Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression

Dan Vanderpool[1]*, Bui Quang Minh[2,3], Robert Lanfear[3], Daniel Hughes[4], Shwetha Murali[4], R. Alan Harris[4,5], Muthuswamy Raveendran[4], Donna M. Muzny[4,5], Mark S. Hibbins[1], Robert J. Williamson[6], Richard A. Gibbs[4,5], Kim C. Worley[4,5], Jeffrey Rogers[4,5], Matthew W. Hahn[1]

[1]Department of Biology and Department of Computer Science, Indiana University, 1001 E. 3rd Street, Bloomington, Indiana, USA.

[2]Research School of Computer Science, Australian National University, 145 Science Road, Canberra, Australian Capital Territory, Australia.

[3]Research School of Biology, Australian National University, 46 Sullivans Creek Road, Canberra, Australian Capital Territory, Australia.

[4]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas, USA.

[5]Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas, USA.

[6] Department of Computer Science and Software Engineering and Department of Biology and Biomedical Engineering, Rose-Hulman Institute of Technology, 5500 Wabash Avenue, Terre Haute, Indiana, USA.

*danvand@indiana.edu

Abstract

Our understanding of the evolutionary history of primates is undergoing continual revision due to ongoing genome sequencing efforts. Bolstered by growing fossil evidence, these data have led to increased acceptance of once controversial hypotheses regarding phylogenetic relationships, hybridization and introgression, and the biogeographical history of primate groups. Among these findings is a pattern of recent introgression between species within all major primate groups examined to date, though little is known about introgression deeper in time. To address this and other phylogenetic questions, here we present new reference genome assemblies for three Old World Monkey species: *Colobus angolensis ssp. palliatus* (the black and white colobus), *Macaca nemestrina* (southern pig-tailed macaque), and *Mandrillus leucophaeus* (the drill). We combine these data with 23 additional primate genomes to estimate both the species tree and individual gene trees using thousands of loci. While our species tree is largely consistent with previous phylogenetic hypotheses, the gene trees reveal high levels of genealogical discordance associated with multiple primate radiations. We use strongly asymmetric patterns of gene tree discordance around specific branches to identify multiple instances of introgression between ancestral primate lineages. In addition, we exploit recent fossil evidence to perform fossil-calibrated molecular dating analyses across the tree. Taken together, our genome-wide data help to resolve multiple contentious sets of relationships among primates, while also providing insight into the biological processes and technical artifacts that led to the disagreements in the first place.

Introduction

Understanding the history of individual genes and whole genomes is an important goal for evolutionary biology. It is only by understanding these histories that we can understand the origin and evolution of traits—whether morphological, behavioral, or biochemical. Until recently, our ability to address the history of genes and genomes was limited by the availability of comparative genomic data. However, genome sequences are now being generated extremely rapidly. In primates alone, there are already 23 species with published reference genome sequences and associated annotations (SI Table 1), as well as multiple species with population samples of whole genomes [1–11]. These data can now be used to address important evolutionary questions.

Several studies employing dozens of loci sampled across broad taxonomic groups have provided rough outlines of the evolutionary relationships and divergence times among primates [12,13]. Due to the rapid nature of several independent radiations within primates, these limited data cannot resolve species relationships within some clades [12–14]. For instance, the New World Monkeys (NWM) experienced a rapid period of diversification ~15-18 million years ago (mya) [15] (Fig 1), resulting in ambiguous relationships among the three Cebidae subfamilies (Cebinae=squirrel monkeys and capuchins, Aotinae=owl monkeys, and Callitrichinae=marmosets and tamarins) [12–14,16–18]. High levels of incomplete lineage sorting (ILS) driven by short times between the divergence of distinct lineages have led to a large amount of gene

tree discordance in the NWM, with different loci favoring differing relationships among taxa. Given the known difficulties associated with resolving short internodes [19–21], as well as the multiple different approaches and datasets used in these analyses, the relationships among cebid subfamilies remain uncertain.

Fig 1. Species tree estimated using ASTRAL III with 1,730 gene trees (the *Mus musculus* outgroup was removed to allow for a visually finer scale). Common names for each species can be found in SI Table 1. Node labels indicate the bootstrap value from a maximum likelihood analysis of the concatenated dataset as well as the local posterior probability from the ASTRAL analysis. Gene concordance factors (gCF) and site concordance factors (sCF) are also reported. Eight fossil calibrations (blue stars; SI Table 6) were used to calibrate node ages. Grey bars indicate the minimum and maximum mean age from independent dating estimates. The inset tree with colored branches shows the maximum likelihood branch lengths estimated using a partitioned analysis of the concatenated alignment. Colors correspond to red = Strepsirrhini, cyan = Tarsiiformes, green = Platyrrhini (New World Monkeys), blue = Cercopithecoidea (Old World Monkeys), orange = Hominoidea (Apes).

In addition to issues of limited data and rapid radiations, a history of hybridization and subsequent gene flow between taxa means that there is no single dichotomously branching tree that all genes follow. Although introgression once was thought to be relatively rare (especially among animals, [22]), genomic studies have uncovered widespread patterns of recent introgression across the tree of life [23]. Evidence for recent or ongoing gene flow is especially common among the primates (e.g. [9,24–27]), sometimes with clear evidence for adaptive introgression (e.g. [28–30]). Whether widespread gene flow among primates is emblematic of their initial radiation (which began 60-75 mya, [13,31–33]) or is a consequence of current conditions—which include

higher environmental occupancy and more secondary contact—remains an open question [34].

Here we report the sequencing and annotation of three new primate genomes, all Old World Monkey (OWM) species: *Colobus angolensis ssp. palliatus* (the black and white colobus), *Macaca nemestrina* (southern pig-tailed macaque), and *Mandrillus leucophaeus* (the drill). Together with the published whole genomes of extant primates, we present a phylogenomic analysis including 26 primate species and several closely related non-primates. Incorporating recently discovered fossil evidence [35], we perform fossil-calibrated molecular dating analyses to estimate divergence times, including dates for the crown primates as well as the timing of more recent splits. Compared to recent hybridization, introgression that occurred between two or more ancestral lineages (represented by internal branches on a phylogeny) is difficult to detect. To get around this limitation, we modify a previously proposed method for detecting introgression [36] and apply it to our whole-genome datasets, finding additional evidence for gene flow among ancestral primates. Finally, we closely examine the genealogical patterns left behind by the NWM radiation, as well as the biases of several methods that have been used to resolve this topology. We use multiple approaches to provide a strongly supported history of the NWM and primates in general, while also highlighting the large amounts of gene tree discordance across the tree caused by ILS and introgression.

Results and Discussion

Primate Genome Sequencing

The three species sequenced here are all Old World Monkeys, and each is closely related to an already-sequenced species. This sampling scheme provides us increased power to detect introgression among each of the sub-clades containing these species. The assembly and annotation of each of the three species sequenced for this project are summarized here, with further details listed in Table 1. A summary of all published genomes used in this study, including links to the assemblies and NCBI BioProjects, is available in SI Table 2. All species were sequenced using standard methods according to Illumina Hi-seq protocols. Additional long-read sequencing was performed using Pacific Biosciences technology for *Macaca nemestrina*.

Table 1. Genomes sequenced in this study and associated assembly and annotation metrics.

Species name	Assembly Accession	Assembly Total length	Number of scaffolds	Scaffold N50 (mb)	Contig N50 (kb)	Protein-coding genes	BUSCO
Colobus angolensis ssp. palliatus (the black and white colobus)	GCF_000951035.1	2,970,124,662	13,124	7.84	38.36	20,222	95.82%
Macaca nemestrina (pig-tailed macaque)	GCF_000956065.1	2,948,703,511	9,733	15.22	106.89	21,017	95.98%
Mandrillus leucophaeus (drill)	GCF_000951045.1	3,061,992,840	12,821	3.19	31.35	20,465	95.45%

BUSCO percentages reflect the complete and fragmented genes relative to the Euarchontoglires ortholog database v9.

The sequencing effort for *Colobus angolensis ssp. palliatus* produced 514 Gb of data, which are available in the NCBI Short Read Archive (SRA) under the accession SRP050426 (BioProject PRJNA251421). The biological sample used for sequencing was kindly provided by Dr. Oliver Ryder (San Diego Zoo). Assembly of these data resulted in a total assembly length of 2.97 Gb in 13,124 scaffolds (NCBI assembly

Cang.pa_1.0; GenBank accession GCF_000951035.1) with an average per base coverage of 86.8X. Subsequent annotation via the NCBI Eukaryotic Genome Annotation Pipeline (annotation release ID: 100) resulted in the identification of 20,222 protein-coding genes and 2,244 non-coding genes. An assessment of the annotation performed using BUSCO 3.0.2 [37] in conjunction with the Euarchontoglires ortholog database 9 (https://busco-archive.ezlab.org/v3/datasets/euarchontoglires_odb9.tar.gz) indicated that 95.82% single-copy orthologs (91.68% complete, 4.13% fragmented) were present among the annotated protein-coding genes. Comprehensive annotation statistics for *C. angolensis ssp. palliatus* with links to the relevant annotation products available for download can be viewed at https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Colobus_angolensis_palliatus/100/.

For *Macaca nemestrina*, 1,271 Gb of data were produced (SRA accession SRP045960; BioProject PRJNA251427) resulting in an assembled genome length of 2.95 Gb in 9,733 scaffolds (Mnem_1.0; GenBank accession GCF_000956065.1). This corresponds to an average per base coverage of 113.1X when both short and long-read data are combined (Materials and Methods). The biological sample used for sequencing was kindly provided by Drs. Betsy Ferguson and James Ha (Washington National Primate Research Center). The NCBI annotation resulted in 21,017 protein coding genes and 13,163 non-coding genes (annotation release ID: 101). A BUSCO run to assess the completeness of the annotation (as above) indicated that 95.98% single-copy orthologs (92.23% complete, 3.75% fragmented) were present among the

annotated protein-coding genes. Comprehensive annotation statistics for *M. nemestrina* with links to the relevant annotation products available for download can be viewed at https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Macaca_nemestrina/101/.

Sequencing of *Mandrillus leucophaeus* libraries resulted in 334.1 Gb of data (SRA accession SRP050495; BioProject PRJNA251423) that once assembled resulted in a total assembly length of 3.06 Gb in 12,821 scaffolds (Mleu.le_1.0; GenBank accession GCF_000951045.1) with an average coverage of 117.2X per base. The biological sample used for sequencing was kindly provided by Dr. Oliver Ryder (San Diego Zoo). The NCBI annotation produced of 20,465 protein coding genes and 2,300 non-coding genes (annotation release ID: 100). A BUSCO run to assess the completeness of the annotation (as above) indicated that 95.45% single-copy orthologs (91.38% complete, 4.07% fragmented) were present among the annotated protein-coding genes. The full annotation statistics with links to the associated data can be viewed at https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Mandrillus_leucophaeus/100/.

