THE *10KTREES* WEBSITE: A NEW ONLINE RESOURCE FOR PRIMATE, CARNIVORA, CETARTIODACTYLA AND PERISSODACTYLA PHYLOGENY

Christian Arnold, Luke Matthews, and Charles Nunn

Department of Human Evolutionary Biology

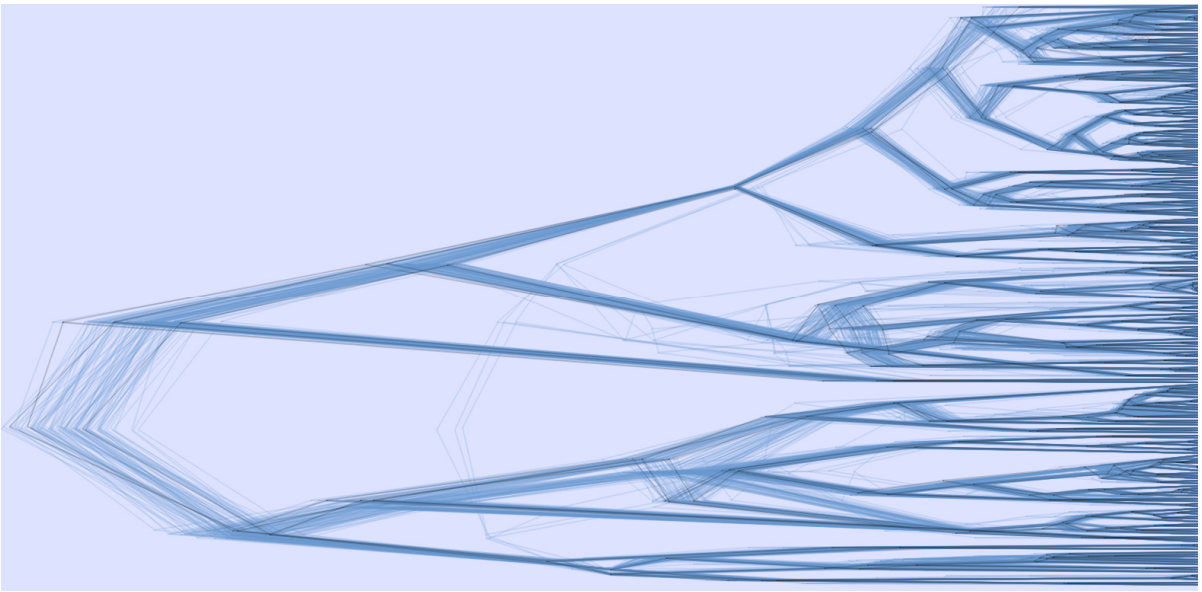
Harvard University

11 Divinity Avenue

Cambridge, MA 02138

http://www.fas.harvard.edu/~primecol

http://10kTrees.fas.harvard.edu



Documentation

Last updated: June 2012

This document provides additional information for the *10kTrees Project*.

If you have questions or comments, feel free to contact me, Christian Arnold (carnold@fas.harvard.edu). I will be happy to answer any questions related to this project as well as questions related to the web-implementation.

<u>If you use trees from this website, please cite the following reference:</u>

Arnold, C., L. J. Matthews, and C. L. Nunn. 2010. The *10kTrees Website*: A New Online Resource for Primate Phylogeny. *Evolutionary Anthropology* **19**:114-118.

Table of Contents

1. Project Description

The *10kTrees Website* is a new web resource for conducting comparative studies of primates, carnivorans, odd-toed ungulates, and even-toed ungulates and cetaceans. The comparative method plays a central role in efforts to uncover the adaptive basis for primate behavior, morphology and life history traits and has undergone a revolution in the past 20 years. With a phylogeny for a group of organisms, it is now possible to address fundamental questions about correlated trait evolution, the factors that drive diversification of lineages, and the pattern and process of evolutionary change.

The true history (i.e. tree topology and timing of speciation events) is never known with certainty, however, and relationships should be continually reassessed as new data become available. This last fact recommends against the continued use of older phylogenies, as better data are now available. Furthermore, when conducting a comparative test, it is desirable to incorporate the current level of uncertainty for specific nodes and branch lengths. Different trees can produce different results during comparative analysis, which argues against conditioning comparative tests on a single hypothesis of evolutionary relationships when that hypothesis is legitimately uncertain (Lutzoni et al. 2001). A major development in phylogenetics research involves the use of statistical methods that control for phylogenetic uncertainty (Huelsenbeck et al. 2001; Lutzoni et al. 2001; Pagel and Lutzoni 2002). These *Bayesian* methods provide a way to sample a set of trees in proportion to their posterior probabilities by using Markov chain Monte Carlo (MCMC). This allows researchers to run analyses on an entire set of trees rather than using a single tree; thus, results are no longer conditioned on a single tree being correct.

Using the *10kTrees Website*, users can download up to 10,000 phylogenies for primates, carnivorans, odd-toed and even-toed ungulates, and cetaceans. These phylogenies (with branch lengths) are sampled from a Bayesian phylogenetic analysis of genetic data. The website provides a variety of options, which are further described in section 6 of this document. Moreover, we designed the website so that it can be easily updated as new versions of the phylogeny become available. We also expect that the website itself will evolve to provide more tools for primate comparative biology (see section 8 and the *News* section of the website).

The overarching goal of 10kTrees is to produce a set of phylogenetic trees that is appropriate for comparative research and reflects current uncertainty in the understanding of

primate evolutionary relationships. We regularly update the dataset to accommodate the ever-increasing amount of available sequence data as well as tree inference methods. Thus, this project evolves as new resources become available to expand phylogenetic inference to more species and strengthen our understanding of phylogenetic relationships more generally.

2. Methodological Details for the Primates Part of the Website

In what follows, we provide details on each of the versions, beginning with the most recent version.

2.1. Version 3

Version 3 of the Primates part is our biggest dataset so far. The trees include 301 primate species and are based on more genes than version 2. Importantly, all the species that were missing in version 2 compared to version 1 are now included in version 3. For the first time, Version 3 includes two extinct species with sequenced DNA (*Homo sapiens neanderthalensis* and *Archaeolemur majori*).

