancestral effective population sizes, and selection as major genomic and functional determinants that have shaped the patterns of ancestral primate diversity.

Results

ILS is pervasive on most branches of the primate tree

We applied CoalHMM to the internal branches of the primate tree for 50 species used in Shao *et al.* (*15*) and shown in Fig. 1A, using combinations of quartets of species from the genome-wide alignment (see the supplementary materials, section 4). After filtering out ambiguously aligned regions, we used posterior decoding to infer segments of the alignment best supported by either the species topology or any of the two possible discordant topologies. Figure 1A shows the level of autosomal ILS detected on individual branches of the phylogeny. Branch lengths represent estimated genomic divergence times obtained by dividing substitution rates of the ExaML Gamma model by an estimate of the yearly mutation rate of each branch (supplementary materials, sections 3 and 7). We found appreciable genome-wide ILS proportions between 5 and 64% on 29 of the 49 branches, which implies that, on these branches, a large proportion of the genome follows a different gene genealogy from that of the species tree (Fig. 1A). The length distribution of the genome segments supporting the discordant topologies (i.e., topologies V2 and V3 in Fig. 1A, inset) depends mainly on the effective population size of the examined branch and is expected to follow a geometric distribution. Except for a deficiency of very short segments, this assumption is generally met in our analysis (fig. S7). We also show that the mean length of segments supporting both the species topology and the discordant topologies varies substantially among nodes, with mean lengths for discordant segments between 100 and 1000 base pairs for individual branches (fig. S7). This shows that single genes, which typically cover >20 kb in the genome, rarely have just one phylogenetic history when ILS is prominent.

A previous study based on the phylogenies of 1700 genes concluded that hybridization

[1]Bioinformatics Research Centre, Aarhus University, DK-8000 Aarhus C, Denmark. [2]BGI-Research, BGI-Wuhan, Wuhan 430074, China. [3]Institute of Animal Sex and Development, ZhejiangWanli University, Ningbo 315104, China. [4]BGI-Research, BGI-Shenzhen, Shenzhen 518083, China. [5]Evolutionary & Organismal Biology Research Center, Zhejiang University School of Medicine, Hangzhou 310058, China. [6]Women's Hospital, School of Medicine, Zhejiang University, Shangcheng District, Hangzhou 310006, China. [7]Max Planck Institute for Evolutionary Biology, Plön, Germany. [8]Institute of Evolution Sciences of Montpellier (ISEM), CNRS, University of Montpellier, IRD, EPHE, 34095 Montpellier, France. [9]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China. [10]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, Yunnan 650223, China. [11]National Resource Center for Non-Human Primates, Kunming Primate Research Center, and National Research Facility for Phenotypic and Genetic Analysis of Model Animals (Primate Facility), Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650107, China. [12]Kunming Natural History Museum of Zoology, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China. [13]Liangzhu Laboratory, Zhejiang University Medical Center, Hangzhou 311121, China. [14]Villum Centre for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark.
***Corresponding author. Email: mheide@birc.au.dk (M.H.S.); guojiezhang@zju.edu.cn (G.Z.)**
†These authors contributed equally to this work.

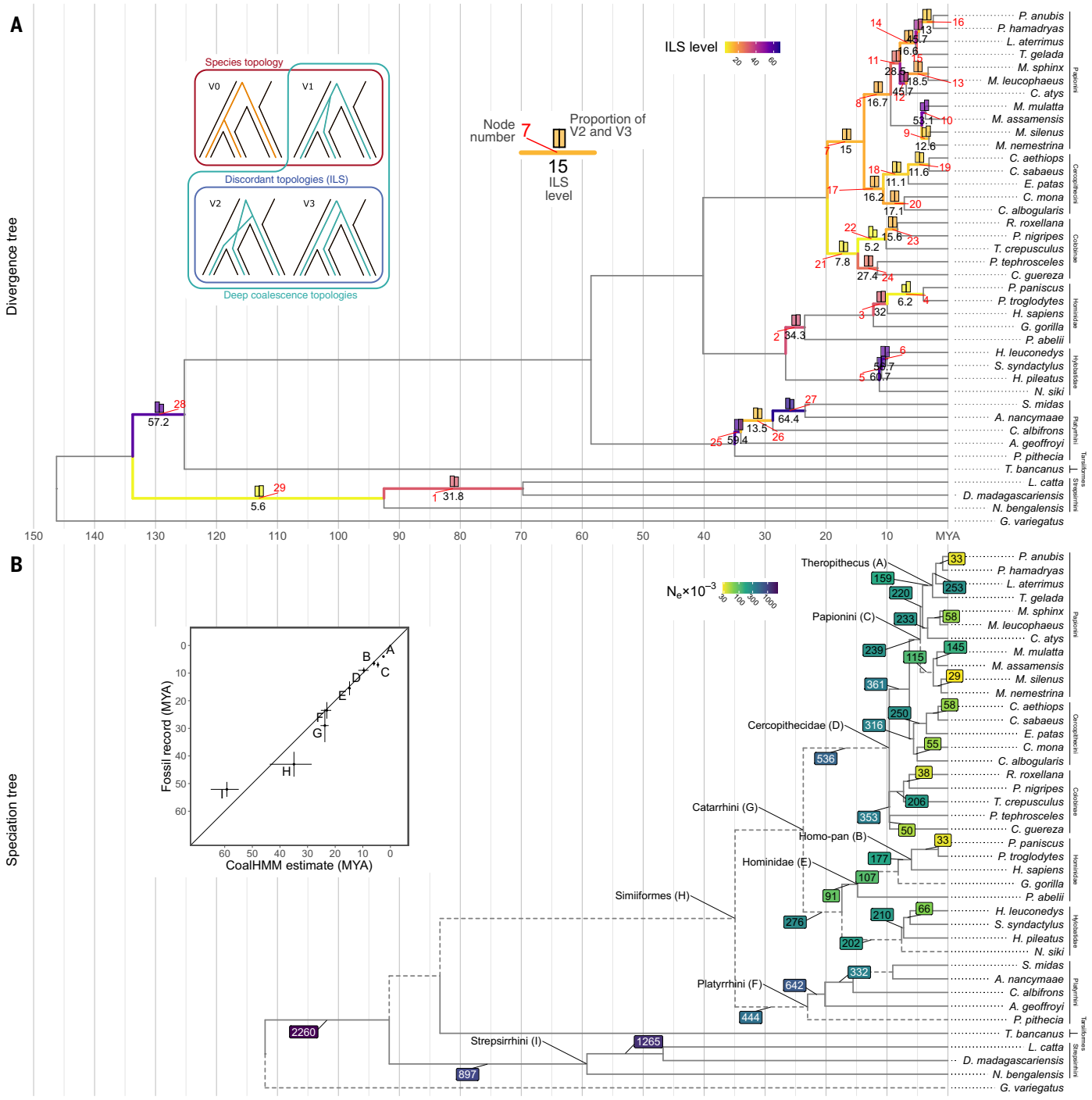


Fig. 1. Phylogenetic tree of primates, with scaled divergence times or speciation times as branch lengths. (**A**) Divergence time tree scaled with estimated mutation rate for individual branches (supplementary materials, section 7). Percentage ILS (the sum of V2 and V3 topologies; see inset) is plotted as branch color and marked with numbers for those branches with >5% of ILS. Only the subset of 38 species that were used to infer the ILS of the colored branches are plotted for clarity. The two columns on top of each branch show the relative frequency of bases attributed to V2 and V3, respectively. The numbers in red denote individual branches referenced in subsequent figures. The taxonomic classification is shown to the right of the phylogeny. (**B**) Speciation time tree with branch lengths in units of million years (MYA) as estimated from CoalHMM and scaled with estimated ancestral mutation rates for individual branches (supplementary materials, section 7). The annotations in colored rectangles refer to the inferred ancestral effective population sizes. Branches without enough information to infer speciation times using CoalHMM (i.e., branches with <5% ILS) are shown as dashed lines. Here, speciation times are instead estimated by subtracting an assumed population size (the CoalHMM estimate of the ancestral population size of the closest branch) from the divergence time rescaled by mutation rate per generation (supplementary materials, section 10). The inset panel shows the correlation between the split times estimated by CoalHMM and the dated fossil record. Each point corresponds to an evolutionary node in the right panel. Horizontal lines correspond to the bootstrapped standard deviation of the estimated branch length, and vertical lines represent the standard deviation of the fossil date estimates (data are shown in table S4).