Phylogenetic Relationships Among Primates

To investigate phylogenetic relationships among primates, we selected the longest isoform for each protein-coding gene from 26 primate species and 3 non-primate species (SI Table 1). After clustering, aligning, trimming, and filtering (Materials and Methods) there were 1,730 single-copy orthologs present in at least 27 of the 29 species (see SI Table 3 for the orthogroup, protein name, chromosome, and location of

each single-copy ortholog in the human genome). The cutoffs used to filter the dataset ensure high species coverage while still retaining a large number of orthologs. The coding sequences of these orthologs have an average length of 1,018 bp and 178 parsimony-informative characters per gene. Concatenation of these loci resulted in an alignment of 1,761,114 bp, with the fraction of gaps/ambiguities varying from 4.04% (*Macaca mulatta*) to 18.37% (*Carlito syrichta*) (SI Table 4).

We inferred 1,730 individual gene trees from nucleotide alignments using maximum likelihood in IQ-TREE 2 and then inferred a species tree using these gene tree topologies as input to ASTRAL III ([39]; Materials and Methods). We used the mouse, *Mus musculus*, as an outgroup to root the species tree. This approach resulted in a topology (which we refer to as "ML-ASTRAL"; Fig 1) that largely agrees with previously published phylogenies [12,13]. We also used IQ-TREE to carry out a maximum likelihood analysis of the concatenated nucleotide alignment (a topology we refer to as "ML-CONCAT"). This analysis resulted in a topology that differed from the ML-ASTRAL tree only with respect to the placement of *Aotus nancymaae* (owl monkey): rather than sister to the *Saimiri*+*Cebus* clade (as in Fig 1), the ML-CONCAT tree places *Aotus* sister to *Callithrix jacchus*, a minor rearrangement around a very short internal branch (Fig 1). All branches of the ML-ASTRAL species tree are supported by maximum local posteriors, the default support values provided by ASTRAL III [40], except for the branch that defines *Aotus* as sister to the *Saimiri*+*Cebus* clade (0.46 local posterior probability). Likewise, each branch in the ML-CONCAT tree is supported by

100% bootstrap values, including the branch uniting *Aotus* and *Callithrix*. We return to this conflict in the next section.

There has been some contention as to the placement of the mammalian orders Scandentia (treeshrews) and Dermoptera (colugos) [41–50]. The controversy concerns whether Dermoptera is sister to Primates, Scandentia is sister to Primates, or Dermoptera and Scandentia are sister groups. As expected, both the ML-ASTRAL and ML-CONCAT trees place these two groups outside the Primates with maximal statistical support (i.e. local posterior probabilities of 1.0 and bootstrap values of 100%; Fig 1); they also both point to Dermoptera as the closest sister lineage to the Primates [12,51–53]. However, while support values such as the bootstrap or posterior probability provide statistical confidence in the species tree topology, there can be large amounts of underlying gene tree discordance even for branches with 100% support (e.g. [54–56]). To assess discordance generally, and the relationships among the Primates, Scandentia, and Dermoptera in particular, we used IQ-TREE to calculate both gene (gCF) and site (sCF) concordance factors [57] for each internal branch of the topology in Figure 1. These two measures represent the fraction of genes and sites, respectively, that are in agreement with the species tree for any particular branch.

Examining concordance factors helps to explain previous uncertainty in the relationships among Primates, Scandentia, and Dermoptera (Fig 1). Although the bootstrap support is 100% and the posterior probability is 1.0 on the branch leading to the Primate common ancestor, the gene concordance factor is 45% and the site concordance factor is 39%. These values indicate that, of decisive gene trees (n=1663),

only 45% of them contain the branch that is in the species tree; this branch reflects the Primates as a single clade that excludes Scandentia and Dermoptera. While the species tree represents the single topology supported by the most gene trees (hence the strong statistical support for this branch), the concordance factors also indicate that a majority of gene tree topologies differ from the estimated species tree. In fact, the gCF value indicates that 55% of trees do not support a monophyletic Primate order, with either Dermoptera, Scandentia, or both lineages placed within Primates. Likewise, the sCF value indicates that only 39% of decisive sites in the total alignment support the branch uniting all primates, with 30% favoring Dermoptera as sister to the Primate sub-order Strepsirrhini and 31% placing Dermoptera sister to the Primate sub-order Haplorrhini. Similarly, only a small plurality of genes and sites have histories that place Dermoptera as sister to the Primates rather than either of the two alternative topologies (gCF=37, sCF=40; Fig 1), despite the maximal statistical support for these relationships. While discordance at individual gene trees can result from technical problems in tree inference (e.g. long-branch attraction, low phylogenetic signal, poorly aligned sequences, or model misspecification), it also often reflects biological causes of discordance such as incomplete lineage sorting and introgression. We further address the possible role of technical errors in generating patterns of discordance in the section entitled "Sources of Gene Tree Discordance" below.

Within the Primates, the phylogenetic affiliation of tarsiers (represented here by *Carlito syrichta*) has been debated since the first attempts by Buffon (1765) and Linnaeus (1767-1770) to systematically organize described species [58]. Two prevailing

hypotheses group tarsiers (Tarsiiformes) with either lemurs and lorises (the “prosimian” hypothesis, [59]) or with Simiiformes (the “Haplorrhini” hypothesis [60], where Simiiformes = Apes+OWM+NWM). The ML-ASTRAL and ML-CONCAT analyses place Tarsiiformes with Simiiformes, supporting the Haplorrhini hypothesis (Fig 1). The strepsirrhines come out as a well-supported group sister to the other primates. Again, our inference of species relationships is consistent with previous genomic analyses [61,62], but also highlights the high degree of discordance in this part of the tree. The rapid radiation of mammalian lineages that occurred in the late Paleocene and early Eocene [32] encompassed many of the basal primate branches, including the lineage leading to Haplorrhini. The complexity of this radiation is likely the reason for low gCF and sCFs (39.5% and 36%, respectively) for the branch leading to Haplorrhini, and perhaps explains why previous studies recovered conflicting resolutions for the placement of tarsiers [31,63,64].

The remaining branches of the species tree that define major primate clades all have remarkably high concordance with the underlying gene trees (gCF > 80%), though individual branches within these clades do not. The gCFs for the branches defining these clades are: Strepsirrhini (lemurs+lorises) = 84.5, Catarrhini (OWM+Apes) = 90.0, Platyrrhini (NWM) = 96.6, Hominoidea (Apes) = 82.7, and Cercopithecidae (OWM) = 92.3 (Fig 1). High gene tree/species tree concordance for these branches is likely due to a combination of more recent divergences (increasing gene tree accuracy) and longer times between branching events [65]. Within these clades, however, we see

multiple recent radiations. One of the most contentious has been among the New World Monkeys, a set of relationships we address next.

ML Concatenation Affects Resolution of the New World Monkey Radiation

Sometime during the mid-to-late Eocene (~45-34 mya), a small number of primates arrived on the shores of South America [15,66]. These monkeys likely migrated from Africa [66] and on arrival underwent multiple rounds of extinction and diversification [15]. Three extant families from this radiation now make up the New World Monkeys (Platyrrhini, Fig 1). Because of the rapidity with which these species spread and diversified across the new continent, relationships at the base of the NWM have been hard to determine [12–14,16–18].

As reported above, the concatenated analysis (ML-CONCAT) gives a different topology than the gene tree-based analysis (ML-ASTRAL). Specifically, the ML-CONCAT analysis supports a symmetrical tree, with *Aotus* sister to *Callithrix* (Fig 2A). In contrast, ML-ASTRAL supports an asymmetrical (or "caterpillar") tree, with *Aotus* sister to a clade comprised of *Saimiri*+*Cebus* (Fig 2B). There are reasons to have doubts about both topologies. It is well known that carrying out maximum likelihood analyses of concatenated datasets can result in incorrect species trees, especially when the time between speciation events is short [67,68]. In fact, the specific error that is made in these cases is for ML concatenation methods to prefer a symmetrical four-taxon tree over an asymmetrical one, exactly as is observed here. Gene tree-based methods such as ASTRAL are not prone to this particular error, as long as the underlying gene trees

are all themselves accurate [69,70]. However, if there is bias in gene tree reconstruction, then there are no guarantees as to the accuracy of the species tree. In addition, the ML-ASTRAL tree is supported by only a very small plurality of gene trees: there are 442 trees supporting this topology, compared to 437 supporting the ML-CONCAT topology and 413 supporting the third topology (Fig 2D). This small excess of supporting gene trees also explains the very low posterior support for this branch in the species tree (Fig 1). Additionally, a polytomy test [71], implemented in ASTRAL and performed using ML gene trees, failed to reject the null hypothesis of "polytomy" for the branch uniting *Aotus*+(*Saimiri,Cebus*) (P=0.47).

Fig 2. **The three most frequent topologies of New World Monkeys.** A) Tree 1 is the symmetrical topology inferred by the maximum likelihood concatenated analysis (ML-CONCAT) of 1,730 loci (1.76 Mb). B) Tree 2 is the asymmetrical topology inferred by ASTRAL III using either maximum likelihood (ML-ASTRAL) or maximum parsimony (MP-ASTRAL) gene tree topologies. Using maximum parsimony on the concatenated alignment also returns this tree (MP-CONCAT). C) Tree 3 is the alternative resolution recovered at high frequency in all gene tree analyses, though it is not the optimal species tree using any of the methods. D) Number of gene trees supporting each of the three resolutions of the NWM clade when maximum likelihood is used to infer gene tree topologies. There are 1,637 decisive gene trees for these splits. E) Gene tree counts when maximum parsimony is used to infer gene tree topologies. F) Number of parsimony informative sites in the concatenated alignment supporting each of the three resolutions.

To investigate these relationships further, we carried out additional analyses. The trees produced from concatenated alignments can be biased in situations with high ILS when maximum likelihood is used for inference, but this bias does not affect parsimony methods [21,72]. Therefore, we analyzed exactly the same concatenated 1.76 Mb alignment used as input for ML, but carried out a maximum parsimony analysis in PAUP* [73]. As would be expected given the known biases of ML methods, the

maximum parsimony tree (which we refer to as “MP-CONCAT”) returns the same tree as ML-ASTRAL, supporting an asymmetric topology of NWMs (Fig 2B). Underlying this result is a relatively large excess of parsimony-informative sites supporting this tree (Fig 2F), which results in maximal bootstrap values for every branch. The two most diverged species in this clade (*Saimiri* and *Callithrix*) are only 3.26% different at the nucleotide level, so there should be little effect of multiple substitutions on the parsimony analysis.