2.1.1. Data Collection

For the third version of the dataset, we collected data for eleven mitochondrial and six autosomal genes that were generally available in GenBank across 301 primate species and the outgroup species *Galeopterus variegates* (Sunda flying lemur). We used the Phylota browser (release 1.5, Sanderson et al. 2008) for data collection and to identify the genes for which sufficient data were available and automatically downloaded all available sequences for each of the species in this dataset using the bioinformatics pipeline FAST (Arnold 2012, in prep.). We strictly excluded all sequences that were annotated as pseudogenes or hypothetical or working draft etc., similar to the particular gene of interest, or ambiguous (in annotation) in general. If multiple sequences from a particular gene were available for the same species, we selected the longest sequence (while controlling for ambiguous codes, such as *N,* which can stand for any of the four bases). If the whole mitochondrial genome for a particular species was available, we always extracted and selected the sequences for the genes of interest from the mitochondrion, rather than taking the sequences that were available on GenBank (if any were available at all). This substantially improved the quality of the sequences and the subsequent alignments.

Table 1. Summary of the data collected for Version 3.

Gene Name (abbr.)	**Full name**	**Genomic position**	**Number of species for which seq. are available**
12S rRNA	12S ribosomal rRNA	MIT	179
16S rRNA	16S ribosomal rRNA	MIT	140
CCR5	C-C chemokine receptor type 5	CHR	76
COX1	Cytochrome c oxidase subunit I	MIT	119
COX2	Cytochrome c oxidase subunit II	MIT	157
COX3	Cytochrome c oxidase subunit III	MIT	63
CYTB	Cytochrome B	MIT	228
IRBP	Interphotoreceptor retinoid-binding protein	CHR	51
MC1R	Melanocortin 1 receptor	CHR	73
ND1	NADH dehydrogenase subunit 1	MIT	66
ND3	NADH dehydrogenase subunit 3	MIT	141
ND4	NADH dehydrogenase subunit 4	MIT	168
ND4L	NADH dehydrogenase subunit 4L	MIT	152
ND5	NADH dehydrogenase subunit 5	MIT	74
PRP	Major prion protein (encoded by the *PRNP* gene)	CHR	46
SRY	Sex-determining Region Y	Y-CHR	99
TSPY	Testis-specific Y-encoded protein 1 (encoded by the *TSPY1* gene)	Y-CHR	62

Notes: MIT stands for mitochondrial, CHR for chromosome in general, while Y-CHR stands for Y-chromosome.

<u>The following list summarizes the data collection for Version 2:</u>

Number of species: 302 (only 301 are listed on the website, as we pruned the outgroup species from the trees after the tree inference)

Total number of available sequences: 1894 (out of 301*17 = 5117 total)

Percentage of missing data: 63.0% (69.0% if missing data within genes are also counted)

2.1.2. Multiple sequence alignments

For creating multiple sequence alignments (MSA) for each of the genes, we used Muscle 3.7 (default parameters except the following: -cluster1 neighborjoining -maxtrees 5 -noanchors -cluster2 neighborjoining -distance1 kmer20_4). As it has been repeatedly demonstrated that alignment quality may have a substantial impact on the inferred tree (Kjer 1995; Morrison and Ellis 1997; Ogden and Rosenberg 2006; Smythe et al. 2006; Talavera and Castresana 2007), we eliminated poorly aligned positions and divergent regions of the alignment using the program Gblocks (Castresana 2002) with the settings -b5=h, -t=d, and -b2=0.6 * *number of sequences*. These positions may not be homologous or may have been saturated by multiple substitutions. Gblocks selects blocks in a similar way as it is usually done manually by hand. However, it follows a reproducible set of conditions, making the phylogenetic analyses of large datasets reliable, feasible, and also more accurate, especially because sequences in GenBank may be of poor quality. The multiple sequence alignments for each gene can be downloaded on the website. For some of the genes (e.g., 12S rRNA or 16S rRNA), we manually improved the quality of the alignment or eliminated regions of high divergence and / or a large number of gaps before Gblocks was used.

2.1.3. Phylogenetic constraints

In Version 3, due to the increased number of available genes and sequences (as compared to Version 1), we only constrained four major nodes (see the file "Phylogenetic constraints for the tree inference" on the website). See section 2.3.3 for an explanation why we defined constraints in our analysis.

2.1.4. Tree inference

For the tree inference, we used the program MrBayes 3.2 (Ronquist and Huelsenbeck 2003). We used the species *Galeopterus variegatus* (Sunda flying lemur) as the outgroup, as it has been shown that this species is the closest living relative to the order Primates (Janecka et al. 2007). We ran a Bayesian analysis with three runs and 8 chains in each run. We used different substitution models (general time reversible (GTR) model (Rodriguez et al. 1990) and the HKY model, with a proportion of invariable sites and a gamma-shaped rate variation across sites,

Table 2) for each of the genes in a partitioned dataset (while all mitochondrial genes were in one partition), which were identified in the program JModelTest (Posada 2008) and Phyml (Guindon and Gascuel 2003). If the best-suited substitution model determined by JModelTest was not available in MrBayes, we selected the model with the best AIC score among the models that are implemented in MrBayes. The analysis was run for 60 million generations, with trees sampled every 5,000 generations. To accommodate for the long-tree problem[1] (Marshall 2009), we changed the prior for branch length mean to *Unconstrained:Exponential(100)*, which is 1/10 of the default value[2]. We also assessed the heating (changed to 0.005) and unlinked the model parameters across partitions.

Table 2. Best substitution models for each partition as selected by JModelTest.

Gene	Substitution model	Number of free parameters
CCR5	GTR+I+G	10
IRBP	HKY+G	7
MC1R	GTR+G	9
Mitochondrial genes	GTR+I+G	10
PRP	GTR+G	9
SRY	GTR+I+G	10
TSPY	GTR+G	9

Notes: The gene names are abbreviated; see Table 1 for full names.

After tree inference, we chose a burn-in of 8,666 trees (43.33 million generations) for each of the three runs; thus, 10,000 trees contributed to the Bayesian tree block. Although the analysis seemed to converge before 43.33 million generations, we chose this value so that we had exactly 10,000 post-burnin trees left. We determined the burn-in and verified that the runs converged with the program Tracer (available at http://tree.bio.ed.ac.uk/software/tracer/). We summarized these topologies by constructing a 50% majority rule consensus tree, which is also

[1] The long-tree problem can be summarized as follows. Bayesian analyses may become trapped in regions of parameter space that are characterized by unrealistically long trees and distorted partition rate multipliers. Fortunately, however, this does typically not affect topological relationships.

[2] Various users reported in the internet that this modification was sufficient to solve the problem. The overall tree length of all four independent runs was very similar, which indicates that the analysis does not show the long-tree problem.

available on the *10kTrees* website. Branch lengths were calculated as the mean branch length from all trees in the posterior distribution in which the branch was present.