events are as common in the deeper branches of the primate tree as they are today between related extant species of many primate groups (*16*, *17*). To estimate to what extent the phylogenetic incongruence that the model attributes to ILS is affected by widespread hybridization on the deeper branches, we investigated the relative frequency, nucleotide divergence, and length of the genomic fragments assigned to the two discordant topologies on each internal branch. If explained by ILS, the three measures should all be equal for the two discordant topologies, whereas hybridization is expected to cause one of the discordant topologies to be more frequent, and the genomic segments supporting the predominant topology should be, on average, longer and less divergent than those supporting the other discordant topology. On most of the 29 internal branches, we observe near-equal proportions of genomic positions assigned to the two discordant topologies (see the proportions of V2 versus V3 in Fig. 1A), and we find that the fragments have very similar size distributions (fig. S7). After correcting for different substitution rates (supplementary materials, section 9), we also find that segments with the two discordant topologies are close to equally divergent (figs. S17 and S19). Exceptions to these general patterns are found within the recent macaque, gibbon, and lesser apes divergences. In these cases, evidence of introgression has also been reported previously (*16*, *18*, *19*). However, even in those cases, ILS is the predominant cause of incongruent genealogies in the primate tree (*20*) (supplementary materials, section 9). It is possible that hybridizations occurred between related species in deeper branches, as is observed in several extant genera. However, if a pair of hybridizing species did not both leave extant descendant species (as is likely because most species die out), this would not have been distinguishable from deep coalescences in causing ILS. Thus, we cannot completely exclude that gene flow occurred at ancestral branches—only that it did not leave detectable evidence of ancient hybridization.

The level of ILS generally increases with shorter internal branch lengths (Fig. 1A). In the taxon sampling of our present dataset, we find that ILS is particularly ubiquitous in Old World monkeys, which have undergone rapid speciation events. Notably, however, 32% ILS is estimated even on a very long and deep branch within Strepsirrhini and 57% on the branch separating tarsiers from Strepsirrhini (branch 1 and branch 28, respectively; Fig. 1A), which suggests very large ancestral population sizes in these nodes that can also be predicted from the short size of the ILS fragments (fig. S7). Furthermore, the very high levels of ILS in gibbons and Old World monkeys, particularly macaques and baboons, explain the long-standing difficulty to resolve their phylogenetic relationships (*16*, *19*, *21*, *22*).

Speciation times and ancestral effective population sizes in the primate tree

The reconstruction of the dated history of a group of species is typically based on genomic divergence rates turned in divergence times through fossil calibrations (*23*–*25*). However, the genomic divergence times in species with large populations and long generation times can be much further back in time than the time when species actually split. The expected time for genomic coalescence on an ancestral branch is 2 × *Ne* generations older than the times of speciation. For an ancient population with an *Ne* of 200,000 and a generation time of 10 years, the average expected genomic divergence time would be 4 million years further back in time than the actual species split time. The analysis of incongruences produced by ILS via CoalHMM allows direct estimation of speciation times as opposed to divergence times as well as estimation of the ancestral effective population sizes.

We used the estimated parameters using CoalHMM together with simulations and a random forest model to derive ancestral effective population sizes and speciation times in all nodes with >5% of ILS (supplementary materials, section 10, and fig. S20). We then rescaled the parameters by estimated yearly mutation rates, which we derived from the relationship between pedigree-based yearly mutation rate and generation time, and the relationship between inferred body mass of extant and ancestral species and generation time (*26*, *27*) (supplementary materials, section 7). The resulting tree (fig. S21) was close to ultrametric and was linearized to make the speciation time tree shown in Fig. 1B (and that in fig. S22).

We infer ancestral effective population sizes that vary more than an order of magnitude within the primate phylogeny. In the few cases where ancestral effective population sizes of primate lineages have been estimated by other approaches, they are in good agreement with our estimates (*21*, *28*–*30*). For instance, Warren *et al.* (*29*) have estimated effective population sizes in the ancestors of the *Chlorocebus* lineage at around 40,000 using a multiple sequentially Markovian coalescent (MSMC) approach, when we infer an ancestral population size of 58,000, and Schrago and Seuánez (*21*) have estimated *Ne* in the ancestors of *Aotus* and Callitrichinae to >240,000 using a MSMC approach, when we infer an ancestral population size of 330,000. Most estimated ancestral *Ne* values are higher than effective population sizes estimated for primates today. This might reflect the fact that the ancestors of primates had smaller body sizes (*31*), which is known to be associated with larger population sizes, or that lineages with small population sizes are more likely to go extinct, leaving no descendants to sample from (*32*). As expected, population size estimates are negatively correlated with the median segment size of the discordant topologies (fig. S24). We also find an expected negative correlation between our estimate of ancestral *Ne* and the efficiency of purifying selection measured as dN/dS (the ratio of nonsynonymous to synonymous mutations) on the ancestral branches (fig. S25; $P = 0.0015$) and an expected negative correlation between average segment length and dN/dS (fig. S26).

Our inferred species split times are generally in good agreement with independent estimates from the fossil record when these exist (Fig. 1B, inset, and table S4), which supports that our approach can also infer speciation times on nodes that lack fossil evidence without the need for fossil calibration. Previous studies extrapolating the speciation time on the basis of pedigree-based mutation rates back in time have generally led to estimated times much further back in time than those suggested by the fossil record (*6*, *33*, *34*). We see two reasons for this. First, the large effective population sizes imply that divergence-based estimates of split times are several million years further back in time than the actual species split times. Second, our analysis rescales branch lengths by yearly mutation rates dependent on body size and generation time.