As mentioned above, gene tree-based methods (such as ASTRAL) are not biased when accurate gene trees are used as input. However, in our initial analyses we used maximum likelihood to infer the individual gene trees. Because protein-coding genes are themselves often a combination of multiple different underlying topologies [74], ML gene trees may be biased, and using them as input to gene tree-based methods may still lead to incorrect inferences of the species tree [75]. Therefore, we used the same 1,730 loci as above to infer gene trees using maximum parsimony with MPBoot [76]. Although the resulting topologies still possibly represent the average over multiple topologies contained within a protein-coding gene, using parsimony ensures that this average tree is not a biased topology. These gene trees were used as input to estimate a species tree using ASTRAL; we refer to this as the “MP-ASTRAL” tree. Once again, the methods that avoid known biases of ML lend further support to an asymmetric tree, placing *Aotus* sister to the *Saimiri*+*Cebus* clade (Fig 2B). In fact, the gene trees inferred with parsimony now show a much greater preference for this topology, with a clear plurality of gene trees supporting the species tree (473 vs. 417 supporting the second-most common tree; Fig 2E). As a consequence, the local

posterior for this branch in the MP-ASTRAL tree is 0.92 and the polytomy test performed using MP gene trees rejects (p = 0.037) the null hypothesis of "polytomy" for the branch uniting *Aotus*+(*Saimiri,Cebus*). The increased number of concordant gene trees using parsimony suggests that the gene trees inferred using ML may well have been suffering from the biases of concatenation when multiple trees are brought together (as observed in the Great Apes, [74]), reducing the observed levels of concordance.

A recent analysis of NWM genomes found *Aotus* sister to *Callithrix*, as in the ML-CONCAT tree, despite the use of gene trees to build the species tree [18]. However, the outgroup used in this analysis is a closely related species (*Brachyteles arachnoides*) that diverged during the NWM radiation and that shares a recent common ancestor with the ingroup taxa [12,13]. If the outgroup taxon used to root a tree shares a more recent common ancestor with subsets of ingroup taxa at an appreciable number of loci, the resulting tree topologies will be biased. A similar problem likely arose in previous studies that have used the Scandentia or Dermoptera as outgroups to Primates. In general, this issue highlights the difficulty in choosing outgroups: though we may have 100% confidence that a lineage lies outside our group of interest in the species tree, a reliable outgroup must also not have any discordant gene trees that place it inside the ingroup.

Sources of Gene Tree Discordance

As previously mentioned, there are both biological reasons for gene tree discordance (e.g. ILS or introgression) and technical reasons (e.g. long-branch attraction, homoplasy, low phylogenetic signal, poorly aligned sequences, or model misspecification). All of these phenomena may be reflected in gene and site concordance factors, but the proportion of discordance attributable to biological vs. technical factors is often difficult to ascertain. We therefore performed additional analyses to assess the impact of error on estimates of concordance factors.

In order to determine the degree to which short alignments or genes with low phylogenetic signal contribute to inaccurate gene trees, we recalculated gCFs and sCFs using the genes with the 200 longest alignments in our dataset (lengths ranging from 1,640 bp to 6,676 bp, with 116 to 2,101 parsimony informative sites). The resulting gCFs for the branch leading to the Primate common ancestor increases from 45% to 66% while the sCFs remain unchanged (SI Fig 1). For the branch placing Dermoptera sister to Primates, using trees estimated from the 200 longest alignments resulted in a modest increase in gCFs from 37% to 45%. Overall, the gCFs for the 200 longest genes were higher for all branches in the tree, with the average gCF increasing from 65.18% to 79.74%. The consistent increase in gCF but not sCF when using longer genes points to errors in gene tree inference as a small, but significant, factor in our dataset.

Using a single outgroup (mouse), could potentially lead to biases such as long-branch attraction near the base of the tree. To ameliorate these concerns, we performed an additional analysis using 150 randomly chosen single-copy orthologs, with pika (*Ochotona princeps*) included as a second outgroup. As in the full dataset,

maximum likelihood and parsimony were both applied to a concatenated dataset, and gene trees were also inferred via both ML and parsimony. Parsimony analysis of the concatenated alignment resulted in the same topology as in Fig 1, while a maximum likelihood analysis produced the same topology as the full ML-CONCAT tree from 1,730 loci, preferring a symmetric tree for the NWM clade. To assess the effect of including an additional outgroup on concordance factors, we calculated gCFs and sCFs using the 150 single-copy orthologs both with (SI Fig 2A) and without (SI Fig 2B) pika (using ML gene trees). In contrast to expectations about any error introduced by long-branch attraction, we observe slightly lower gCFs near the base of the tree when pika is included (SI Fig 2). Site concordance factors are not affected by the inclusion of pika. These analyses indicate that including additional outgroups when analyzing the full dataset is unlikely to reduce concordance factors or to change inferences of the species tree.

Technical errors leading to discordance should be more prominent deeper in the tree, as there is more opportunity for long-branch attraction, homoplasy, poor alignments, or model misspecification to cause problems. To determine whether concordance factors for deep branches in the primate tree are disproportionately affected by error, we looked for a correlation between concordance factors and the age of each bifurcation in the tree. For gCFs we found no correlation with node age (r^2=0.0094), while sCFs were slightly negatively correlated (r^2=0.2998, SI Fig 3). The negative correlation found between sCFs and node age is consistent with the expectation that substitutions occurring on deeper branches of the tree are more likely

to suffer from the effects of multiple substitutions (homoplasy). While there may still be technical factors affecting gCFs, true discordance throughout the tree is high enough to mask any such effect.

A recent simulation study [77] reported that negative selection, in combination with large differences in effective population size, can generate strong enough asymmetries in gene tree topologies that the most common topology does not match the species tree. Such an effect, if real, would mislead both gene-tree-based and concatenation-based approaches to species tree inference. However, previous theoretical results predict that there should be no effect of negative selection on the distribution of tree topologies [78–81] and the new results were obtained using custom simulation software. To clarify this issue, we used the open-source simulator SLiM [82] to study non-recombining loci under the most extreme parameters used by He et al. [77] (see Materials and Methods). We found no evidence for the bias in gene tree frequencies recently reported (SI Fig 4A). However, we observed fewer than one mutation per locus at the end of our simulations under the parameters exactly replicating He et al. [77], suggesting we may not have generated sufficient deleterious variation to observe the effect. To address this, we simulated the same conditions but with the deleterious mutation rate increased two orders of magnitude, and still did not observe a bias in topology frequencies (SI Fig 4B). Our results therefore indicate that weak negative selection does not generate gene tree discordance, consistent with population genetic theory [78–81].

Strongly Supported Divergence Times Using Fossil Calibrations

421 Fossil-constrained molecular dating was performed using 10 independent
422 datasets, each of which consisted of 40 protein-coding genes randomly selected
423 (without replacement) and concatenated. The resulting datasets had an average
424 alignment length of 39,374 bp (SD=2.6x10^3, SI Table 5). Although individual discordant
425 trees included in this analysis may have different divergence times, the difference in
426 estimates of dates should be quite small [83]. We used eight dated fossils (blue stars in
427 Fig 1) from 10 studies for calibration (SI Table 6). The most recent of these fossils is
428 ~5.7 mya [84], while the most ancient is 55.8 mya [85]. Each separate dataset and the
429 same set of "soft" fossil constraints, along with the species tree in Figure 1, were used
430 as input to PhyloBayes 3.3 [86] which was run twice to assess convergence (Materials
431 and Methods).

432 We observed tight clustering of all estimated node ages across datasets and
433 independent runs of PhyloBayes (Fig 3 and SI Table 6). In addition, the ages of most
434 major crown nodes estimated here are largely in agreement with previously published
435 age estimates (Table 2). Some exceptions include the age of the crown Strepsirrhini
436 (47.4 mya) and Haplorrhini (59.0 mya) which are more recent than many previous
437 estimates for these nodes (range in the literature is Strepsirrhini = 51.6 - 68.7,
438 Haplorrhini = 60.6 – 81.3, see Table 2). The crown nodes for Catarrhini, Hominoidea,
439 and Cercopithecidae (28.4, 21.4, and 16.8 mya, respectively) all fall within the range of
440 variation recovered in previous studies (Table 2).

441 **Fig 3**. **Mean node ages for independent Phylobayes dating runs on 10 different**
442 **datasets (each dataset was run twice).** Box plots show the median, interquartile
443 range, and both minimum and maximum values of the mean nodes ages. An additional

run was performed with no sequence data to ascertain the prior on node divergence times in the presence of fossil calibrations (pink asterisks). Some prior ages were too large to include in the plot while still maintaining detail; these ages are given as numeric values. The species tree topology is from Figure 1, 95% highest posterior density (HPD) intervals for each node are reported in SI Table 7.

Table 2. Mean crown node divergence times estimated in this study compared with mean divergences times estimated by eight prior studies.

Node	This Study	This Study, No Max*	This Study, Concord†	Herrera *et al.* [32]	Kistler *et al.* [33]	Perez *et al.* [17]	Springer *et al.* [13]	Meredith *et al.* [45]	Perelman *et al.* [12]	Wilkinson *et al.* [87]	Chatterjee *et al.* [31]
Primates	61.7	67.5	63	63.9	68	NA	67.8	71.5	87.2	84.5	63.7
Strepsirrhini	47.4	50.2	48.4	61.4	59	NA	54.2	55.1	68.7	49.8	51.6
Haplorrhini	59.0	63.8	59.8	61.9	67	60.6	61.2	62.4	81.3	NA	NA
Catarrhini	28.4	29.0	27.2	32.1	33	27.8	25.1	20.6	31.6	31.0	29.3
Hominoidea	21.4	21.6	19.9	NA	21	18.44	17.4	14.4	20.3	NA	21.5
Cercopithecidae	16.8	16.9	14.2	NA	24	13.4	13.2	NA	17.6	14.1	23.4

Estimates were calculated by averaging the mean times across all runs for 10 independent datasets.

*Refers to the average divergence time of the crown node for the indicated taxonomic group when the 65.8 my maximum constraint was removed from the Primate node.

†Refers to the average divergence time of the crown node for the indicated taxonomic group when divergence times were estimated using the most concordant gene trees.