2.1.5. Dating the trees

For the dated tree, we inferred node ages using the mean molecular branch lengths (nucleotide substitutions per site) from the Bayesian search and six fossil calibration points employed by previous phylogenetic studies (Table 3, Godinot 2006; Hodgson et al. 2009; Seiffert et al. 2003; Yang and Yoder 2003; Yoder and Yang 2004). We conducted molecular dating with the software r8s (Sanderson 2002) using the penalized likelihood algorithm (Sanderson 2002) with a smoothing parameter of 100, chosen because this value best recovered dates inferred from phylogenetic analyses of smaller taxonomic samples but with more extensive sequence data (Hodgson et al. 2009; Yang and Yoder 2003; Yoder and Yang 2004).

Table 3. Fossil calibration ranges used to date the consensus molecular phylogeny.

MRCA node	**Min. Age (ma)**	**Max. Age (ma)**	**Source**
Homo- Pan	5	8	Haile-Selassie (2001), Senut *et al*. (2001), Vignaud *et al*. (2002), Brunet *et al*. (2002)
Homo- Pongo	12.5	18	Kelley (2002)
Papio- Theropithecus	3.5	6.5	Leakey (1993)
extant *Catarrhini*	21.0	30.0	Young & MacLatchy (2004), Benefit & McCrossin (2002)
Cebus- Saimiri	12.5	NA	Hartwig & Meldrum (2002)
Loris- Galago	38	42	Seiffert *et al*. (2003)

Notes: MRCA stands for most recent common ancestor.

A new feature of Version 3 (compared to Version 2 and Version 1) is that two extinct species (*Homo sapiens neanderthalensis* and *Archaeolemur majori*) are included in the trees. For dating the trees, we set the node age (i.e., the age of the tip) for *Homo sapiens neanderthalensis* to 0.03 (i.e., it disappeared around 30,000 years ago) and for *Archaeolemur majori* to 0.0073 (i.e., it disappeared around 730 years ago, Mittermeier et al. 1994).

2.2. Version 2

2.2.1. Data Collection

For the second version of the dataset, we collected data for six mitochondrial and three autosomal genes that were generally available in GenBank across 230 primate species and the outgroup species *Galeopterus variegates* (Sunda flying lemur). During data collection, we only included a gene if sequences were available for at least 65 different species. In conjunction with manually collecting sequences, we used the Phylota browser (release 1.01, Sanderson et al. 2008) for data collection and to identify the genes for which sufficient data were available. We excluded all sequences that were annotated as pseudogenes, similar to the particular gene of interest, or ambiguous in general. If multiple sequences from a particular gene were available for the same species, we selected the longest sequence (while controlling for ambiguous codes, such as *N,* which can stand for any of the four bases).

Table 4. Summary of the data collection for Version 2.

Gene Name (abbr.)	**Full name**	**Genomic position**	**Number of species for which seq. are available**
12S rRNA	12S ribosomal rRNA	MIT	168
16S rRNA	16S ribosomal rRNA	MIT	119
CCR5	Chemokine (C-C motif) receptor 5	CHR	70
Cluster of additional mitochondrial genes (COIIII, ND3, ND4L, ND4, various tRNA genes)	Cytochrome oxidase subunit III, NADH dehydrogenase subunit 3, 4, and 4L, tRNA-His, tRNA-Ser, tRNA-Gly, tRNA-Arg, tRNA-Leu	MIT	91
COX1	Cytochrome c oxidase subunit I	MIT	84
COX2	Cytochrome c oxidase subunit II	MIT	147
CYTB	Cytochrome B	MIT	182
MC1R	Melanocortin 1 receptor	CHR	69
SRY	Sex-determining Region Y	Y-CHR	77

Notes: MIT stands for mitochondrial, CHR for chromosome in general, while Y-CHR stands for Y-chromosome.

The following list summarizes the data collection for Version 2:

Number of species: 231 (only 230 are listed on the website, as we pruned the outgroup species from the trees after the tree inference)

Total number of available sequences: 1007 (out of 231*9 = 2079 total)

Percentage of missing data: 51.6% (59.0% if missing data within genes are also counted)

2.1.2. Multiple sequence alignments

For creating multiple sequence alignments (MSA) for each of the genes, we used Muscle 3.7 (default parameters except the following: -cluster2 neighborjoining, -distance1 kmer20_4). As it has been repeatedly demonstrated that alignment quality may have a substantial impact on the inferred tree (Kjer 1995; Morrison and Ellis 1997; Ogden and Rosenberg 2006; Smythe et al. 2006; Talavera and Castresana 2007), we eliminated poorly aligned positions and divergent regions of the alignment using the program Gblocks (Castresana 2002) with the settings -b5=h, -t=d, and -b2=0.6 * *number of sequences*. These positions may not be homologous or may have been saturated by multiple substitutions. Gblocks selects blocks in a similar way as it is usually done manually by hand. However, it follows a reproducible set of conditions, making the phylogenetic analyses of large datasets reliable, feasible, and also more accurate, especially because sequences in GenBank may be of bad quality. The multiple sequence alignments for each gene can be downloaded on the website.

2.2.3. Phylogenetic constraints

In Version 2, due to the increased number of available genes and sequences, we only constrained one major node (placement of the Tarsiers as sister group to monkeys and apes, see file “Phylogenetic constraints for the tree inference” on the website). See section 2.2.3 for an explanation why we defined constraints in our analysis.

2.2.4. Tree inference

For the tree inference, we used the program MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003). We used the species *Galeopterus variegatus* (Sunda flying lemur) as the outgroup, as it has

been shown that this species is the closest living relative to the order Primates (Janecka et al. 2007). We ran a Bayesian analysis with two runs and 16 chains in each run, but discarded one run after the analysis[3]. We used different substitution models (general time reversible (GTR) model (Rodriguez et al. 1990) and the SYM model (Zharkikh 1994), with a proportion of invariable sites and a gamma-shaped rate variation across sites, Table 5) for each of the genes (gene clusters) in a partitioned dataset, which were identified in the program JModelTest (Posada 2008) and Phyml (Guindon and Gascuel 2003). The analysis was run for 50.7 million generations, with trees sampled every 4,000 generations. We also assessed the heating (changed to 0.01) and unlinked the model parameters across partitions.

Table 5. Best substitution models for each partition as selected by JModelTest.