Highly variable frequency of ILS along the genome

Under selective neutrality, ILS is expected to occur at random along the genome. However, if natural selection, either directly or indirectly, affects the coalescent process of a genomic region, the sorting of lineages with deep coalescence will not be random (*12*, *35*, *36*). We painted all the genomes of the 29 ancestral branches by the level of ILS in 100-kb windows displayed as horizon plots (*37*, *38*) (fig. S8) and found many regions that experienced either high or low levels of ILS in the same genomic positions across several ancestral nodes in the primate phylogeny. We therefore integrated the ILS inference across the 29 branches using normalized ILS scores displayed in a single horizon plot showing the general pattern of ILS with the human genome coordinates as reference (Fig. 2A). This integrated signal of ILS shows that certain regions have consistently high or low levels of ILS. As an example, ILS is reduced in a large genomic region from 40 to 60 Mb on chromosome 3 (chr3) (Fig. 2B), which suggests either repeated selective sweeps or strong background selection (*11*, *36*). By contrast, the human lymphocyte antigen–major histocompatibility complex (HLA-MHC) cluster on position 27 to 33 Mb on chr6 has several regions showing

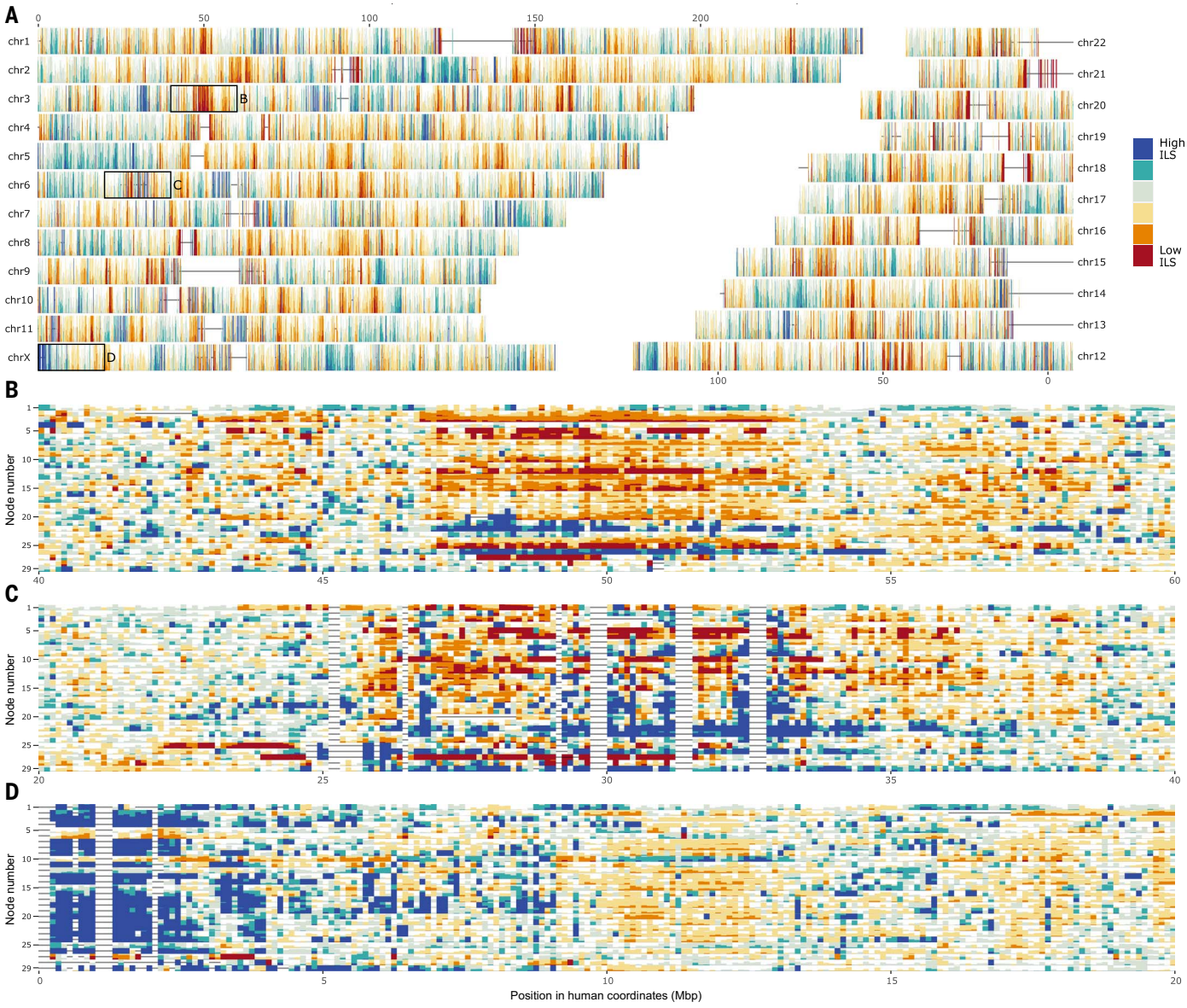


Fig. 2. Genome-wide distribution of ILS levels. (**A**) Horizon plot of the mean z-standardized ILS values in 100-kb windows (*x* coordinates in megabases). Red colors represent regions low in ILS, and blue colors represent high-ILS regions. Missing data are represented by a horizontal line. Regions marked with a rectangle in (A) are zoomed in. (**B** to **D**) A low-ILS region in chr3 (B), the MHC in chr6 (C), and the PAR region of the X chromosome (D). (B) to (D) are all horizon plots for all of the 29 individual nodes, where each node is mapped to Fig. 1A, inset. Mbp, mega–base pairs.

extremely high ILS, likely as a result of balancing selection (Fig. 2C). Additionally, the pseudoautosomal region (PAR) on position 0 to 2.7 Mb on the X chromosome also contains much higher ILS than the rest of the X chromosome (Fig. 2D) and, in many nodes, much higher ILS than the autosomal average. These and many other consistent patterns suggest that there are genomic and/or functional determinants of ILS that persist across the primate phylogeny.

Determinants of the variation in ILS along the genome

Recombination is not expected to directly affect the amount of ILS but can do so indirectly because the amount of recombination determines the efficacy of both positive and negative selection and, thus, the amount of diversity that is lost because of selection at linked genomic positions. A general observation of a positive correlation between nucleotide diversity and the recombination rate in extant species, including humans, has been interpreted as evidence for both the action of linked selection and as a mutagenic effect of recombination (*39*–*41*). ILS patterns will not be affected by the latter, so we investigated how ILS depends on recombination rate by extrapolating the human pedigree–based recombination map (*42*) at a 100-kb scale to the whole primate phylogeny. We inferred ILS levels and the corresponding relative local *Ne* as a function of recombination rate divided into ten bins (fig. S15 and supplementary materials, section 8). We find that the *Ne* of genomic regions with the highest recombination rate is typically 1.3-fold to fourfold larger than that in the lowest recombination bin (Fig. 3A), which implies that linked selection has removed a large proportion of the diversity in the ancestral species. Additionally, the extent of the effect of linked selection on genetic variation that we observe is likely underestimated because the present-day human recombination map is an imperfect proxy of the recombination landscape in ancestral species separated by tens of million years from humans.