Our estimate for the most recent common ancestor of the extant primates (i.e. the last common ancestor of Haplorrhini and Strepsirrhini) is 61.7 mya, which is slightly more recent than several studies [13,31,33,88] and much more recent than other studies [12,87,89] (Table 2). However, our estimate is in good agreement with Herrera *et al.* [32], who used 34 fossils representing extinct and extant lineages (primarily Strepsirrhines) to infer divergence times among primates, concluding that the split occurred approximately 64 mya. Despite limited overlap in taxon sampling, one similarity between our study and that of Herrera *et al.* is that we have both used the maximum constraint of 65.8 my on the ancestral primate node suggested by Benton *et*

469 *al.* [90], which likely contributes to the more recent divergence. It is worth noting that the
470 soft bounds imposed in our analysis permit older ages to be sampled from the Markov
471 chain, but these represented only a small fraction (median 3.37%) of the total sampled
472 states after burn-in (SI Table 6). To determine the effects of imposing the 65.8 my
473 maximum constraint on the Primate node, we analyzed all 10 datasets for a third time
474 with this constraint removed and report the divergence time of major primate clades in
475 Table 2 ("No Max" entries). However, it may be that using genes that have gene trees
476 most similar to the topology being dated will reduce bias caused by concatenation [74].
477 To determine whether using concordant loci has an impact on the estimated dates, we
478 constructed an eleventh dataset consisting of ~43 kb from the 20 loci most similar to the
479 species tree in Figure 1 (as determined by Robinson-Foulds distances). There was no
480 consistent difference in the dates estimated with this dataset ("Concord" entries in Table
481 2).

482 There are several caveats to our age estimates that should be mentioned.
483 Maximum age estimates for the crown node of any given clade are defined by the oldest
484 divergence among sampled taxa in the clade. This limitation results in underestimates
485 for nearly all crown node ages as, in practice, complete taxon sampling is difficult to
486 achieve. Fossil calibrations are often employed as minimum constraints in order to
487 overcome the limitations imposed by taxon sampling, allowing older dates to be
488 estimated more easily. On the other hand, the systematic underestimation of crown
489 node ages due to taxon sampling is somewhat counteracted by the overestimation of
490 speciation times due to ancestral polymorphism. Divergence times estimated from

sequence data represent the coalescence times of sequences, which are necessarily older than the time at which two incipient lineages diverged [91,92]. This overestimation will have a proportionally larger effect on recent nodes (such as the *Homo/Pan* split; Fig 3, node 15), but the magnitude can be no larger than the average level of polymorphism in ancestral populations and will be additionally reduced by post-divergence gene flow.

Introgression During the Radiation of Primates

There is now evidence for recent inter-specific gene flow between many extant primates, including introgression events involving humans [24], gibbons [93,94], baboons [9,27], macaques [95,96], and vervet monkeys [10], among others. While there are several widely used methods for detecting introgression between closely related species (see chapters 5 and 9 in [97]), detecting ancient gene flow is more difficult. One of the most popular methods for detecting recent introgression is the *D* test (also known as the "ABBA-BABA" test; [98]). This test is based on the expectation that, for any given branch in a species tree, the two most frequent alternative resolutions should be present in equal proportions. However, the *D* test uses individual SNPs to evaluate support for alternative topologies, and explicitly assumes an infinite sites model of mutation (i.e. no multiple hits). As this assumption will obviously not hold the further back in time one goes, a different approach is needed.

Fortunately, Huson *et al*. [36] described a method that uses gene trees themselves (rather than SNPs) to detect introgression. Using the same expectations as in the *D* test, these authors looked for a deviation from the expected equal numbers of

512 alternative tree topologies using a test statistic they refer to as Δ. As far as we are
513 aware, Δ has only rarely been used to test for introgression in empirical data, possibly
514 because of the large number of gene trees needed to assess significance, or the
515 assumptions of the parametric method proposed to obtain *P*-values. Here, given our
516 large number of gene trees and large number of internal branches to be tested, we
517 adapt the Δ test for genome-scale data.

518 To investigate patterns of introgression within primates, we used 1,730 single-
519 copy loci to test for deviations from the null expectation of Δ on each of the 24 internal
520 branches of the primate phylogeny (Materials and Methods). To test whether deviations
521 in Δ were significant (i.e. $\Delta > 0$), we generated 2000 resampled datasets of 1,730 gene
522 tree topologies each. *P*-values were calculated from *Z*-scores generated from these
523 resampled datasets. Among the 17 branches where at least 5% of topologies were
524 discordant, we found 7 for which Δ had $P<0.05$.

525 To further verify these instances of potential introgression, for each of these
526 seven branches we increased the number of gene trees used, as well as the alignment
527 length for each locus, by subsampling a smaller set of taxa. We randomly chose four
528 taxa for each internal branch tested that also had this branch as an internal branch, and
529 then aligned all orthologs present in a single copy in each taxon. These steps resulted
530 in ~3,600-6,400 genes depending on the branch being tested (SI Table 8). Additionally,
531 because instances of hybridization and introgression are well documented among
532 macaques [96,99,100], we similarly re-sampled orthologs from the three *Macaca*
533 species in our study.

We recalculated Δ using the larger gene sets and found significant evidence (after correcting for m=17 multiple comparisons by using a cutoff of P = 0.00301) for six introgression events, all of which occurred among the Papionini (Fig 4 and see next paragraph). Within the Hominoidea, we found Δ = 0.0518 for the branch leading to the great apes (P = 0.030). The asymmetry in gene tree topologies here suggests gene flow may have happened between gibbons (represented by *Nomascus*) and the ancestral branch leading to the African hominoids (humans, chimpanzees, and gorillas), but, like the D test, Δ cannot tell us the direction of introgression. Although currently separated by significant geographic distances (African apes south of the Sahara Desert and gibbons all in southeast Asia), it is worth noting that fossil hominoids dating from the early to late Miocene had a broad distribution extending from southern Africa to Europe and Asia [101]. Support for introgression between ancestral hominins and ancestral chimpanzees has been previously reported [102]; our four-taxon analyses found marginal support for this conclusion (Δ = 0.0917, P = 0.055).

Fig 4. Introgression among Papionini taxa (the species tree is unrooted for clarity). Arrows indicate that a significant Δ was found in our four taxon tests and identify the two lineages inferred to have exchanged genes (values underlying these tests are listed in SI Table 8). Among the Papionini, there was evidence of introgression between African taxa (*Papio*, *Theropithecus*, and *Cercocebus*) and Asian *Macaca* species (light grey arrows). Introgression events likely occurred between African taxa and the ancestral *Macaca*, which had a wide distribution across northern Africa prior to the radiation throughout Asia 2-3 mya [103]. More recent instances of introgression are inferred between macaque species and among the African Papionini (dark grey arrows).

Within the OWM, ~40% of Cercopithicine species are known to hybridize in nature [34]. Consistent with this, *Macaca nemestrina* and *M. fascicularis* showed a strong signature of gene flow in our data (Δ = 0.1761, P = 1.377e-09). These two

560 species have ranges that currently overlap (SI Fig 5). In contrast to the clear signal of
561 recent gene flow in the macaques, we detected a complex pattern of ancient
562 introgression between the African Papionini (*Cercocebus, Mandrillus, Papio, and*
563 *Theropithecus*) and the Asian Papionini (*Macaca*) (Fig 4). The Δ test was significant
564 using multiple different subsamples of four taxa, suggesting multiple ancestral
565 introgression events. An initial attempt to disentangle these events using Phylonet
566 v3.8.0 [104] with the seven Papionini species and an outgroup was unsuccessful, as
567 Phylonet failed to converge on an optimal network for these taxa. An attempt to infer the
568 network with SNaQ [105] gave similarly ambiguous results. When there are multiple
569 episodes of gene flow within a clade, even complex computational machinery may be
570 unable to infer the correct combination of events.

571 As an alternative approach, we used four-taxon trees to estimate Δ for each
572 *Macaca* species paired with two African Papionini (one from the *Papio*+*Theropithecus*
573 clade and one from the *Mandrillus* +*Cercocebus* clade; see SI Table 8) and an
574 outgroup. Significant introgression was detected using each of the *Macaca* species and
575 three of the four African Papionini species (*Cercocebus, Theropithecus,* and *Papio*).
576 These results suggest gene flow between the ancestor of the three *Macaca* species in
577 our analysis and the ancestors of the three African Papionini in our analysis, or one
578 introgression event involving the ancestor of all four African species coupled with a
579 second event that masked this signal in *Mandrillus*. This second event may either have
580 been biological (additional introgression events masking the signal), or technical
581 (possibly the lack of continuity or completeness of the *Mandrillus* reference genome

582 sequence), but in either case we could not detect introgression in the available drill
583 sequence. The latter scenario would fit better with the current geographic distributions
584 of these species, as they are on two different continents. However, the fossil record
585 indicates that by the late Miocene to late Pleistocene the ancestral distribution of the
586 genus *Macaca* covered all of North Africa, into the Levant, and as far north as the U.K.
587 (SI Fig 5; [106]). The fossil record for *Theropithecus* indicates several species had
588 distributions that overlapped with *Macaca* during this time, including in Europe and as
589 far east as India (SI Fig 5, [107,108]). Ancestral macaques and ancestral papionins may
590 therefore have come into contact in the area of the Mediterranean Sea. The Sahara
591 Desert is also responsible for the current disjunct distributions of many of these species.
592 However, this region has experienced periods of increased rainfall or "greenings" over
593 the past several million years [109–111]. Faunal migration through the Sahara, including
594 by hominins, is hypothesized to have occurred during these green periods
595 [110,112,113] resulting in successive cycles of range expansion and contraction [114].
596 Hybridization and introgression could have occurred between the ancestors of these
597 groups during one of these periods.

598 Our results on introgression come with multiple caveats, both about the events
599 we detected and the events we did not detect. As with the *D* test, there are multiple
600 alternative explanations for a significant value of Δ besides introgression. Ancestral
601 population structure can lead to an asymmetry in gene tree topologies [115] though it
602 requires a highly specific, possibly unlikely population structure. For instance, if the
603 ancestral population leading to *Macaca nemestrina* was more closely related to *M.*

604 *fascicularis* than was the ancestral population leading to its sister species, *M. mulatta*
605 (Fig 4), then there could be an unequal number of alternative topologies. Similarly, any
606 bias in gene tree reconstruction that favors one alternative topology over the other could
607 potentially lead to a significant value of Δ. While this scenario is unlikely to affect recent
608 divergences using SNPs, well known biases that affect topology reconstruction deeper
609 in the tree (such as long-branch attraction) could lead to gene tree asymmetries.
610 However, we did not observe any significant Δ-values for branches more than ~10 my
611 old. One alternative approach to avoid biases in reconstruction could be the use of
612 transposon insertions or other rare genomic changes (cf. [116,117]). Future analyses
613 that compare these different approaches to detecting introgression would be especially
614 useful.