Gene	Substitution model	Number of free parameters
CYTB	GTR+I+G	10
COX1	GTR+I+G	10
COX2	GTR+I+G	10
12S rRNA	SYM+I+G	7
16S rRNA	SYM+I+G	7
Cluster of additional mitochondrial genes (see Table 1)	GTR+I+G	10
MC1R	GTR+G	9
CCR5	GTR+G	9
SRY	GTR+I	9

Notes: The gene names are abbreviated; see Table 1 for full names.

After tree inference, we chose a burn-in of 2,676 trees (10.7 million generations); thus, 10,000 trees contributed to the Bayesian tree block. Although the analysis seemed to converge before 10.7 million generations, we chose this value so that we had exactly 10,000 trees in the posterior sample. We determined the burn-in with the program Tracer (available at http://tree.bio.ed.ac.uk/software/tracer/). We summarized these topologies by constructing a

[3] Although the two runs converged on the same topology, estimates of the model parameters differed slightly between the runs. We thus selected the run that yielded more reliable results, based on different statistics and posterior probability distributions in the program Tracer and MrBayes.

50% majority rule consensus tree. Branch lengths were calculated as the mean branch length from all trees in the posterior distribution in which this branch is present.

2.2.5. Dating the trees

We used the same procedure and fossil calibration points as in Version 3 (see 2.1.5).

2.3. Version 1

Version 1 is a "beta" version of the *10kTrees* project, and was used as preliminary data for an NSF proposal. We believe that this version is suitable for comparative studies, and are even using it for our comparative research projects. However, as Version 2 and Version 3 are now available, we recommend using Version 2 or Version 3. We nevertheless still provide the option to use Version 1, and in what follows, details about the dataset and analysis are given.

2.3.1. Data Collection

For the first version of the dataset, we collected data for four mitochondrial genes and one autosomal gene that were generally available in GenBank across 189 primate species and the outgroup species *Galeopterus variegates (*Sunda flying lemur).

Table 6. Summary of the data collection for Version 1.

Gene (abbr.)	Full name	Position	No. of species for which seq. are available	Length	Average Length
CYTB	Cytochrome B	MIT	145	267-1162	940
COX1	Cytochrome c oxidase subunit I	MIT	66	505-1554	1017
COX2	Cytochrome c oxidase subunit II	MIT	109	210-746	627
ND1	NADH dehydrogenase subunit 1	MIT	29	934-957	955
SRY	Sex-determining Region Y	Y-CHR	71	347-832	770

Notes: MIT stands for mitochondrial, while Y-CHR stands for Y-chromosome.

The following list summarizes the data collection for Version 1:

Number of species: 190 (only 189 are listed on the website, as we pruned the outgroup species from the trees after the tree inference)

Total number of available sequences: 420 (out of 190*5 = 950 total)

Percentage of missing data: 56% (64% if missing data within genes are also counted)

2.3.2. Multiple sequence alignments

For creating multiple sequence alignments (MSA) for each of the genes, we used Muscle 3.7 with the default parameters. As it has been repeatedly demonstrated that alignment quality may have a substantial impact on the inferred tree (Kjer 1995; Morrison and Ellis 1997; Ogden and Rosenberg 2006; Smythe et al. 2006; Talavera and Castresana 2007), we manually excluded poorly aligned sites or sites with a high percentage of missing data (especially at the beginning and end of the MSA). The multiple sequence alignments for each gene can be downloaded on the website.

2.3.3. Phylogenetic constraints

We constrained 29 major nodes that were well characterized by at least three genomic Alu insertion events (Ray and Batzer 2005; Ray et al. 2005; Roos et al. 2004; Salem et al. 2003; Schmitz et al. 2001; Xing et al. 2005; Xing et al. 2007). Given the amount of sequence data available on Genbank for such a broad taxonomic sample, constraints based on insertion events were necessary to reduce phylogenetic uncertainty at deep nodes with short branches. The constraints eliminate all uncertainty at those nodes, but we think this is reasonable because Alu insertion events are generally regarded as more reliable cladistic indicators that are less prone to homoplasy than DNA sequence data (Ray and Batzer 2005; Ray et al. 2005; Xing et al. 2007). Had we not so constrained these deep nodes, then the limited available sequence data would have produced high levels of uncertainty, but this uncertainty would not have been reflective of the current state of knowledge of primate phylogeny. Only the actual history of evolutionary relationships is the truly relevant phylogeny for comparative methods, and controlling across unjustifiably variable phylogenies is known to produce elevated type 1 error and to reduce statistical power (Symonds 2002).

2.3.4. Tree inference

For the tree inference, we used the program MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003). We used the species *Galeopterus variegatus* (Sunda flying lemur) as outgroup, as it has been shown that this species is the closest living relative to the order Primates (Janecka et al. 2007). We ran a Bayesian analysis with two runs and 8 chains in each run. We used a GTR+I+G substitution model (general time reversible (GTR) model (Rodriguez et al. 1990) with a proportion of invariable sites and a gamma-shaped rate variation across sites) for each of the five genes in a partitioned dataset, which was identified as the best substitution model in the program FindModel (Tao et al. 2005). The analysis was run for 8 million generations, with trees sampled every 1000 generations. We assessed the heating (changed to 0.02).

After tree inference, we chose a burn-in of 2,000 trees on each of the two runs; thus, 12,000 trees contributed to the Bayesian tree block. We determined the burn-in with the program Tracer (available at http://tree.bio.ed.ac.uk/software/tracer/). We summarized these topologies by constructing a 50% majority rule consensus tree. Branch lengths were calculated as the mean branch length from all trees in the posterior distribution in which this branch is present.

2.3.5. Dating the trees

We used the same procedure and fossil calibration points as in Version 3 (see 2.1.5).

2.4. Version Comparison

2.4.1. Overview

Table 7. Comparison of Version 1, Version 2, and Version 3.