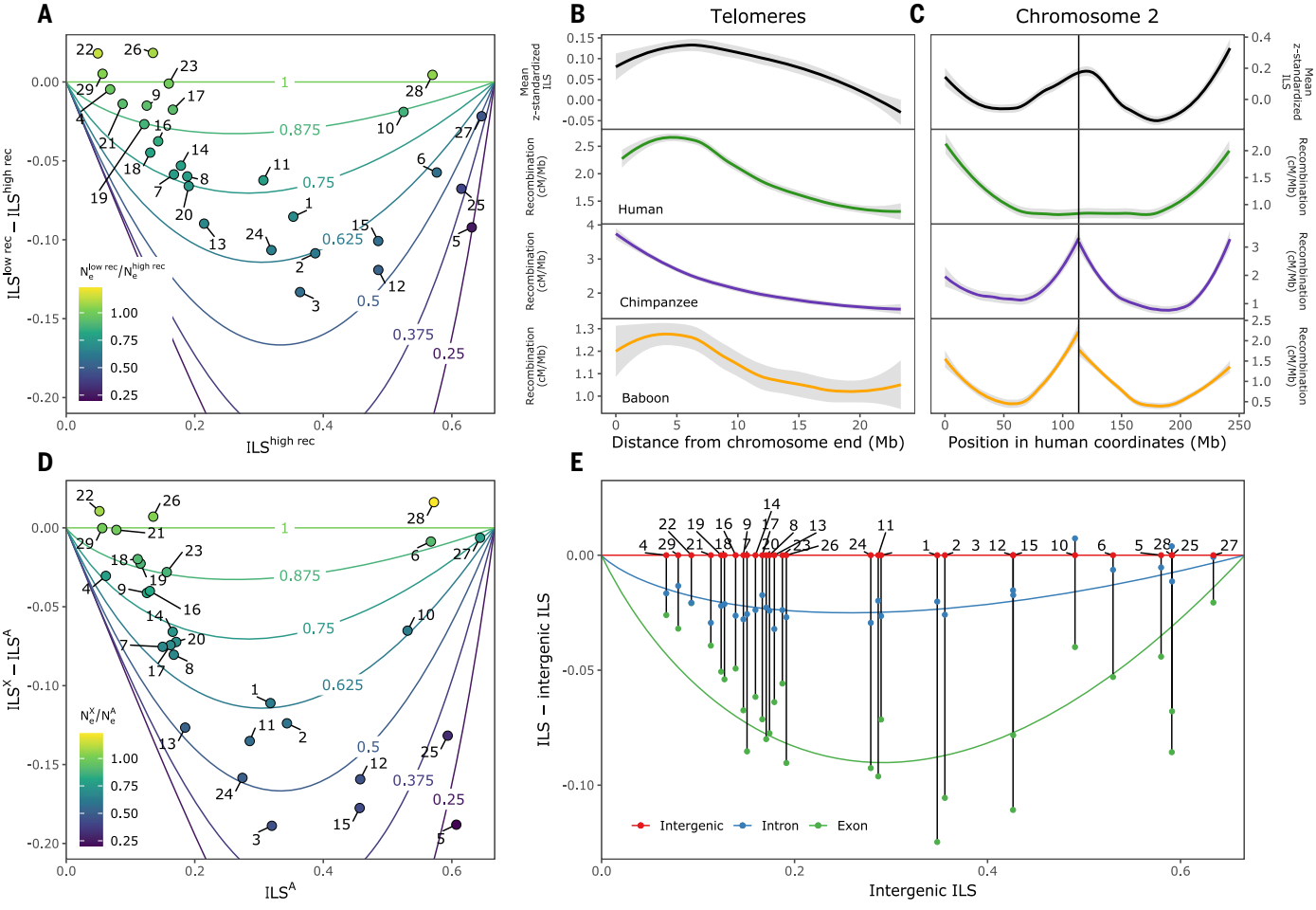


Fig. 3. Determinants of variation in ILS and corresponding *Ne*. (**A**) Difference in the proportion of ILS between the lowest recombination and the highest recombination deciles against the proportion of ILS in the highest recombination decile. Each numbered point represents a node in the phylogeny mapped to Fig. 1A, inset. The color and lines represent the relative change in *Ne* between the low and high recombination deciles, calculated using eq. S3 in the supplementary materials, section 8. (**B** and **C**) Comparison of the mean z-standardized proportion of ILS across 29 branches and the human (green), chimp (purple), and baboon (orange) recombination maps in the telomeres (B) and in chr2 (C). In (C), the fusion point is represented by a vertical line. (**D**) Difference between the ILS proportion of chromosome X and autosomes, where each numbered point is a node in the phylogeny mapped to Fig. 1A, inset. The color and lines represent the relative change in *Ne* between chromosome X and autosomes, calculated using eq. S3 in the supplementary materials, section 8. (**E**) Difference between the proportion of ILS in either exons (green) or introns (blue) and intergenic regions (red). Each numbered point and the corresponding vertical line represent one of 29 nodes in the phylogeny, mapped to Fig. 1A, inset. The colored lines represent fitted models that translate into a constant reduction in *Ne* across nodes.

Telomeres recombine more frequently than the rest of the genome (*42*–*45*). The integrated signal across all nodes and autosomal telomeres (Fig. 3B) shows a peak in the telomeric ILS that agrees with human (*42*), chimpanzee (*45*), and olive baboon (*46*) recombination maps at the tips of the chromosomes. Moreover, there is an increased signal of ILS at around position 114 Mb of chr2 (in human coordinates) (Fig. 3C), which corresponds to the remnants of an ancient telomere-telomere fusion affecting only the human lineage (*47*). Notably, we can only detect the corresponding peak in recombination in this region using the recombination map of nonhuman primate species, which suggests that, although big chromosomal rearrangements might markedly change the present-day recombination patterns, ILS can still be used to infer the ancestral recombination landscape in the primate phylogeny (*48*).

We next contrasted the ILS on the X chromosome with that on the autosomes. Because males only carry a single copy of the X chromosome in primates, and, consequently, it has a smaller effective population size, the X chromosome is expected to have lower ILS. We find that the X chromosome has an overall lower amount of ILS compared with the autosomal average (Fig. 3D and fig. S6), with the decrease corresponding to the *Ne* of chromosome X being between 50 and 75% of that of the autosomes. Under random mating and unbiased sex ratio, the Ne_X/Ne_A ratio is expected to equal 75% (*49*). However, in primates, males typically have the highest variance of reproductive success (*50*), which is at odds with our observed ratios smaller than 0.75 (Fig. 3D).

Previous surveys of chromosome X to autosome diversity have also often reported ratios below 0.75—e.g., 0.6 in non-African humans (*51*), 0.4 in gorillas, 0.5 in orangutans (*52*), and 0.3 in macaques (*53*). These observations have often been ascribed to differences in male and female mutation rates and recent bottleneck effects affecting the X chromosome diversity more than the autosomal diversity (*54*). However, sex differences in mutation rates should not affect ILS inference, and bottlenecks are unlikely as a general explanation throughout the primate phylogeny. We thus conclude that the large reduction in ILS on the X chromosome is likely a result of linked selection targeting the X chromosome to a

larger extent than the autosomes, as has been reported previously in the human-chimpanzee ancestral species (*12*). The 1.5- to 2.7-Mb PAR of the X chromosome is very high in ILS in most ancestral species (fig. S8). This is consistent with its very high recombination rate in males—~22 times the genome average rate, which minimizes the effect of linked selection—and its high polymorphism in great apes (*55*).

The strong positive correlation between ILS and recombination (fig. S15) suggests that positive and negative selection events had a strong impact on the removal of diversity in the ancestral species. These selective events are more likely enriched in genes, so we contrasted the amount of ILS in coding regions, introns, and intergenic regions (Fig. 3E). We find that, for all internal branches, $ILS_{exon} < ILS_{intron} < ILS_{intergenic}$. We estimate that a constant average reduction in the *Ne* of exons of 31% compared with the *Ne* in intergenic regions across the primate nodes would amount to the observed decrease in exonic ILS ($P < 2 \times 10^{16}$; SD = 1.4). Additionally, introns have an estimated average reduction in *Ne* of 10% compared with intergenic regions ($P < 2 \times 10^{16}$; SD = 0.5), which we interpret as a direct effect of their closer physical proximity to exons, leaving intronic ILS more strongly affected by linked selection than intergenic ILS.

ILS and gene function

Finally, we investigated whether certain gene categories are more likely to experience high levels of ILS than others—either because they experience less purifying selection and adaptive evolution or because they are more likely to be under balancing selection. We performed gene ontology enrichment tests with ILS as the response variable (supplementary materials, section 12).