615 There are also multiple reasons why our approach may have missed
616 introgression events, especially deeper in the tree. All methods that use asymmetries in
617 gene tree topologies miss gene flow between sister lineages, as such events do not
618 lead to changes in the proportions of underlying topologies. Similarly, equal levels of
619 gene flow between two pairs of non-sister lineages can mask both events, while even
620 unequal levels will lead one to miss the less-frequent exchange. More insidiously,
621 especially for events further back in time, extinction of the descendants of hybridizing
622 lineages will make it harder to detect introgression (though extinction of donor lineages
623 is less of a problem than extinction of lineages receiving migrants). Internal branches
624 closer to the root will be on average longer than those near the tips because of
625 extinction [118], and therefore introgression between non-sister lineages would have to

626 occur longer after speciation in order to be detected. For instance, gene flow among
627 Strepsirrhine species has been detected in many previous analyses of more closely
628 related species (e.g. [119–122]) but the deeper relationships among the taxa sampled
629 here may have made it very difficult to detect introgression. Nevertheless, our analyses
630 were able to detect introgression between many primate species across the phylogeny.

631 Conclusions

632 Several previous phylogenetic studies of primates have included hundreds of
633 taxa, but fewer than 70 loci [12,13]. While the species tree topologies produced by
634 these studies are nearly identical to the one recovered in our analysis, the limited
635 number of loci meant that it was difficult to assess gene tree discordance accurately. By
636 estimating gene trees from 1,730 single-copy loci, we were able to assess the levels of
637 discordance present at each branch in the primate phylogeny. Understanding
638 discordance helps to explain why there have been longstanding ambiguities about
639 species relationships near the base of primates and in the radiation of New World
640 Monkeys. Our analyses reveal how concatenation of genes—or even of exons—can
641 mislead maximum likelihood phylogenetic inference in the presence of discordance, but
642 also how to overcome these biases. Discordance also provides a window into
643 introgression among lineages, and here we have found evidence for exchange among
644 several species pairs. Each instance of introgression inferred from the genealogical
645 data is plausible insofar as it can be reconciled with current and ancestral species
646 distributions.

Materials and Methods

Source Material and Sequencing

For the sequencing of the *Colobus angolensis palliatus* genome, paired-end (100 bp) libraries were prepared using DNA extracted from heart tissue (isolate OR3802 from the San Diego Zoo). Sequencing was performed using nine Illumina Hi-seq 2000 lanes and four Illumina Hi-seq 2500 lanes with subsequent assembly carried out using ALLPATHS-LG software (v. 48744) [123]. Additional scaffolding and gap-filling was performed using Atlas-Link v. 1.1 (https://www.hgsc.bcm.edu/software/atlas-link) and Atlas-GapFill v. 2.2. (https://www.hgsc.bcm.edu/software/atlas-gapfill) respectively. Annotation for all three species was carried out using the NCBI Eukaryotic Genome Annotation Pipeline. A complete description of the pipeline can be viewed at https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/.

For the sequencing of the *Macaca nemestrina* genome, DNA was extracted from a blood sample (isolate M95218 from the Washington National Primate Research Center). Paired-end libraries were prepared and sequenced on 20 Illumina Hi-Seq 2000 lanes with the initial assembly performed using ALLPATHS-LG as above. Scaffolding was conducted using Atlas-Link v. 1.1. Additional gap-filling was performed using the original Illumina reads and Atlas-GapFill v. 2.2, as well as long reads generated using the Pacific Biosciences RS (60 SMRT cells) and RSII (50 SMRT cells) platforms. The PacBio reads were mapped to scaffolds to fill remaining gaps in the assembly using PBJelly2 (v. 14.9.9) [124].

For the sequencing of the *Mandrillus leucophaeus* genome, DNA was extracted from heart tissue (isolate KB7577 from the San Diego Zoo). Paired-end libraries were prepared and sequenced on nine Illumina Hi-Seq 2000 lanes with the initial assembly performed using ALLPATHS-LG as above. Additional scaffolding was completed using Atlas-Link v. 1.1 and additional gap-filling in scaffolds was performed using the original Illumina reads and Atlas-GapFill v. 2.2.

Phylogenomic Analyses

The full set of protein-coding genes for 26 primates and 3 non-primates were obtained by combining our newly sequenced genomes with already published data (see SI Table 1 for references and accessions and Tables 1 and SI Table 2 for genome statistics). Ortholog clustering was performed by first executing an all-by-all BLASTP search [125,126] using the longest isoform of each protein coding gene from each species. The resulting BLASTP output was clustered using the mcl algorithm [127] as implemented in FastOrtho [128] with various inflation parameters (the maximum number of clusters was obtained with *inflation=5*). Orthogroups were then parsed to retain those genes present as a single-copy in all 29 taxa (1,180 genes), 28 of 29 taxa (1,558 genes), and 27 of 29 taxa (1,735 genes). We chose to allow up to two missing species per alignment to maximize the data used in our phylogenomic reconstructions while maintaining high taxon-occupancy in each alignment.

Coding sequences (CDS) for each single-copy orthogroup were aligned, cleaned, and trimmed via a multi-step process: First, sequences in each orthogroup were aligned

by codon using GUIDANCE2 [129] in conjunction with MAFFT v7.407 [130] with 60 bootstrap replicates. GUIDANCE2 uses multiple bootstrapped alignments to generate quality scores for each column in the final alignment as well as for each taxon sequence in each alignment. Sequence residues in the resulting MAFFT alignment with GUIDANCE scores < 0.93 were converted to gaps and sites with > 50% gaps were removed using Trimal v1.4.rev22 [131]. Alignments shorter than 200 bp (full dataset) or 300 bp (four-taxon tests for introgression), and alignments that were invariant or contained no parsimony informative characters, were removed from further analyses. Alignments with high numbers of discordant sites were further inspected for errors and removed from the analysis when warranted. This resulted in 1,730 loci for the full analysis (see SI Table 8 for gene counts used in four-taxon tests).

IQ-TREE v2-rc1 was used with all 1,730 aligned loci to estimate a maximum likelihood concatenated (ML-CONCAT) tree with an edge-linked, proportional-partition model and 1,000 ultrafast bootstrap replicates. This strategy uses ModelFinder [134] to automatically find the best-fit model for each ortholog alignment (partition). Branch lengths are shared between partitions, with each partition having its own rate that rescales branch lengths, accommodating different evolutionary rates between partitions. The full IQ-TREE commandline used was: “iqtree -p Directory_of_Gene_Alignments --prefix -m MFP -c 8 -B 1000”. Maximum likelihood gene trees were estimated for each alignment with nucleotide substitution models selected using ModelFinder [134] as implemented in IQ-TREE. The full IQ-TREE commandline used was: “iqtree -s Directory_of_Gene_Alignments --prefix -m MFP -c 8”. We used the resulting maximum

likelihood gene trees to estimate a species tree using ASTRAL III (ML-ASTRAL) [38]. Parsimony gene trees were generated using MPboot [75] and used to estimate a species tree using ASTRAL III (MP-ASTRAL), while PAUP* [72] was used to estimate the concatenated parsimony tree (MP-CONCAT) with 500 bootstrap replicates. IQ-TREE was used to calculate both gene concordance factors (gCFs) and site concordance factors (sCFs), with sCFs estimated from 300 randomly sampled quartets using the commandline: "iqtree --cf-verbose --gcf 1730_GENETREE.treefile -t Species_tree_file --df-tree --scf 300 -p Directory_of_Gene_Alignments -c 4".

Effects of selection on gene tree distributions

We performed 100 replicate simulations for each mutation rate condition using SLiM version 3.3.1 [81], with tree sequence recording turned on and no neutral mutations. Each replicate simulation consisted of 50 non-recombining loci of 1 kb each, with free recombination between loci, for three populations with the phylogenetic relationship ((p2,p3),p1). These simulations closely match the population genetic parameters under the most extreme asymmetry condition reported in He et al. [76], with population sizes, selection coefficients, and mutation rates rescaled two orders of magnitude for performance (SLiM recipes are available via Data Dryad https://doi.org/10.5061/dryad.rfj6q577d [135]). These parameters include: a per-locus deleterious mutation rate of 3×10^{-7} per generation; a population-scaled selection coefficient (Ns) of -7.5; an internal branch subtending p2 and p3 of $0.01N$ generations (where N is the population size of p1 and p2); a population size for p3 that is 0.04 times that of p1 and p2; and tip branch lengths of $8N$ generations. In the higher mutation rate

condition, the per-locus rate was increased to 3 x 10^{-5} per generation. We randomly sampled one chromosome from each population at each locus at the end of the simulation, and obtained the genealogy of these samples recorded in the tree sequence at the locus.

Introgression Analyses

For each internal branch of the primate tree where the proportion of discordant trees was > 5% of the total, concordance factors were used to calculate the test statistic Δ, where:

$$\Delta = \frac{Number\ of\ DF1\ trees - Number\ of\ DF2\ trees}{Number\ of\ DF1\ trees + Number\ of\ DF2\ trees}$$

Where *DF1* trees represent the most frequent discordant topology and *DF2* trees are the second most frequent discordant topology. This is a normalized version of the statistic proposed by Huson *et al.* [36], which only included the numerator of this expression. Note also that, by definition, Δ here is always equal to or greater than 0. To test whether deviations from zero were significant (i.e. $\Delta > 0$), we calculated Δ for 2,000 pseudo-replicate datasets generated by resampling gene trees with replacement. The resulting distribution was used to calculate *Z*-scores and the resulting *P*-values for the observed Δ value associated with each branch tested [136]. Of the 17 internal branches where > 5% of topologies were discordant, 7 were significant at $P < 0.05$, and selected for more extensive testing. For each of the 7 significant branches in the all-Primates

753 tree, 4 taxa were selected that included the target branch as an internal branch. Single-
754 copy genes present in each taxon were aligned as previously described. Alignments
755 with no variant or parsimony-informative sites were removed from the analysis and gene
756 trees were estimated using maximum likelihood in IQ-TREE 2. The test statistic, Δ, was
757 calculated and significance was again determined using 2,000 bootstrap replicates with
758 the *P*-value threshold for significance corrected for multiple comparisons (*m*=17) using
759 the Dunn–Šidák correction [137,138].

760 **Molecular Dating**

761 Molecular dating analyses were performed on 10 datasets consisting of 40 CDS
762 alignments each sampled randomly without replacement from the 1,730 loci used to
763 estimate the species tree. Gene alignments were concatenated into 10 supermatrices
764 ranging from 36.7 kb – 42.7 kb in length (see SI Table 5 for the length of each
765 alignment). Each dataset was then analyzed using PhyloBayes 3.3 [85] with sequences
766 modeled using a site-specific substitution process with global exchange rates estimated
767 from the data (CAT-GTR; [139]). Among-site rate-variation was modeled using a
768 discrete gamma distribution with six rate categories. A relaxed molecular clock [140]
769 with eight, soft-bounded, fossil calibrations (see SI Table 6) was used to estimate
770 divergence times on the fixed species tree topology (Fig 1), the analyses were executed
771 using the following command line: pb -x 1 15000 -d Alignment.phy -T Tree_file.tre -r
772 outgroup_file.txt -cal 8_fossil.calib -sb -gtr -cat -bd -dgam 6 -ln -rp 90 90. Each dataset
773 was analyzed for 15,000 generations, sampling every 10 generations, with 5,000
774 generations discarded as burn-in. Each dataset was analyzed twice to ensure

convergence of the average age estimated for each node (Fig 3 shows the node age for both runs). To determine the effect of including a maximum constraint on the root of the Primates, we analyzed each dataset a third time with this constraint removed. Both the constrained and unconstrained node ages for major groups within the Primates are reported in Table 2.