	Version 1	**Version 2**	**Version 3**
Species	187	231	301
Genes	4 mitochondrial (COI, COII, CYTB and ND1) and 1 autosomal gene (SRY)	6 mitochondrial (12S rRNA, 16S rRNA, COI, COII, CYTB, cluster of other mitochondrial genes) and 3 autosomal genes (SRY, CCR5, MC1R)	11 mitochondrial (12S rRNA, 16S rRNA, COI, COII, COIII, CYTB, ND1, ND3, ND4, ND4L, ND5) and 6 autosomal genes (SRY, CCR5, MC1R, PRP, TSPY, IRBP)
Genetic loci	2	4	7
Total No. of Sites	5134	9079	17972
Collected sequences	413 out of 935 total (55.8% missing data)	1007 out of 2079 total (51.6% missing data)	1894 out of 5117 total (63.0% missing data)
No. of constraints	29	1	4
Number of generations	8 millions for each of the two runs	50.7 millions for one run	60 millions for each of the three runs
Sampling frequency	every 1,000 generations	every 4,000 generations	every 5,000 generations
Number of chains	8 (one cold chain and 7 heated chains)	16 (one cold chain and 15 heated chains)	8 (one cold chain and 7 heated chains)
Burn-in	2 million generations	10.7 million generations	43.33 million generations
Computing time	~ 48 days (16 processors in parallel, ~ 3 days each)	~ 2 years (32 processors in parallel, ~ 3 weeks each)	~ 3.5 years (24 processors in parallel, ~ 7.7 weeks each)

2.4.2. Comparison of Version 1 and Version 2

The following two tables list the species differences for Version 1 and Version 2. Specifically, the left column lists species that are included in Version 1, but not in Version 2 (due to the different thresholds regarding gene availability when a species is included), and the right column lists species that are included in Version 2, but not in Version 1 (due to increased availability of sequence data).

Table 8. Species comparison for Version 1 and 2.

Included in Version 1, but not in Version 2	**Included in Version 2, but not in Version 1**
Alouatta belzebul	*Arctocebus aureus*
Alouatta guariba	*Ateles geoffroyi panamensis*
Aotus brumbacki	*Ateles geoffroyi vellerosus*
Aotus nigriceps	*Ateles geoffroyi yucatanensis*
Aotus vociferans	*Avahi occidentalis*
Cacajao melanocephalus	*Callicebus donacophilus*
Callicebus hoffmannsi	*Callithrix emiliae*
Callicebus personatus	*Cercocebus torquatus atys*
Callicebus torquatus	*Cercopithecus cephus cephus*
Cebus olivaceus	*Cercopithecus cephus ngottoensis*
Cercopithecus pogonias	*Cercopithecus erythrotis*
Galago matschiei	*Cheirogaleus crossleyi*
Phaner furcifer	*Chlorocebus pygerythrus*
Pithecia irrorata	*Chlorocebus sabaeus*
Presbytis comata	*Chlorocebus tantalus*
Saguinus bicolor	*Eulemur fulvus albocollaris*
Saguinus leucopus	*Eulemur fulvus collaris*
Saguinus mystax	*Eulemur fulvus fulvus*
Saguinus tripartitus	*Eulemur fulvus mayottensis*
Saimiri boliviensis	*Eulemur fulvus rufus*
Saimiri ustus	*Eulemur fulvus sanfordi*
Trachypithecus geei	*Eulemur macaco macaco*
	Lepilemur aeeclis
	Lepilemur ankaranensis
	Lepilemur mitsinjoensis
	Lepilemur randrianasoli
	Lepilemur sahamalazensis
	Lepilemur seali
	Lophocebus aterrimus
	Loris lydekkerianus malabaricus
	Loris tardigradus nordicus
	Macaca brunnescens
	Macaca hecki
	Macaca leonina
	Macaca nemestrina leonina
	Macaca nemestrina nemestrina
	Macaca nemestrina siberu
	Macaca nigrescens
	Macaca pagensis
	Microcebus berthae
	Microcebus bongolavensis
	Microcebus danfossi
	Microcebus griseorufus
	Microcebus jollyae
	Microcebus lehilahytsara
	Microcebus lokobensis
	Microcebus mittermeieri

	Microcebus myoxinus
	Microcebus ravelobensis
	Microcebus sambiranensis
	Microcebus simmonsi
	Microcebus tavaratra
	Nomascus concolor
	Pongo abelii
	Propithecus coquereli
	Propithecus edwardsi
	Rungwecebus kipunji
	Saguinus fuscicollis
	Saguinus imperator
	Saimiri boliviensis boliviensis
	Trachypithecus poliocephalus
	Varecia rubra

Table 9. Species in Version 1 that have a different taxonomical name or classification in Version 2.

Name of species in Version 1	**Name of corresponding species in Version 2**
Alouatta_palliata coibensis	*Alouatta palliata*
Aotus lemurinus	*Aotus lemurinus griseimembra*
Ateles belzebuth chamek	*Ateles belzebuth*
Ateles belzebuth marginatus	
Gorilla gorilla	*Gorilla gorilla gorilla*
Hapalemur griseus	*Hapalemur griseus alaotrensis*
	Hapalemur griseus griseus
	Hapalemur griseus meridionalis
	Hapalemur griseus occidentalis
Pan troglodytes	*Pan troglodytes schweinfurthii*
	Pan troglodytes troglodytes
	Pan troglodytes verus
Pongo pygmaeus	*Pongo pygmaeus pygmaeus*
Propithecus verreauxi	*Propithecus verreauxi verreauxi*
Varecia variegata	*Varecia variegata variegata*

2.4.3. Comparison of Version 1 and Version 3

The following two tables list the species differences for Version 1 and Version 3. Specifically, the left column lists species that are included in Version 1, but not in Version 3 (due to the different thresholds regarding gene availability when a species is included), and the right column lists species that are included in Version 3, but not in Version 1 (due to increased availability of sequence data).

Table 10. Species comparison for Version 1 and 3.

Included in Version 1, but not in Version 3	**Included in Version 3, but not in Version 1**
All species in Version 1 are included in Version 3 (see also Table 11)	*Aotus azarai boliviensis*
	Aotus lemurinus griseimembra
	Archaeolemur majori
	Arctocebus aureus
	Avahi cleesei
	Avahi occidentalis
	Avahi unicolor
	Cacajao calvus
	Callicebus donacophilus
	Callithrix emiliae
	Callithrix mauesi
	Cebus xanthosternos
	Cercocebus torquatus atys
	Cercopithecus albogularis
	Cercopithecus campbelli
	Cercopithecus cephus cephus
	Cercopithecus cephus ngottoensis
	Cercopithecus erythrogaster
	Cercopithecus erythrotis
	Cheirogaleus crossleyi
	Chiropotes satanas
	Chlorocebus pygerythrus
	Chlorocebus pygerythrus cynosurus
	Chlorocebus sabaeus
	Chlorocebus tantalus
	Colobus angolensis palliatus
	Colobus satanas
	Colobus vellerosus
	Eulemur fulvus albocollaris
	Eulemur fulvus collaris
	Eulemur fulvus fulvus
	Eulemur fulvus mayottensis
	Eulemur fulvus rufus
	Eulemur fulvus sanfordi
	Eulemur macaco macaco
	Galago granti
	Gorilla beringei
	Hapalemur griseus alaotrensis
	Hapalemur griseus griseus
	Hapalemur griseus meridionalis
	Hapalemur griseus occidentalis
	Homo sapiens neanderthalensis
	Lepilemur aeeclis
	Lepilemur ankaranensis
	Lepilemur hubbardorum
	Lepilemur manasamody