We identify the most significant gene ontology terms enriched for either high or low ILS genes across the primate nodes and plot the gene ontology terms as a function of their average dN/dS ratio (Fig. 4A). As expected, more selectively constrained gene categories have significantly lower ILS than the genic average (correlation coefficient, $r = 0.35$; $P = 2.68 \times 10^{-10}$). These include many house-keeping gene categories and genes categories associated with chromosome organization and regulation. The *PIAS3* gene involved in transcriptional modulation is an example of consistently low ILS (Fig. 4B, left; other examples are in fig. S27).

Notably, the two gene ontologies with the highest ILS are "cornification" and/or "keratinization" and "immune response regulation." Corfinication (enriched for high ILS in 12 nodes) and keratinization (enriched for high ILS in 17 nodes) are tightly related gene ontology terms that include epidermal and keratinization genes. Primates exhibit an extraordinary degree of color variation across

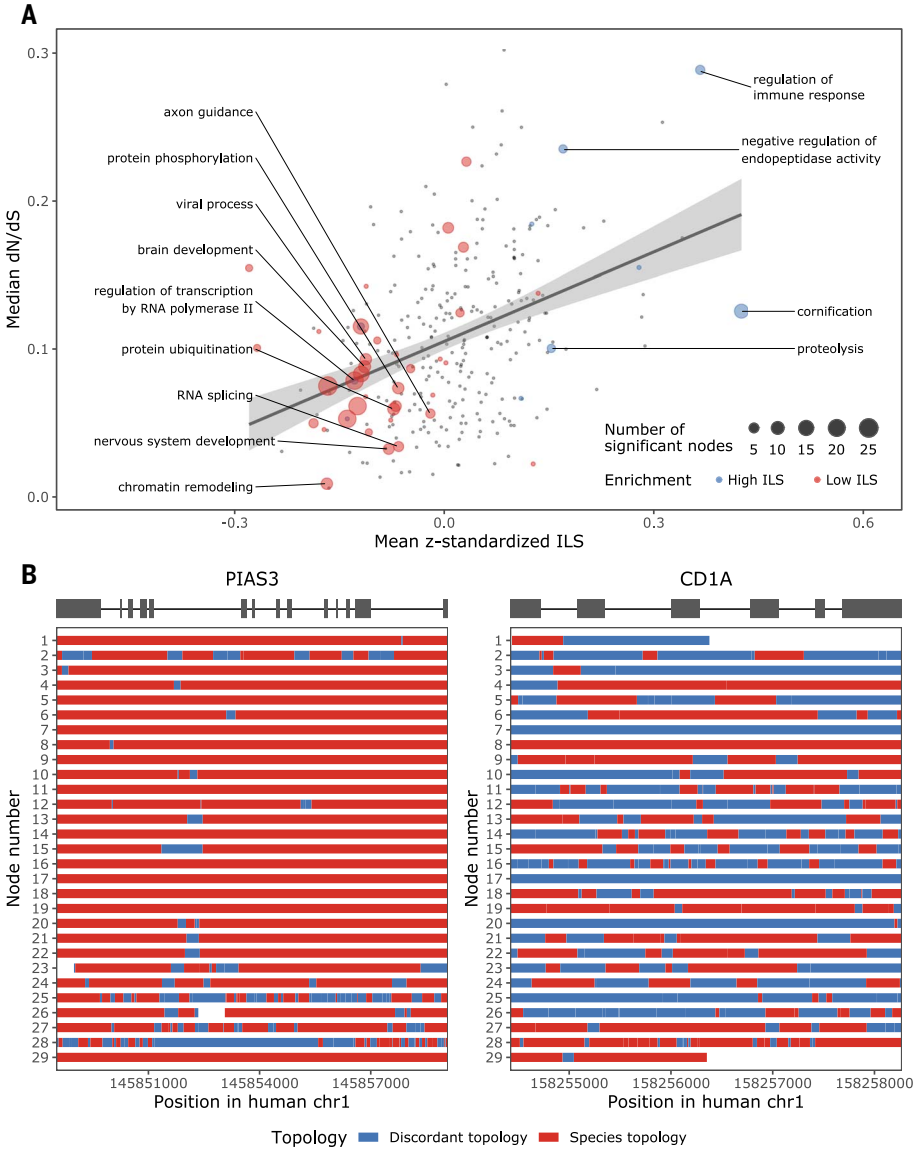


Fig. 4. ILS and gene function. (**A**) relationship between the ILS and the median dN/dS for each gene ontology term. Each data point corresponds to one gene ontology term, where the median dN/dS across all 29 nodes is plotted on the *y* axis, and the mean z-standardized ILS across all 29 nodes is represented on the *x* axis. Blue points are gene ontology terms that are significantly enriched for high ILS in at least one node, and red points correspond to gene ontology terms significantly enriched for low ILS. The size of the data points represents the number of nodes for which that gene ontology term has been significantly detected in the enrichment test. (**B**) Examples of genes with consistently low ILS (*PIAS3*, left) or consistently high ILS (*CD1A*, right). Each row corresponds to the inferred topologies (V0 or V1 in blue, and V2 or V3 in red) per genomic position for each node in the primate phylogeny. The top gray bar represents exons (in thick lines) and introns (in thin lines).

and within species and even in different parts of the body (*56*, *57*), which highlights the importance of the phenotypic evolution of skin in primates. This high diversity of coloration is crucial as social and sexual signaling and is often under stabilizing selection or positive selection that is closely linked with the high variations of primates in ecological niches, color vision, mating, and social systems (*58*). Additionally, some of the keratin gene families exhibit high levels of gene duplication and high functional diversification in primates (*59*–*61*).

Immune response regulation genes have been reported to evolve under balancing selection in primates (*62*), consistent with their enrichment in high-ILS genes. The MHC in chr6 is an outstanding region enriched for ILS, especially in Old World monkeys. Many other genes related to the immune response in genomic locations other than the HLA are also high in ILS. The detailed ILS pattern for the

CD1A gene (chr1) involved with innate immune response (Fig. 4B, right; other examples in fig. S28) reveals a higher ILS proportion in this gene above the average across the 29 nodes. Other examples are the ULBP family and killer cell immunoglobulin-like receptor (KIR) proteins. This last family is highly diverse, and it is consistent with patterns of balancing selection in several present-day human populations (*63*, *64*) and other primates (*65*).

Conclusion

The inference of ILS on many nodes in the primate phylogeny allows us to estimate speciation times and ancestral population sizes directly from genomic divergence data. We found that the effective population sizes have been very large in early primate evolution, at least in most lineages that have descendants today. This explains why the genomic divergence times estimates are much further back in time than the actual speciation times and why estimates of speciation events from trio-based germline mutation rates are often further back in time than the dating with fossil records.

The high levels of ILS in most nodes of the primate phylogeny made it possible to investigate the forces that shape genetic diversity along the genome in a complementary way to what has been done extensively using genome diversity data for individual species. We find that ILS depends strongly on the recombination rate, likely illustrating that a large part of genetic diversity is being removed by selection at linked sites. This dependency may partly explain Lewontin's paradox that the difference in genetic diversity across species is smaller than predicted from differences in neutral effective population sizes (*66*, *67*). The prevalence of natural selection at linked sites influencing diversity in ancestral nodes and thus ILS is also clear from the reduced ILS in introns compared with intergenic regions. The X chromosome appears to undergo more natural selection than the autosomes, perhaps as a consequence of male hemizygosity or possibly its strong role in male reproduction. Finally, ILS patterns also illuminate gene categories under balancing selection, particularly related to cornification or keratinization and immune functions, often experiencing different genealogical history compared with the speciation process.