All data deposited in the Dryad repository:https://doi.org/10.5061/dryad.rfj6q577d [135].

Acknowledgements

We thank Yue Liu for assistance in assembling the genomes, and Fábio Mendes and Gregg Thomas for helpful advice.

References

1. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2011;473: 544–544. doi:10.1038/nature09991

2. Fawcett GL, Raveendran M, Deiros DR, Chen D, Yu F, Harris RA, et al. Characterization of single-nucleotide variation in Indian-origin rhesus macaques (*Macaca mulatta*). BMC Genomics. 2011;12: 311. doi:10.1186/1471-2164-12-311

3. Higashino A, Sakate R, Kameoka Y, Takahashi I, Hirata M, Tanuma R, et al. Whole-genome sequencing and analysis of the Malaysian cynomolgus macaque (*Macaca fascicularis*) genome. Genome Biol. 2012;13: R58. doi:10.1186/gb-2012-13-7-r58

4. Kuhlwilm M, Han S, Sousa VC, Excoffier L, Marques-Bonet T. Ancient admixture from an extinct ape lineage into bonobos. Nat Ecol Evol. 2019;3: 957–965. doi:10.1038/s41559-019-0881-7

5. Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, et al. Comparative and demographic analysis of orang-utan genomes. Nature. 2011;469: 529–533. doi:10.1038/nature09687

6. de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science. 2016;354: 477–481. doi:10.1126/science.aag2602

7. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature. 2013;499: 471–475. doi:10.1038/nature12228

8. Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, et al. The bonobo genome compared with the chimpanzee and human genomes. Nature. 2012;486: 527–531. doi:10.1038/nature11128

9. Rogers J, Raveendran M, Harris RA, Mailund T, Leppälä K, Athanasiadis G, et al. The comparative genomics and complex population history of *Papio* baboons. Sci Adv. 2019;5: eaau6947. doi:10.1126/sciadv.aau6947

10. Svardal H, Jasinska AJ, Apetrei C, Coppola G, Huang Y, Schmitt CA, et al. Ancient hybridization and strong adaptation to viruses across African vervet monkey populations. Nat Genet. 2017;49: 1705–1713. doi:10.1038/ng.3980

11. Zhou X, Wang B, Pan Q, Zhang J, Kumar S, Sun X, et al. Whole-genome sequencing of the snub-nosed monkey provides insights into folivory and evolutionary history. Nat Genet. 2014;46: 1303–1310. doi:10.1038/ng.3137

12. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, et al. A molecular phylogeny of living primates. PLOS Genet. 2011;7: e1001342. doi:10.1371/journal.pgen.1001342

13. Springer MS, Meredith RW, Gatesy J, Emerling CA, Park J, Rabosky DL, et al. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLOS One. 2012;7: e49521. doi:10.1371/journal.pone.0049521

14. Wang X, Lim BK, Ting N, Hu J, Liang Y, Roos C, et al. Reconstructing the phylogeny of new world monkeys (Platyrrhini): evidence from multiple non-coding loci. Curr Zool. 2019;65: 579–588. doi:10.1093/cz/zoy072

15. Silvestro D, Tejedor MF, Serrano-Serrano ML, Loiseau O, Rossier V, Rolland J, et al. Early arrival and climatically-linked geographic expansion of New World monkeys from tiny African ancestors. Syst Biol. 2018;68: 78–92. doi:10.1093/sysbio/syy046

16. Jameson Kiesling NM, Yi SV, Xu K, Gianluca Sperone F, Wildman DE. The tempo and mode of New World monkey evolution and biogeography in the context of phylogenomic analysis. Mol Phylogenet Evol. 2015;82 Pt B: 386–399. doi:10.1016/j.ympev.2014.03.027

17. Perez SI, Tejedor MF, Novo NM, Aristide L. Divergence times and the evolutionary radiation of New World monkeys (Platyrrhini, Primates): an analysis of fossil and molecular data. PLOS One. 2013;8: e68029. doi:10.1371/journal.pone.0068029

18. Schrago CG, Seuánez HN. Large ancestral effective population size explains the difficult phylogenetic placement of owl monkeys. Am J Primatol. 2019;55: e22955. doi:10.1002/ajp.22955

19. Degnan JH, Rosenberg NA. Discordance of species trees with their most likely gene trees. PLOS Genet. 2006;2: e68. doi:10.1371/journal.pgen.0020068

20. Huang H, Knowles LL. What is the danger of the anomaly zone for empirical phylogenetics? Syst Biol. 2009;58: 527–536. doi:10.1093/sysbio/syp047

21. Mendes FK, Hahn MW. Why Concatenation Fails Near the Anomaly Zone. Syst Biol. 2018;67: 158–169. doi:10.1093/sysbio/syx063

22. Mallet J. Hybridization as an invasion of the genome. Trends Ecol Evol. 2005;20: 229–237. doi:10.1016/j.tree.2005.02.010

23. Mallet J, Besansky N, Hahn MW. How reticulated are species? BioEssays. 2016;38: 140–149. doi:10.1002/bies.201500149

24. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328: 710–722. doi:10.1126/science.1188021

25. Lima MGM, Silva-Júnior J de SE, Černý D, Buckner JC, Aleixo A, Chang J, et al. A phylogenomic perspective on the robust capuchin monkey (*Sapajus*) radiation: First evidence for extensive population admixture across South America. Mol Phylogenet Evol. 2018;124: 137–150. doi:10.1016/j.ympev.2018.02.023

26. de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science. 2016;354: 477–481. doi:10.1126/science.aag2602

27. Wall JD, Schlebusch SA, Alberts SC, Cox LA, Snyder-Mackler N, Nevonen KA, et al. Genomewide ancestry and divergence patterns from low-coverage sequencing data reveal a complex history of admixture in wild baboons. Mol Ecol. 2016;25: 3469–3483. doi:10.1111/mec.13684

28. Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014;512: 194–197. doi:10.1038/nature13408

29. Racimo F, Sankararaman S, Nielsen R, Huerta-Sanchez E. Evidence for archaic adaptive introgression in humans. Nat Rev Genet. 2015;16: 359–371. doi:10.1038/nrg3936

30. Racimo F, Gokhman D, Fumagalli M, Ko A, Hansen T, Moltke I, et al. Archaic adaptive introgression in *TBX15/WARS2*. Mol Biol Evol. 2017;34: 509–524. doi:10.1093/molbev/msw283

31. Chatterjee HJ, Ho SYW, Barnes I, Groves C. Estimating the phylogeny and divergence times of primates using a supermatrix approach. BMC Evol Biol. 2009;9: 259. doi:10.1186/1471-2148-9-259

32. Herrera JP, Dávalos LM. Phylogeny and divergence times of lemurs inferred with recent and ancient fossils in the tree. Syst Biol. 2016;65: 772–791. doi:10.1093/sysbio/syw035

33. Kistler L, Ratan A, Godfrey LR, Crowley BE, Hughes CE, Lei R, et al. Comparative and population mitogenomic analyses of Madagascar's extinct, giant 'subfossil' lemurs. J Hum Evol. 2015;79: 45–54. doi:10.1016/j.jhevol.2014.06.016

34. Tung J, Barreiro LB. The contribution of admixture to primate evolution. Curr Opin Genet Dev. 2017;47: 61–68. doi:10.1016/j.gde.2017.08.010

35. Stevens NJ, Seiffert ER, O'Connor PM, Roberts EM, Schmitz MD, Krause C, et al. Palaeontological evidence for an Oligocene divergence between Old World monkeys and apes. Nature. 2013;497: 611–614. doi:10.1038/nature12161

36. Huson DH, Klöpper T, Lockhart PJ, Steel MA. Reconstruction of reticulate networks from gene trees. Proceedings of RECOMB 2005: The 9th Annual International Conference Research in Computational Molecular Biology. Berlin: Springer; 2005. pp. 233–249. doi:10.1007/11415770_18

37. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2018;35: 543–548. doi:10.1093/molbev/msx319

38. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020. doi:10.1093/molbev/msaa015

903 39. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species
904 tree reconstruction from partially resolved gene trees. BMC Bioinformatics.
905 2018;19: 153. doi:10.1186/s12859-018-2129-y

906 40. Sayyari E, Mirarab S. Fast coalescent-based computation of local branch support
907 from quartet frequencies. Mol Biol Evol. 2016;33: 1654–1668.
908 doi:10.1093/molbev/msw079

909 41. Adkins RM, Honeycutt RL. Molecular phylogeny of the superorder Archonta. Proc
910 Natl Acad Sci. 1991;88: 10317–10321. doi:10.1073/pnas.88.22.10317

911 42. Arnason U, Adegoke JA, Bodin K, Born EW, Esa YB, Gullberg A, et al.
912 Mammalian mitogenomic relationships and the root of the eutherian tree. Proc
913 Natl Acad Sci. 2002;99: 8151–8156. doi:10.1073/pnas.102164299

914 43. Bloch JI, Boyer DM. Grasping primate origins. Science. 2002;298: 1606–1610.
915 doi:10.1126/science.1078249

916 44. Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, et al. Parallel
917 adaptive radiations in two major clades of placental mammals. Nature. 2001;409:
918 610–614. doi:10.1038/35054544

919 45. Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, et al.
920 Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal
921 diversification. Science. 2011;334: 521–524. doi:10.1126/science.1211028

922 46. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. Molecular
923 phylogenetics and the origins of placental mammals. Nature. 2001;409: 614–618.
924 doi:10.1038/35054550

925 47. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, et al.
926 Resolution of the early placental mammal radiation using Bayesian phylogenetics.
927 Science. 2001;294: 2348–2351. doi:10.1126/science.1067179

928 48. Novacek MJ. Mammalian phylogeny: shaking the tree. Nature. 1992;356: 121–
929 125. doi:10.1038/356121a0

930 49. O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, et al.
931 The placental mammal ancestor and the post-K-Pg radiation of placentals.
932 Science. 2013;339: 662–667. doi:10.1126/science.1229237

933 50. Poux C, Douzery EJP. Primate phylogeny, evolutionary rate variations, and
934 divergence times: a contribution from the nuclear gene *IRBP*. Am J Phys
935 Anthropol. 2004;124: 1–16. doi:10.1002/ajpa.10322