	Lepilemur mitsinjoensis
	Lepilemur otto
	Lepilemur randrianasoli
	Lepilemur sahamalazensis
	Lepilemur seali
	Lophocebus aterrimus
	Loris lydekkerianus
	Macaca brunnescens
	Macaca hecki
	Macaca leonina
	Macaca munzala
	Macaca nemestrina leonina
	Macaca nemestrina siberu
	Macaca nigrescens
	Macaca pagensis
	Microcebus berthae
	Microcebus bongolavensis
	Microcebus danfossi
	Microcebus griseorufus
	Microcebus jollyae
	Microcebus lehilahytsara
	Microcebus lokobensis
	Microcebus macarthurii
	Microcebus mamiratra
	Microcebus mittermeieri
	Microcebus myoxinus
	Microcebus ravelobensis
	Microcebus sambiranensis
	Microcebus simmonsi
	Microcebus tavaratra
	Mirza zaza
	Nomascus concolor
	Nomascus nasutus
	Nomascus siki
	Nycticebus bengalensis
	Nycticebus javanicus
	Nycticebus menagensis
	Phaner furcifer pallescens
	Piliocolobus foai
	Piliocolobus gordonorum
	Piliocolobus kirkii
	Piliocolobus pennantii
	Piliocolobus preussi
	Piliocolobus rufomitratus
	Piliocolobus tephrosceles
	Piliocolobus tholloni
	Pithecia pithecia
	Pongo abelii
	Procolobus verus

	Propithecus coquereli
	Propithecus deckenii
	Propithecus edwardsi
	Pygathrix cinerea
	Rungwecebus kipunji
	Saguinus fuscicollis
	Saguinus fuscicollis melanoleucus
	Saguinus imperator
	Saguinus niger
	Tarsius dentatus
	Tarsius lariang
	Trachypithecus delacouri
	Trachypithecus germaini
	Trachypithecus laotum
	Trachypithecus poliocephalus
	Varecia rubra
	Varecia variegata variegata
	Tarsius lariang
	Trachypithecus delacouri
	Trachypithecus germaini
	Trachypithecus laotum
	Trachypithecus poliocephalus
	Varecia rubra

Table 11. Species in Version 1 that have a different taxonomical name or classification in Version 3.

Name of species in Version 1	**Name of corresponding species in Version 3**
Alouatta palliata coibensis	*Alouatta palliata*
Ateles belzebuth chamek	*Ateles belzebuth*
Ateles belzebuth marginatus	
Gorilla gorilla	*Gorilla gorilla gorilla*
	Gorilla gorilla graueri
Pan troglodytes	*Pan troglodytes schweinfurthii*
	Pan troglodytes troglodytes
	Pan troglodytes vellerosus
	Pan troglodytes verus
Varecia variegata	*Varecia variegata variegata*

2.4.4. Comparison of Version 2 and Version 3

The following tables list the species differences for Version 2 and Version 3. Specifically, the left column lists species that are included in Version 2, but not in Version 3 (due to the different thresholds regarding gene availability when a species is included), and the right column lists species that are included in Version 3, but not in Version 2 (due to increased availability of sequence data).

Table 12. Species comparison for Version 2 and 3.

Included in Version 2, but not in Version 3	**Included in Version 3, but not in Version 2**
Macaca nemestrina nemestrina	*Alouatta belzebul*
	Alouatta guariba
	Aotus azarai boliviensis
	Aotus brumbacki
	Aotus lemurinus
	Aotus nigriceps
	Aotus vociferans
	Archaeolemur majori
	Avahi cleesei
	Avahi unicolor
	Cacajao calvus
	Cacajao melanocephalus
	Callicebus hoffmannsi
	Callicebus personatus
	Callicebus torquatus
	Callithrix mauesi
	Cebus olivaceus
	Cebus xanthosternos
	Cercopithecus albogularis
	Cercopithecus campbelli
	Cercopithecus erythrogaster
	Cercopithecus pogonias
	Chiropotes satanas
	Chlorocebus pygerythrus cynosurus
	Colobus angolensis palliatus
	Colobus satanas
	Colobus vellerosus
	Galago granti
	Galago matschiei
	Gorilla beringei
	Gorilla gorilla graueri
	Hapalemur griseus
	Homo sapiens neanderthalensis
	Lepilemur hubbardorum
	Lepilemur manasamody
	Lepilemur otto
	Macaca munzala
	Microcebus macarthurii
	Microcebus mamiratra
	Mirza zaza
	Nomascus nasutus
	Nomascus siki
	Nycticebus bengalensis
	Nycticebus javanicus
	Nycticebus menagensis
	Pan troglodytes vellerosus
	Phaner furcifer

	Phaner furcifer pallescens
	Piliocolobus foai
	Piliocolobus gordonorum
	Piliocolobus kirkii
	Piliocolobus pennantii
	Piliocolobus preussi
	Piliocolobus rufomitratus
	Piliocolobus tephrosceles
	Piliocolobus tholloni
	Pithecia irrorata
	Pithecia pithecia
	Presbytis comata
	Procolobus verus
	Propithecus deckenii
	Pygathrix cinerea
	Saguinus bicolor
	Saguinus fuscicollis melanoleucus
	Saguinus leucopus
	Saguinus mystax
	Saguinus niger
	Saguinus tripartitus
	Saimiri ustus
	Tarsius dentatus
	Tarsius lariang
	Trachypithecus delacouri
	Trachypithecus geei
	Trachypithecus germaini
	Trachypithecus laotum
	Trachypithecus vetulus

Table 13. Species in Version 2 that have a different taxonomical name or classification in Version 3.

Name of species in Version 2	**Name of corresponding species in Version 3**
Ateles geoffroyi panamensis	*Ateles geoffroyi*
Ateles geoffroyi vellerosus	
Ateles geoffroyi yucatanensis	
Loris lydekkerianus malabaricus	*Loris lydekkerianus*
Loris tardigradus nordicus	*Loris tardigradus*
Pongo pygmaeus pygmaeus	*Pongo pygmaeus*
Propithecus verreauxi verreauxi	*Propithecus verreauxi*
Saimiri boliviensis boliviensis	*Saimiri boliviensis*

.