Materials and methods summary

Data, alignment, and species tree

Our dataset consists of 50 primate species, including 27 newly sequenced ones and an outgroup, *Galeopterus variegatus*. For detailed information on sequencing and assembling, see the accompanying paper (*15*). We generated pairwise genome alignments using LASTZ (v1.04.00) for each species versus the human genome then using MULTIZ (v11.2) for multiway alignments. After removing columns of the alignment containing gaps in any of the species, we randomly chose half of the columns to run ExaML with the GAMMA model with 100 bootstraps. We report the tree with the highest maximum likelihood (fig. S1).

CoalHMM

We designed a divide-and-conquer, automated CoalHMM pipeline to fit a hidden Markov model where hidden states are four different topologies (Fig. 1A, inset), namely the species tree topology (V0), the deep coalescent topology following the species tree (V1), or one of two alternative topologies incongruent with the species tree (V2 and V3) (*14*). We defined each branch with a quartet of genomes and extracted them from the 51-way alignment using MafFilter (*68*). We removed columns containing only gaps and merged consecutive blocks that were <200 nucleotides apart. Chunks of <2000 nucleotides were filtered out, and blocks were divided into groups containing roughly 1 Mb alignment each (fig. S2B). CoalHMM was first run in a subset of 1-Mb groups of blocks, and the means of each of the estimated population parameters (tau1, tau2, theta1, theta2, c2, rho, and all the GTR model values) were recovered and used as starting parameters for the second CoalHMM run on all the other 1-Mb groups of blocks (fig. S2D). The posterior probabilities for each of the four hidden states were collected for each 1-Mb run and mapped to human coordinates (fig. S2E). All the code for processing the files and running CoalHMM is unified using a gwf workflow (https://gwf.app/), which can be accessed via https://github.com/rivasiker/autocoalhmm.

Genomic determinants of ILS

We used the latest deCODE human recombination map from Halldorsson *et al.* (*42*), the chimpanzee recombination map from Auton *et al.* (*45*), and the olive baboon recombination map from Sørensen *et al.* (*46*) to divide the genome into 10 equally sized recombination bins at a 100-kb resolution. We then calculated the mean ILS for each bin.

We retrieved intron and exon information from the knownGene UCSC Genome Browser table for hg38 (*69*–*71*) and kept only protein-coding genes that appear in the knownCanonical UCSC Genome Browser table (*72*). After trimming for size (supplementary materials, section 8), ILS level was calculated for exons and introns separately.

Introgression

We compared the level of divergence between sister species for segments of the genome attributed with the four different topologies to assess whether the level of incongruences that we report could be influenced by introgression. We used MafFilter to extract, filter, and concatenate segments and computed the percentage of mismatch as a measure of divergence between the sister species of the alignment (i.e., species 1 and species 2 for the segments assigned to the states V0 or V1, species 1 and species 3 for the segments assigned to the states V2, and species 2 and species 3 for the segments assigned to the states V3). We performed nonparametric Tukey-Kramer tests to compare the distribution of V0 segments versus V2 versus V3 segments divergence.

Population parameters reconstruction

The population parameters tau1, tau2, theta1, theta2, c2, and rho (defined as in fig. S20) outputted by CoalHMM are biased because of the use of a restricted set of four possible topologies to model the continuity of possible coalescence times (*14*), so we developed a machine learning–based procedure to learn how the different combinations of parameter values influence the bias of each parameter and then used this knowledge to predict the bias on real data. Briefly (but see supplementary materials, section 10), we ran CoalHMM on alignment blocks simulated under a grid of known combinations of population parameters using msprime (*73*). We then used the simulated versus estimated population parameters to train a random forest model and estimated the bias in our data on the basis of the estimates outputted by CoalHMM on the primate dataset.

dN/dS

We recovered 9972 coding gene alignments and filtered for orthologous genes where at least 41 out of the 50 primate species and the outgroup were present. Protein alignments were aligned using PRANK (*74*) and then filtered by Gblocks (*75*). Nucleotide alignments were generated by applying the protein alignment and site selection to the corresponding nucleotide sequences. We estimated branch-specific dN/dS ratios using the branch model of Codeml from PAML 4 (*23*). Results are reported in table S5.

Gene ontology

A gene ontology (GO) enrichment test was carried out for both high-ILS and low-ILS genes in each node using GOATOOLS (*76*). Gene annotations were downloaded from the National Center for Biotechnology Information's file transfer protocol (FTP) server (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz). For each branch, genes were assigned to be high in ILS if their exonic ILS was in the top 30%, whereas genes were classified as low ILS if they were in the bottom 30%. The significance level for the enrichment test was set to 0.05 after false discovery rate correction. A full list of the enriched gene ontology terms can be found in table S6.