936 51. Janečka JE, Miller W, Pringle TH, Wiens F, Zitzmann A, Helgen KM, et al.
937 Molecular and genomic data identify the closest living relative of primates.
938 Science. 2007;318: 792–794. doi:10.1126/science.1147555

939 52. Mason VC, Li G, Minx P, Schmitz J, Churakov G, Doronina L, et al. Genomic
940 analysis reveals hidden biodiversity within colugos, the sister group to primates.
941 Sci Adv. 2016;2: e1600633. doi:10.1126/sciadv.1600633

942 53. Schmitz J, Ohme M, Suryobroto B, Zischler H. The colugo (*Cynocephalus*
943 *variegatus*, Dermoptera): the primates' gliding sister? Mol Biol Evol. 2002;19:
944 2308–2312. doi:10.1093/oxfordjournals.molbev.a004054

945 54. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, et al. Whole-genome
946 analyses resolve early branches in the tree of life of modern birds. Science.
947 2014;346: 1320–1331. doi:10.1126/science.1253451

948 55. Pease JB, Haak DC, Hahn MW, Moyle LC. Phylogenomics reveals three sources
949 of adaptive variation during a rapid radiation. PLOS Biol. 2016;14: e1002379.
950 doi:10.1371/journal.pbio.1002379

951 56. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong
952 phylogenetic signals. Nature. 2013;497: 327–331. doi:10.1038/nature12130

953 57. Minh BQ, Hahn MW, Lanfear R. New methods to calculate concordance factors
954 for phylogenomic datasets. bioRxiv. 2018; 487801. doi:10.1101/487801

955 58. Yoder AD. The phylogenetic position of genus *Tarsius*: whose side are you on? In:
956 Wright PC, Simons EL, Gursky S, editors. Tarsiers Past, Present, and Future.
957 Rutgers University Press; 2003.

958 59. Gregory WK. On the classification and phylogeny of the Lemuroidea. Bull Geol
959 Soc Am. 1915; 426–446.

960 60. Pocock RI. On the external characters of the lemurs and of *Tarsius*. Proc Zool Soc
961 Lond. 1918;88: 19–53. doi:10.1111/j.1096-3642.1918.tb02076.x

962 61. Hartig G, Churakov G, Warren WC, Brosius J, Makalowski W, Schmitz J.
963 Retrophylogenomics place tarsiers on the evolutionary branch of anthropoids. Sci
964 Rep. 2013;3: 1756. doi:10.1038/srep01756

965 62. Jameson NM, Hou Z-C, Sterner KN, Weckle A, Goodman M, Steiper ME, et al.
966 Genomic data reject the hypothesis of a prosimian primate clade. J Hum Evol.
967 2011;61: 295–305. doi:10.1016/j.jhevol.2011.04.004

63. Hayasaka K, Gojobori T, Horai S. Molecular phylogeny and evolution of primate mitochondrial DNA. Mol Biol Evol. 1988;5: 626–644. doi:10.1093/oxfordjournals.molbev.a040524

64. Jaworski CJ. A reassessment of mammalian alpha A-crystallin sequences using DNA sequencing: implications for anthropoid affinities of tarsier. J Mol Evol. 1995;41: 901–908. doi:10.1007/bf00173170

65. Whitfield JB, Lockhart PJ. Deciphering ancient rapid radiations. Trends Ecol Evol. 2007;22: 258–265. doi:10.1016/j.tree.2007.01.012

66. Bond M, Tejedor MF, Campbell KE, Chornogubsky L, Novo N, Goin F. Eocene primates of South America and the African origins of New World monkeys. Nature. 2015;520: 538–541. doi:10.1038/nature14120

67. Kubatko LS, Degnan JH. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol. 2007;56: 17–24. doi:10.1080/10635150601146041

68. Roch S, Steel M. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor Popul Biol. 2015;100C: 56–62. doi:10.1016/j.tpb.2014.12.005

69. Warnow T. Concatenation analyses in the presence of incomplete lineage sorting. PLOS Curr Tree Life. 2015;7. doi:10.1371/currents.tol.8d41ac0f13d1abedf4c4a59f5d17b1f7

70. Bryant D, Hahn MW. The concatenation question. In: Scornavacca C, Delsuc F, Galtier N, editors. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book; 2020. p. 3.4:1–3.4:23. Available: https://hal.archives-ouvertes.fr/hal-02535651

71. Sayyari E, Mirarab S. Testing for polytomies in phylogenetic species trees using quartet frequencies. Genes. 2018;9: 132. doi:10.3390/genes9030132

72. Liu L, Edwards SV. Phylogenetic analysis in the anomaly zone. Syst Biol. 2009;58: 452–460. doi:10.1093/sysbio/syp034

73. Swofford DL. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer Assoc Sunderland Mass. 2002;Version 4.

74. Mendes FK, Livera AP, Hahn MW. The perils of intralocus recombination for inferences of molecular convergence. Philos Trans R Soc Lond B Biol Sci. 2019;374: 20180244. doi:10.1098/rstb.2018.0244

75. Springer MS, Gatesy J. The gene tree delusion. Mol Phylogenet Evol. 2016;94: 1–33. doi:10.1016/j.ympev.2015.07.018

76. Hoang DT, Vinh LS, Flouri T, Stamatakis A, von Haeseler A, Minh BQ. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. BMC Evol Biol. 2018;18: 11–11. doi:10.1186/s12862-018-1131-3

77. He C, Liang D, Zhang P. Asymmetric distribution of gene trees can arise under purifying selection If differences in population size exist. Mol Biol Evol. 2020;37: 881–892. doi:10.1093/molbev/msz232

78. Golding GB. The effect of purifying selection on genealogies. In: Donnelly P, Tavaré S, editors. Progress in population genetics and human evolution. New York: Springer Verlag; 1997. pp. 271–285.

79. Przeworski M, Charlesworth B, Wall JD. Genealogies and weak purifying selection. Mol Biol Evol. 1999;16: 246–252. doi:10.1093/oxfordjournals.molbev.a026106

80. Slade PF. Most recent common ancestor probability distributions in gene genealogies under selection. Theor Popul Biol. 2000;58: 291–305. doi:10.1006/tpbi.2000.1488

81. Williamson S, Orive ME. The genealogy of a sequence subject to purifying selection at multiple sites. Mol Biol Evol. 2002;19: 1376–1384. doi:10.1093/oxfordjournals.molbev.a004199

82. Haller BC, Messer PW. SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. Mol Biol Evol. 2019;36: 632–637. doi:10.1093/molbev/msy228

83. Mendes FK, Hahn MW. Gene tree discordance causes apparent substitution rate variation. Syst Biol. 2016;65: 711–721. doi:10.1093/sysbio/syw018

84. Brunet M, Guy F, Pilbeam D, Mackaye HT, Likius A, Ahounta D, et al. A new hominid from the Upper Miocene of Chad, Central Africa. Nature. 2002;418: 145–151. doi:10.1038/nature00879

85. Sigé B, Jaeger J-J, Sudre J, Vianey-Liaud M. *Altiatlasius koulchii* n. gen. et sp., primate omomyidé du Paléocène supérieur du Maroc, et les origines des euprimates. Palaeontogr Abt A. 1990; 31–56.

86. Lartillot N, Lepage T, Blanquart S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 2009;25: 2286–2288. doi:10.1093/bioinformatics/btp368

87. Wilkinson RD, Steiper ME, Soligo C, Martin RD, Yang Z, Tavaré S. Dating primate divergences through an integrated analysis of palaeontological and molecular data. Syst Biol. 2011;60: 16–31. doi:10.1093/sysbio/syq054

88. Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, et al. Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol. 1998;9: 585–598. doi:10.1006/mpev.1998.0495

89. Yoder AD, Yang Z. Divergence dates for Malagasy lemurs estimated from multiple gene loci: geological and evolutionary context. Mol Ecol. 2004;13: 757–773. doi:10.1046/j.1365-294X.2004.02106.x

90. Benton MJ, Donoghue PCJ, Asher RJ, Friedman M, Near TJ, Vinther J. Constraints on the timescale of animal evolutionary history. Palaeontol Electron. 2015;18.1.1FC: 1–106. doi:10.26879/424

91. Edwards SV, Beerli P. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution. 2000;54: 1839–1854. doi:10.1111/j.0014-3820.2000.tb01231.x

92. Rogers J. Levels of the genealogical hierarchy and the problem of hominoid phylogeny. Am J Phys Anthropol. 1994;94: 81–88. doi:10.1002/ajpa.1330940107

93. Carbone L, Harris RA, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, et al. Gibbon genome and the fast karyotype evolution of small apes. Nature. 2014;513: 195–201. doi:10.1038/nature13679

94. Veeramah KR, Woerner AE, Johnstone L, Gut I, Gut M, Marques-Bonet T, et al. Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach. Genetics. 2015;200: 295–308. doi:10.1534/genetics.115.174425

95. Hamada Y, San AM, Malaivijitnond S. Assessment of the hybridization between rhesus (*Macaca mulatta*) and long-tailed macaques (*M. fascicularis*) based on morphological characters. Am J Phys Anthropol. 2016;159: 189–198. doi:10.1002/ajpa.22862

96. Osada N, Uno Y, Mineta K, Kameoka Y, Takahashi I, Terao K. Ancient genome-wide admixture extends beyond the current hybrid zone between *Macaca fascicularis* and *M. mulatta*. Mol Ecol. 2010;19: 2884–2895. doi:10.1111/j.1365-294X.2010.04687.x

97. Hahn MW. Molecular population genetics. First Edition. Oxford, New York: Oxford University Press; 2018.

98. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192: 1065–1093. doi:10.1534/genetics.112.145037

99. Fan Z, Zhao G, Li P, Osada N, Xing J, Yi Y, et al. Whole-genome sequencing of Tibetan macaque (*Macaca thibetana*) provides new insight into the macaque evolutionary history. Mol Biol Evol. 2014;31: 1475–1489. doi:10.1093/molbev/msu104

100. Yan G, Zhang G, Fang X, Zhang Y, Li C, Ling F, et al. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol. 2011;29: 1019–1023. doi:10.1038/nbt.1992

101. Koufos GD. Potential hominoid ancestors for Hominidae. In: Henke W, Tattersall I, editors. Handbook of Paleoanthropology. Berlin, Heidelberg: Springer; 2007. pp. 1761–1790. doi:10.1007/978-3-642-39979-4_44

102. Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. Genetic evidence for complex speciation of humans and chimpanzees. Nature. 2006;441: 1103–1108. doi:10.1038/nature04789

103. Fleagle JG. Apes and humans. Primate adaptation and evolution. Elsevier; 2013. pp. 151–168. doi:10.1016/B978-0-12-378632-6.00007-0

104. Than C, Ruths D, Nakhleh L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics. 2008;9: 322. doi:10.1186/1471-2105-9-322