3. Methodological Details for the Odd-toed Ungulates Part of the Website

In what follows, we provide details on each of the versions, beginning with the most recent version.

3.1. Version 1

3.1.1. Data Collection

For the first version of the dataset, we collected data for eleven mitochondrial and four autosomal genes that were generally available in GenBank across all 17 extant odd-toed ungulates species and the outgroup species *Bos taurus* (cattle). During data collection, we only included a gene if sequences were available for at least 7 different species. We used the Phylota browser (Sanderson et al. 2008) (rel. 1.5) for data collection and to identify the genes for which sufficient data were available and automatically downloaded all available sequences for each of the species in this dataset using the bioinformatics pipeline FAST (Arnold 2012, in prep.). We strictly excluded all sequences that were annotated as pseudogenes or hypothetical or working draft etc., similar to the particular gene of interest, or ambiguous (in annotation) in general. If multiple sequences from a particular gene were available for the same species, we selected the longest sequence (while controlling for ambiguous codes, such as *N,* which can stand for any of the four bases). If the whole mitochondrial genome for a particular species was available, we always extracted and selected the sequences for the genes of interest from the mitochondrion, rather than taking the sequences that were available on GenBank (if any were available at all). This substantially improved the quality of the sequences and the subsequent alignments.

Table 14. Summary of the data collection for Version 1.

Gene Name (abbr.)	Full name	Genomic position	Number of species for which seq. are available
12S rRNA	12S ribosomal rRNA	MIT	17
16S rRNA	16S ribosomal rRNA	MIT	10
COX1	Cytochrome c oxidase subunit I	MIT	9
COX2	Cytochrome c oxidase subunit II	MIT	12
COX3	Cytochrome c oxidase subunit III	MIT	8
CYTB	Cytochrome B	MIT	14
MC1R	Melanocortin 1 receptor	CHR	8
ND1	NADH dehydrogenase subunit 1	MIT	8
ND3	NADH dehydrogenase subunit 3	MIT	8
ND4	NADH dehydrogenase subunit 4	MIT	8
ND4L	NADH dehydrogenase subunit 4L	MIT	8
ND5	NADH dehydrogenase subunit 5	MIT	8
PRND	Prion protein 2 (dublet)	CHR	9
PRP	Major prion protein (encoded by the *PRNP* gene)	CHR	8
SRY	Sex-determining Region Y	Y-CHR	7

Notes: MIT stands for mitochondrial, CHR for chromosome in general, while Y-CHR stands for Y-chromosome.

The following list summarizes the data collection for Version 1:

Number of species: 18 (only 17 are listed on the website, as we pruned the outgroup from the trees after the tree inference)

Total number of available sequences: 142 (out of 18*15 = 270 total)

Percentage of missing data: 47.4% (50.3% if missing data within genes are also counted)

3.1.2. Multiple sequence alignments

For creating multiple sequence alignments (MSA) for each of the genes, we used Muscle 3.7 (default parameters except the following: -cluster1 neighborjoining -maxtrees 5 -noanchors -cluster2 neighborjoining -distance1 kmer20_4). As it has been repeatedly demonstrated that alignment quality may have a substantial impact on the inferred tree (Kjer 1995; Morrison and

Ellis 1997; Ogden and Rosenberg 2006; Smythe et al. 2006; Talavera and Castresana 2007), we eliminated poorly aligned positions and divergent regions of the alignment using the program Gblocks (Castresana 2002) with the settings -b5=h, -t=d, and -b2=0.6 * *number of sequences*. These positions may not be homologous or may have been saturated by multiple substitutions. Gblocks selects blocks in a similar way as it is usually done manually by hand. However, it follows a reproducible set of conditions, making the phylogenetic analyses of large datasets reliable, feasible, and also more accurate, especially because sequences in GenBank may be of bad quality. The multiple sequence alignments for each gene can be downloaded on the website. For some of the genes (e.g., 12S rRNA), we furthermore manually improved the quality of the alignment.

3.1.3. Phylogenetic constraints

In Version 1, we did not include any phylogenetic constraints.

3.1.4. Tree inference

For the tree inference, we used the program MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003). We used the species *Bos taurus* (cattle) as the outgroup. We ran a Bayesian analysis with four runs and 8 chains in each run. We used different substitution models (general time reversible (GTR) model (Rodriguez et al. 1990) and the HKY model (Hasegawa et al. 1985) with a proportion of invariable sites and a gamma-shaped rate variation across sites, Table 13) for each of the genes (gene clusters) in a partitioned dataset, which were identified in the program JModelTest (Posada 2008) and Phyml (Guindon and Gascuel 2003). If the best-suited substitution model determined by JModelTest was not available in MrBayes, we selected the model with the lowest AIC score among the models that are implemented in MrBayes. The analysis was run for 15 million generations, with trees sampled every 2,000 generations. To accommodate for the long-tree problem[4] (Marshall 2009), we changed the prior for branch

[4] The long-tree problem can be summarized as follows. Bayesian analyses may become trapped in regions of parameter space that are characterized by unrealistically long trees and distorted partition rate multipliers. Fortunately, however, this does typically not affect topological relationships.

length mean to *Unconstrained:Exponential(100)*, which is 1/10 of the default value[5]. We also assessed the heating (changed to 0.08) and unlinked the model parameters across partitions.

Table 15. Best substitution models for each partition as selected by JModelTest.

Gene Name (abbr.)	Substitution model	Number of free parameters
CYTB	GTR+G	9
COX1	GTR+I+G	10
COX2	HKY+I+G	6
COX3	HKY+I+G	6
12S rRNA	GTR+I+G	10
16S rRNA	GTR+G	9
ND1	HKY+I	5
ND3	HKY+I	5
ND4	GTR+I+G	10
ND4L	HKY+G	5
ND5	HKY+I+G	6
MC1R	HKY+G	5
PRND	GTR	8
PRP	GTR+G	9
SRY	GTR	8

Notes: The gene names are abbreviated; see Table 14 for full names.