REFERENCES AND NOTES

1. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, 1987).
2. T. Mailund, K. Munch, M. H. Schierup, Lineage sorting in apes. *Annu. Rev. Genet.* **48**, 519–535 (2014). doi: 10.1146/annurev-genet-120213-092532; pmid: 25251849
3. A. Suh, L. Smeds, H. Ellegren, The Dynamics of Incomplete Lineage Sorting across the Ancient Adaptive Radiation of Neoavian Birds. *PLOS Biol.* **13**, e1002224 (2015). doi: 10.1371/journal.pbio.1002224; pmid: 26284513
4. K. Wang *et al*., Incomplete lineage sorting rather than hybridization explains the inconsistent phylogeny of the wisent. *Commun. Biol.* **1**, 169 (2018). doi: 10.1038/s42003-018-0176-6; pmid: 30374461
5. F. Alda *et al*., Resolving Deep Nodes in an Ancient Radiation of Neotropical Fishes in the Presence of Conflicting Signals from Incomplete Lineage Sorting. *Syst. Biol.* **68**, 573–593 (2019). doi: 10.1093/sysbio/syy085; pmid: 30521024
6. A. Scally *et al*., Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012). doi: 10.1038/nature10842; pmid: 22398555
7. Z. N. Kronenberg *et al*., High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018). doi: 10.1126/science.aar6343; pmid: 29880660
8. Y. Mao *et al*., A high-quality bonobo genome refines the analysis of hominid evolution. *Nature* **594**, 77–81 (2021). doi: 10.1038/s41586-021-03519-x; pmid: 33953399
9. A. Hobolth, O. F. Christensen, T. Mailund, M. H. Schierup, Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLOS Genet.* **3**, e7 (2007). doi: 10.1371/journal.pgen.0030007; pmid: 17319744
10. A. Siepel, Phylogenomics of primates and their ancestral populations. *Genome Res.* **19**, 1929–1941 (2009). doi: 10.1101/gr.084228.108; pmid: 19801602
11. K. Prüfer *et al*., The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**, 527–531 (2012). doi: 10.1038/nature11128; pmid: 22722832
12. J. Y. Dutheil, K. Munch, K. Nam, T. Mailund, M. H. Schierup, Strong selective sweeps on the X chromosome in the human-chimpanzee ancestor explain its low divergence. *PLOS Genet.* **11**, e1005451 (2015). doi: 10.1371/journal.pgen.1005451; pmid: 26274919
13. N. Patterson, D. J. Richter, S. Gnerre, E. S. Lander, D. Reich, Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103–1108 (2006). doi: 10.1038/nature04789; pmid: 16710306
14. J. Y. Dutheil *et al*., Ancestral population genomics: The coalescent hidden Markov model approach. *Genetics* **183**, 259–274 (2009). doi: 10.1534/genetics.109.103010; pmid: 19581452
15. Y. Shao *et al*., Phylogenomic analyses provide insights into primate evolution. *Science* **380**, eabn6919 (2023). doi: 10.1126/science.abn6919
16. D. Vanderpool *et al*., Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLOS Biol.* **18**, e3000954 (2020). doi: 10.1371/journal.pbio.3000954; pmid: 33270638
17. J. Tung, L. B. Barreiro, The contribution of admixture to primate evolution. *Curr. Opin. Genet. Dev.* **47**, 61–68 (2017). doi: 10.1016/j.gde.2017.08.010; pmid: 28923540
18. N. Osada *et al*., Ancient genome-wide admixture extends beyond the current hybrid zone between *Macaca fascicularis* and *M. mulatta*. *Mol. Ecol.* **19**, 2884–2895 (2010). doi: 10.1111/j.1365-294X.2010.04687.x; pmid: 20579289
19. K. R. Veeramah *et al*., Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach. *Genetics* **200**, 295–308 (2015). doi: 10.1534/genetics.115.174425; pmid: 25769979
20. Y. Song *et al*., Genome-wide analysis reveals signatures of complex introgressive gene flow in macaques (genus *Macaca*). *Zool. Res.* **42**, 433–449 (2021). doi: 10.24272/j.issn.2095-8137.2021.038; pmid: 34114757
21. C. G. Schrago, H. N. Seuánez, Large ancestral effective population size explains the difficult phylogenetic placement of owl monkeys. *Am. J. Primatol.* **81**, e22955 (2019). doi: 10.1002/ajp.22955; pmid: 30779198
22. D. Silvestro *et al*., Early Arrival and Climatically-Linked Geographic Expansion of New World Monkeys from Tiny African Ancestors. *Syst. Biol.* **68**, 78–92 (2019). doi: 10.1093/sysbio/syy046; pmid: 29931325
23. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007). doi: 10.1093/molbev/msm088; pmid: 17483113
24. F. Ronquist *et al*., MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012). doi: 10.1093/sysbio/sys029; pmid: 22357727
25. R. Bouckaert *et al*., BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019). doi: 10.1371/journal.pcbi.1006650; pmid: 30958812
26. L. Bromham, The genome as a life-history character: Why rate of molecular evolution varies between mammal species. *Phil. Trans. R. Soc. B* **366**, 2503–2513 (2011). doi: 10.1098/rstb.2011.0014; pmid: 21807731
27. G. W. C. Thomas *et al*., Reproductive Longevity Predicts Mutation Rates in Primates. *Curr. Biol.* **28**, 3193–3197.e5 (2018). doi: 10.1016/j.cub.2018.08.050; pmid: 30270182
28. C. G. Schrago, The effective population sizes of the anthropoid ancestors of the human–chimpanzee lineage provide insights on the historical biogeography of the great apes. *Mol. Biol. Evol.* **31**, 37–47 (2014). doi: 10.1093/molbev/mst191; pmid: 24124206
29. W. C. Warren *et al*., The genome of the vervet (*Chlorocebus aethiops sabaeus*). *Genome Res.* **25**, 1921–1933 (2015). doi: 10.1101/gr.192922.115; pmid: 26377836
30. R. Burgess, Z. Yang, Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol. Biol. Evol.* **25**, 1979–1994 (2008). doi: 10.1093/molbev/msn148; pmid: 18603620
31. M. E. Steiper, E. R. Seiffert, Evidence for a convergent slowdown in primate molecular rates and its implications for the timing of early primate evolution. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 6006–6011 (2012). doi: 10.1073/pnas.1119506109; pmid: 22474376
32. J. J. O'Grady, D. H. Reed, B. W. Brook, R. Frankham, What are the best correlates of predicted extinction risk? *Biol. Conserv.* **118**, 513–520 (2004). doi: 10.1016/j.biocon.2003.10.002
33. M. Chintalapati, P. Moorjani, Evolution of the mutation rate across primates. *Curr. Opin. Genet. Dev.* **62**, 58–64 (2020). doi: 10.1016/j.gde.2020.05.028; pmid: 32634682
34. F. L. Wu *et al*., A comparison of humans and baboons suggests germline mutation rates do not track cell divisions. *PLOS Biol.* **18**, e3000838 (2020). doi: 10.1371/journal.pbio.3000838; pmid: 32804933
35. A. Hobolth, J. Y. Dutheil, J. Hawks, M. H. Schierup, T. Mailund, Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.* **21**, 349–356 (2011). doi: 10.1101/gr.114751.110; pmid: 21270173
36. K. Munch, K. Nam, M. H. Schierup, T. Mailund, Selective sweeps across twenty millions years of primate evolution. *Mol. Biol. Evol.* **33**, 3065–3074 (2016). doi: 10.1093/molbev/msw199; pmid: 27660295
37. J. Heer, N. Kong, M. Agrawala, "Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations" in *CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2009), pp. 1303–1312.
38. C. Perin, F. Vernier, J.-D. Fekete, "Interactive horizon graphs: Improving the compact visualization of multiple time series" in *CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2013), pp. 3217–3226.
39. M. W. Nachman, Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.* **17**, 481–485 (2001). doi: 10.1016/S0168-9525(01)02409-X; pmid: 11525814
40. F. Pratto *et al*., DNA recombination. Recombination initiation maps of individual human genomes. *Science* **346**, 1256442 (2014). doi: 10.1126/science.1256442; pmid: 25395542
41. B. Arbeithuber, A. J. Betancourt, T. Ebner, I. Tiemann-Boege, Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 2109–2114 (2015). doi: 10.1073/pnas.1416622112; pmid: 25646453
42. B. V. Halldorsson *et al*., Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019). doi: 10.1126/science.aau1043; pmid: 30679340
43. F. Pardo-Manuel de Villena, C. Sapienza, Recombination is proportional to the number of chromosome arms in mammals. *Mamm. Genome* **12**, 318–322 (2001). doi: 10.1007/s003350020005; pmid: 11309665
44. A. Kong *et al*., A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002). doi: 10.1038/ng917; pmid: 12053178
45. A. Auton *et al*., A fine-scale chimpanzee genetic map from population sequencing. *Science* **336**, 193–198 (2012). doi: 10.1126/science.1216872; pmid: 22422862
46. E. F. Sørensen *et al*., Genome-wide coancestry reveals details of ancient and recent male-driven reticulation in baboons. *Science* **380**, eabn8153 (2023). doi: 10.1126/science.abn8153
47. J. W. Ijdo, A. Baldini, D. C. Ward, S. T. Reeders, R. A. Wells, Origin of human chromosome 2: An ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 9051–9055 (1991). doi: 10.1073/pnas.88.20.9051; pmid: 1924367
48. K. Munch, T. Mailund, J. Y. Dutheil, M. H. Schierup, A fine-scale recombination map of the human–chimpanzee ancestor reveals faster change in humans than in chimpanzees and a strong impact of GC-biased gene conversion. *Genome Res.* **24**, 467–474 (2014). doi: 10.1101/gr.158469.113; pmid: 24190946
49. B. Vicoso, B. Charlesworth, Effective population size and the faster-X effect: An extended model. *Evolution* **63**, 2413–2426 (2009). doi: 10.1111/j.1558-5646.2009.00719.x; pmid: 19473388
50. C. Dubuc, A. Ruiz-Lambides, A. Widdig, Variance in male lifetime reproductive success and estimation of the degree of polygyny in a primate. *Behav. Ecol.* **25**, 878–889 (2014). doi: 10.1093/beheco/aru052; pmid: 25024637
51. A. Keinan, J. C. Mullikin, N. Patterson, D. Reich, Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat. Genet.* **41**, 66–70 (2009). doi: 10.1038/ng.303; pmid: 19098910
52. J. Prado-Martinez *et al*., Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013). doi: 10.1038/nature12228; pmid: 23823723
53. N. Osada *et al*., Finding the factors of reduced genetic diversity on X chromosomes of *Macaca fascicularis*: Male-driven evolution, demography, and natural selection. *Genetics* **195**, 1027–1035 (2013). doi: 10.1534/genetics.113.156703; pmid: 24026095
54. J. E. Pool, R. Nielsen, Population size changes reshape genomic patterns of diversity. *Evolution* **61**, 3001–3006 (2007). doi: 10.1111/j.1558-5646.2007.00238.x; pmid: 17971168
55. J. Bergman, M. Heide Schierup, Population dynamics of GC-changing mutations in humans and great apes. *Genetics* **218**, iyab083 (2021). doi: 10.1093/genetics/iyab083; pmid: 34081117
56. S. E. Santana, J. Lynch Alfaro, M. E. Alfaro, Adaptive evolution of facial colour patterns in Neotropical primates. *Proc. Biol. Sci.* **279**, 2204–2211 (2012). doi: 10.1098/rspb.2011.2326; pmid: 22237906
57. H. Rakotonirina, P. M. Kappeler, C. Fichtel, Evolution of facial color pattern complexity in lemurs. *Sci. Rep.* **7**, 15181 (2017). doi: 10.1038/s41598-017-15393-7; pmid: 29123214
58. J. G. Fleagle, *Primate Adaptation and Evolution* (Academic Press, 2013).
59. B. Jackson *et al*., Late cornified envelope family in differentiating epithelia—Response to calcium and ultraviolet irradiation. *J. Invest. Dermatol.* **124**, 1062–1070 (2005). doi: 10.1111/j.0022-202X.2005.23699.x; pmid: 15854049
60. D.-D. Wu, D. M. Irwin, Y.-P. Zhang, Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair. *BMC Evol. Biol.* **8**, 241 (2008). doi: 10.1186/1471-2148-8-241; pmid: 18721477
61. H. Niehues *et al*., Late cornified envelope (LCE) proteins: Distinct expression patterns of LCE2 and LCE3 members suggest nonredundant roles in human epidermis and other epithelia. *Br. J. Dermatol.* **174**, 795–802 (2016). doi: 10.1111/bjd.14284; pmid: 26556599
62. A. Cagan *et al*., Natural selection in the great apes. *Mol. Biol. Evol.* **33**, 3268–3283 (2016). doi: 10.1093/molbev/msw215; pmid: 27795229
63. K. J. Guinan *et al*., Signatures of natural selection and coevolution between killer cell immunoglobulin-like receptors (KIR) and HLA class I genes. *Genes Immun.* **11**, 467–478 (2010). doi: 10.1038/gene.2010.9; pmid: 20200544
64. V. Béziat, H. G. Hilton, P. J. Norman, J. A. Traherne, Deciphering the killer-cell immunoglobulin-like receptor system at super-resolution for natural killer and T-cell biology. *Immunology* **150**, 248–264 (2017). doi: 10.1111/imm.12684; pmid: 27779741
65. J. Bruijnesteijn, N. G. de Groot, R. E. Bontrop, The Genetic Mechanisms Driving Diversification of the *KIR* Gene Cluster in Primates. *Front. Immunol.* **11**, 582804 (2020). doi: 10.3389/fimmu.2020.582804; pmid: 33013938
66. R. C. Lewontin, *The Genetic Basis of Evolutionary Change* (Columbia Univ. Press, 1974).