105. Solís-Lemus C, Ané C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLOS Genet. 2016;12: e1005896. doi:10.1371/journal.pgen.1005896

106. Roos C, Kothe M, Alba DM, Delson E, Zinner D. The radiation of macaques out of Africa: Evidence from mitogenome divergence times and the fossil record. J Hum Evol. 2019;133: 114–132. doi:10.1016/j.jhevol.2019.05.017

107. Belmaker M. The presence of a large cercopithecine (cf. *Theropithecus* sp.) in the 'Ubeidiya formation (Early Pleistocene, Israel). J Hum Evol. 2010;58: 79–89. doi:10.1016/j.jhevol.2009.08.004

108. Hughes JK, Elton S, O'Regan HJ. Theropithecus and "Out of Africa" dispersal in the Plio-Pleistocene. J Hum Evol. 2008;54: 43–77. doi:10.1016/j.jhevol.2007.06.004

109. Larrasoaña JC, Roberts AP, Rohling EJ, Winklhofer M, Wehausen R. Three million years of monsoon variability over the northern Sahara. Clim Dyn. 2003;21: 689–698. doi:10.1007/s00382-003-0355-z

110. Larrasoaña JC, Roberts AP, Rohling EJ. Dynamics of green Sahara periods and their role in hominin evolution. PLOS One. 2013;8: e76514. doi:10.1371/journal.pone.0076514

111. Vaks A, Woodhead J, Bar-Matthews M, Ayalon A, Cliff RA, Zilberman T, et al. Pliocene–Pleistocene climate of the northern margin of Saharan–Arabian Desert recorded in speleothems from the Negev Desert, Israel. Earth Planet Sci Lett. 2013;368: 88–100. doi:10.1016/j.epsl.2013.02.027

112. Coulthard TJ, Ramirez JA, Barton N, Rogerson M, Brücher T. Were rivers flowing across the Sahara during the last interglacial? Implications for human migration through Africa. PLOS One. 2013;8: e74834. doi:10.1371/journal.pone.0074834

113. Sahnouni M, Parés JM, Duval M, Cáceres I, Harichane Z, van der Made J, et al. 1.9-million- and 2.4-million-year-old artifacts and stone tool-cutmarked bones from Ain Boucherit, Algeria. Science. 2018;362: 1297–1301. doi:10.1126/science.aau0008

114. deMenocal PB. African climate change and faunal evolution during the Pliocene–Pleistocene. Earth Planet Sci Lett. 2004;220: 3–24. doi:10.1016/S0012-821X(04)00003-2

115. Slatkin M, Pollack JL. Subdivision in an ancestral species creates asymmetry in gene trees. Mol Biol Evol. 2008;25: 2241–2246. doi:10.1093/molbev/msn172

116. Kuritzin A, Kischka T, Schmitz J, Churakov G. Incomplete Lineage Sorting and Hybridization Statistics for Large-Scale Retroposon Insertion Data. PLOS Comput Biol. 2016;12: e1004812. doi:10.1371/journal.pcbi.1004812

117. Springer MS, Molloy EK, Sloan DB, Simmons MP, Gatesy J. ILS-Aware Analysis of Low-Homoplasy Retroelement Insertions: Inference of Species Trees and Introgression Using Quartets. J Hered. 2020;111: 147–168. doi:10.1093/jhered/esz076

118. Gernhard T. New analytic results for speciation times in neutral models. Bull Math Biol. 2008;70: 1082–1097. doi:10.1007/s11538-007-9291-0

119. Gligor M, Ganzhorn JU, Rakotondravony D, Ramilijaona OR, Razafimahatratra E, Zischler H, et al. Hybridization between mouse lemurs in an ecological transition zone in southern Madagascar. Mol Ecol. 2009;18: 520–533. doi:10.1111/j.1365-294X.2008.04040.x

120. Pastorini J, Zaramody A, Curtis DJ, Nievergelt CM, Mundy NI. Genetic analysis of hybridization and introgression between wild mongoose and brown lemurs. BMC Evol Biol. 2009;9: 32. doi:10.1186/1471-2148-9-32

121. Williams RC, Blanco MB, Poelstra JW, Hunnicutt KE, Comeault AA, Yoder AD. Conservation genomic analysis reveals ancient introgression and declining levels of genetic diversity in Madagascar's hibernating dwarf lemurs. Heredity. 2020;124: 236–251. doi:10.1038/s41437-019-0260-9

122. Wyner YM, Johnson SE, Stumpf RM, Desalle R. Genetic assessment of a white-collared×red-fronted lemur hybrid zone at Andringitra, Madagascar. Am J Primatol. 2002;57: 51–66. doi:10.1002/ajp.10033

123. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci. 2011;108: 1513–1518. doi:10.1073/pnas.1017351108

124. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS ONE. 2012;7: e47768. doi:10.1371/journal.pone.0047768

125. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

126. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-2105-10-421

127. van Dongen S. Graph clustering by flow simulation. PhD thesis, University of Utrecht. 2000. Available: http://www.library.uu.nl/digiarchief/dip/diss/1895620/full.pdf

128. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42: D581–D591. doi:10.1093/nar/gkt1099

129. Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res. 2015;43: W7–14. doi:10.1093/nar/gkv318

130. Katoh K, Standley DM. Mafft multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010

131. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25: 1972–1973. doi:10.1093/bioinformatics/btp348

132. Chernomor O, von Haeseler A, Minh BQ. Terrace aware data structure for phylogenomic inference from supermatrices. Syst Biol. 2016;65: 997–1008. doi:10.1093/sysbio/syw037

133. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018;35: 518–522. doi:10.1093/molbev/msx281

134. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14: 587–589. doi:10.1038/nmeth.4285

135. Vanderpool D. Data from: Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. In: Dryad Digital Repository. [Internet]. 2020. Available: https://doi.org/10.5061/dryad.rfj6q577d.

136. Eaton DAR, Ree RH. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). Syst Biol. 2013;62: 689–706. doi:10.1093/sysbio/syt032

137. Dunn OJ. Confidence intervals for the means of dependent, normally distributed variables. J Am Stat Assoc. 1959;54: 613–621. doi:10.2307/2282541

138. Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc. 1967;62: 626–633. doi:10.2307/2283989

139. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004;21: 1095–1109. doi:10.1093/molbev/msh112

140. Thorne JL, Kishino H, Painter IS. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol. 1998;15: 1647–1657.

Supplementary Information

SI Table 1. Genomes analyzed in this study with the original NCBI release date, the publication for the reference used, and the accession number for the assembly. When possible the most recent version for each genome was used.

SI Table 2. All published genomes used in this study, including links to the assemblies and NCBI BioProjects. Annotation information is included for each genome at the time of download.

SI Table 3. Orthogroup, protein name, human chromosome number, and coordinates for the single-copy human orthologs used in the 1,730 gene analysis. Alignment files are named by orthogroup, allowing the use of this table to identify the protein in each alignment.

SI Table 4. Gaps/Ambiguities by species, and as a percentage of total alignment length. * denotes species sequenced this study.

SI Table 5. Lengths for each 40-locus concatenated alignment used in the molecular dating analyses. Each dataset was analyzed twice until node age estimates converged (15-25k steps) using a log-normal auto-correlated model [140]. Datasets are available via Data Dryad at https://doi.org/10.5061/dryad.rfj6q577d.

SI Table 6. Fossil calibrations employed in this study. Node numbering corresponds to the numbering in Figure 3. Median underflow/overflow for each calibration was calculated from 20 independent runs performed on 10 datasets (2 runs per dataset).

SI Table 7. Mean node age for 20 independent PhyloBayes dating runs. Node numbers correspond to the numbering in Figure 3. The 95% HPD intervals were calculated by averaging the minimum and maximum of the 95% HPD interval for each dating run.

SI Table 8. Quartets used to test for significant Δ values for internal branches of the primate tree. Branches tested correspond to the labeled branches in Figure 3. After correcting for multiple comparisons (Dunn-Šidák, $P = 0.00301$), three internal branches and 8 quartets were found to have significant Δ values, indicating a likely introgression event.

SI Fig 1. Concordance Factors for the species tree in Figure 1 calculated using maximum likelihood gene trees and site patterns from the 200 longest single-copy loci alignments used in the 1,730-gene analysis. In general, gCFs increase while the sCFs remain the same, indicating gene tree error is a likely source of some discordance.

SI Fig 2. Concordance factors calculated using 150 randomly chosen single-copy orthologs, with pika (*Ochotona princeps*) included as an additional outgroup to mouse. (A) gCFs and sCFs for these 150 genes when pika is included. (B) gCFs and sCFs for these same genes when pika is not included. We observe slightly higher gCFs near the

base of the tree with pika excluded (red boxes). Note that these species trees use unit-length branch lengths for readability of branch labels.

SI Fig 3. Gene and site concordance factors plotted as a function of node depth (in millions of years). No correlation was found between gCFs and node depth, while a slightly negative correlation was found between sCFs and node depth. This relationship indicates that homoplasy may act to slightly reduce sCFs deeper in the tree.

SI Fig 4. Forward simulations using SLiM3 with the most extreme parameters used by He *et al.* (2020): population size combination "F" with s=-7.5x10^{-6} and $\Delta\tau$=2000. Our results show no significant difference in the distribution of gene tree topologies in the presence of negative selection (A). This result holds for simulations in which we increased the per locus mutation rate by two orders of magnitude (B).

SI Fig 5. Present day species distributions for four African Papionini (*Papio, Theropithecus, Mandrillus*, and *Cercocebus*) and three Asian *Macaca* species included in the introgression analysis. The ancestral *Macaca* distribution (grey shading) is inferred from *Macaca* fossil localities in Africa and Europe as reviewed in Roos et al. (2019). The ancestral *Macaca* distribution likely represents only a fraction of the species range from the late Miocene to the late Pleistocene in Africa and Europe. The contemporary distribution of the African *Macaca sylvanus* (bright green) is included for reference; the current distribution of *Macaca nemestrina* is completely contained within that of *Macaca fascicularis*. Fossil localities for *Theropithecus* species hypothesized to overlap contemporaneously with various ancestral *Macaca* are included. Citations for

spatial data of extant species: *Macaca nemestrina* (Richardson et al., 2008), *Macaca fascicularis* (Ong & Richardson, 2008), *Macaca sylvanus* (Butynski et al., 2008), *Macaca mulatta* (Timmins et al., 2008), *Theropithecus gelada* (Gippoliti et al., 2019), *Papio anubis* (Kingdon et al., 2008), *Cercocebus atys* (Oates et al., 2016), and *Mandrillus leucophaeus* (Oates & Butynski, 2008). Base map was obtained from the public domain map database Natural Earth (http://www.naturalearthdata.com/downloads/).