67. R. B. Corbett-Detig, D. L. Hartl, T. B. Sackton, Natural selection constrains neutral diversity across a wide range of species. *PLOS Biol.* **13**, e1002112 (2015). doi: 10.1371/journal.pbio.1002112; pmid: 25859758
68. J. Y. Dutheil, in *Statistical Population Genomics*, J. Y. Dutheil, Ed., vol. 2090 of *Methods in Molecular Biology* (Humana, 2020), pp. 21–48.doi: 10.1007/978-1-0716-0199-0_2
69. W. J. Kent *et al*., The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). doi: 10.1101/gr.229102; pmid: 12045153
70. D. Karolchik *et al*., The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004). doi: 10.1093/nar/gkh103; pmid: 14681465
71. F. Hsu *et al*., The UCSC known genes. *Bioinformatics* **22**, 1036–1046 (2006). doi: 10.1093/bioinformatics/btl048; pmid: 16500937
72. J. Harrow *et al*., GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012). doi: 10.1101/gr.135350.111; pmid: 22955987
73. J. Kelleher, A. M. Etheridge, G. McVean, Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Comput. Biol.* **12**, e1004842 (2016). doi: 10.1371/journal.pcbi.1004842; pmid: 27145223
74. A. Löytynoja, in *Multiple Sequence Alignment Methods*, D. Russell, Ed., vol. 1079 of *Methods in Molecular Biology* (Humana Press, 2014), pp. 155–170.
75. G. Talavera, J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007). doi: 10.1080/10635150701472164; pmid: 17654362
76. D. V. Klopfenstein *et al*., GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018). doi: 10.1038/s41598-018-28948-z; pmid: 30022098
77. I. Rivas-González, rivasiker/autocoalhmm: v1.0.0, version 1.0.0, Zenodo (2022); https://doi.org/10.5281/zenodo.7277715.

ACKNOWLEDGMENTS

We thank GenomeDK for the computations; T. Bataillon, A. Hobolth, and M. Coll Macià for their valuable insights; and the two anonymous reviewers whose comments helped improve the manuscript. **Funding:** The study was supported by grants NNF18OC0031004 from the Novo Nordisk Foundation and 6108-00385 from the Independent Research Fund Denmark, Natural Sciences, to M.H.S. This work was also supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31020000), the International Partnership Program of the Chinese Academy of Sciences (no. 152453KYSB20170002), and the Villum Foundation (no. 25900) to G.Z. **Author contributions:** I.R.-G. and M.R. performed conceptualization, methodology, software, formal analysis, visualization, and writing. F.L. and L.Z. performed methodology, software, and formal analysis. Y.S. and D.W. performed data generation. J.Y.D. performed methodology, software, and revision. K.M. performed revision. M.H.S. and G.Z. performed conceptualization, writing, and supervision. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The primate alignment can be retrieved from Shao *et al*. (*15*). The code used to run CoalHMM on the primate alignment is available on Github (https://github.com/rivasiker/autocoalhmm) and Zenodo (*77*). The ILS tracts in 100-kb windows for each node can be retrieved from file S1.

SUPPLEMENTARY MATERIALS

View/request a protocol for this paper from *Bio-protocol*.

Submitted 28 November 2021; accepted 19 January 2023
10.1126/science.abn4409

Pervasive incomplete lineage sorting illuminates speciation and selection in primates

Iker Rivas-Gonzlez, Marjolaine Rousselle, Fang Li, Long Zhou, Julien Y. Dutheil, Kasper Munch, Yong Shao, Dongdong Wu, Mikkel H. Schierup, and Guojie Zhang

Science, **380** (6648), eabn4409.
DOI: 10.1126/science.abn4409