After tree inference, we chose a burn-in of 5001 trees (approximately 10 million generations). Thus, in all four runs, a total of 10,000 trees contributed to the Bayesian tree block. Although the analysis clearly seemed to converge before 10 million generations, we chose this value so that we had exactly 10,000 trees remaining in the posterior sample (note that this somewhat arbitrary decision is not an issue, since convergence was *before* this value). We determined the burn-in with the program Tracer (available at http://tree.bio.ed.ac.uk/software/tracer/). Furthermore, we verified that our Bayesian analysis reached (apparent) stationarity with the online tool AWTY (http://ceb.csit.fsu.edu/awty/) (Nylander et al. 2008), Tracer, and the convergence diagnostics from MrBayes (in particular,

[5] Various users reported in the internet that this modification was sufficient to solve the problem. The overall tree length of all four independent runs was very similar, which indicates that the analysis does not show the long-tree problem.

the "potential scale reduction factor"). We summarized these topologies by constructing a 50% majority rule consensus tree. Branch lengths were calculated as the mean branch length from all trees in the posterior distribution in which this branch was present.

3.3.5. Dating the trees

For the dated tree, we inferred node ages using the mean molecular branch lengths (nucleotide substitutions per site) from the Bayesian search and three fossil calibration points, which we extracted from the *Paleobiology Database* (http://paleodb.org) (Table 16). We conducted molecular dating with the software r8s (Sanderson 2002) using the penalized likelihood method in combination with the TN algorithm (Sanderson 2002) with a smoothing parameter of 100, chosen because this value best recovered dates inferred from phylogenetic analyses of smaller taxonomic samples but with more extensive sequence data (Hodgson et al. 2009; Yang and Yoder 2003; Yoder and Yang 2004). Additionally, we set some parameters to non-default values to improve robustness and convergence of the results (*num_restarts=5, nun_time_guesses=5, checkGradient=yes*). Lastly, it was necessary to collapse internal branches of length 0 or very close to 0.

Table 16. Fossil calibration ranges used to date the consensus molecular phylogeny.

MRCA node	Min. Age (ma)	Max. Age (ma)	Source
Equus asinus – Equus caballus przewalskii	5.3	7.2	http://paleodb.org
Rhinoceros sondaicus - Rhinoceros unicornis	20.4	23	http://paleodb.org
Tapirus bairdii – Tapirus indicus	28.4	33.9	http://paleodb.org

Notes: MRCA stands for most recent common ancestor.

4. Methodological Details for the Carnivorans Part of the Website

In what follows, we provide details on each of the versions, beginning with the most recent version.

4.1. Version 1

4.1.1. Data Collection

For the first version of the dataset, we collected data for 14 mitochondrial and 15 autosomal genes that were generally available in GenBank across carnivoran species and the outgroup species *Equus caballus* (horse). During data collection, we only included a gene if sequences were available for at least 70 different species. We used the Phylota browser (release 1.5, Sanderson et al. 2008) for data collection and to identify the genes for which sufficient data were available and automatically downloaded all available sequences for each of the species in this dataset using in-house bioinformatics pipelines. We strictly excluded all sequences that were annotated as pseudogenes or hypothetical or working draft etc., similar to the particular gene of interest, or ambiguous (in annotation) in general. If multiple sequences from a particular gene were available for the same species, we selected the longest sequence (while controlling for ambiguous codes, such as *N,* which can stand for any of the four bases). If the whole mitochondrial genome for a particular species was available, we always extracted and selected the sequences for the genes of interest from the mitochondrion, rather than taking the sequences that were available on GenBank (if any were available at all). This substantially improved the quality of the sequences and the subsequent alignments.

Table 17. Summary of the data collection for Version 1.

Gene Name (abbr.)	Full name	Genomic position	Number of species for which seq. are available
12S rRNA	12S ribosomal rRNA	MIT	112
16S rRNA	16S ribosomal rRNA	MIT	114
ADORA3	adenosine A3 receptor	CHR	108
APOB	apolipoprotein B	CHR	118
ATPASE6	ATPase 6	MIT	83
ATPASE8	ATPase 8	MIT	84
BDNF	brain derived neurotrophic factor	CHR	135
BRCA1	breast and ovarian cancer susceptibility protein 1, exon 9	CHR	70
CHRNA1	nicotinic cholinergic receptor alpha polypeptide 1 precursor	CHR	157
COX1	Cytochrome c oxidase subunit I	MIT	115
COX2	Cytochrome c oxidase subunit II	MIT	107
COX3	Cytochrome c oxidase subunit III	MIT	84
CYTB	Cytochrome B	MIT	225
GHR	growth hormone receptor	CHR	127
IRBP	interphotoreceptor retinoid-binding protein	CHR	135
ND1	NADH dehydrogenase subunit 1	MIT	83
ND2	NADH dehydrogenase subunit 12	MIT	150
ND3	NADH dehydrogenase subunit 3	MIT	83
ND4	NADH dehydrogenase subunit 4	MIT	90
ND4L	NADH dehydrogenase subunit 4L	MIT	83
ND5	NADH dehydrogenase subunit 5	MIT	130
PNOC	prepronociceptin	CHR	138
RAG1	recombination activating protein 1, exon 1	CHR	105
RAG2	recombination activating protein 2	CHR	134
RHO	rhodopsin	CHR	75
SRY	sex-determining Region Y	Y-CHR	74
TMEM20	transmembrane protein 20	CHR	74
TRANSTHYRETIN	transthyretine, intron 1	CHR	70
WILLEBRAND	von Willebrand factor	CHR	91

Notes: MIT stands for mitochondrial, CHR for chromosome in general, while Y-CHR stands for Y-chromosome.

The following list summarizes the data collection for Version 1:

Number of species: 253 (only 252 are listed on the website, as we pruned the outgroup from the trees after the tree inference)

Total number of available sequences: 3,154 (out of 253*29 = 7,337 total)

Percentage of missing data: 57.0% (58.9% if missing data within genes are also counted)

4.1.2. Multiple sequence alignments

For creating multiple sequence alignments (MSA) for each of the genes, we used clustalw2 (default parameters except the following: *-CLUSTERING=NJ -OUTPUTTREE=nexus -ITERATION=TREE -NUMITER=5 -ALIGN -CONVERT -PIM -OUTORDER=INPUT -OUTPUT=FASTA –TREE*). The multiple sequence alignments for each gene can be downloaded on the *10kTrees Website*. We performed a manual quality control for all the genes after the alignment, and for some of the genes (e.g., 12S rRNA), we furthermore manually improved the quality of the alignment, for example by removing ambiguously aligned regions.

4.1.3. Phylogenetic constraints

In Version 1, we did not include any phylogenetic constraints.

4.1.4. Tree inference

For the tree inference, we used the program MrBayes 3.2 (Ronquist and Huelsenbeck 2003). MrBayes versions prior to 3.2 tend to mix very slowly across different tree lengths, as the only proposals it uses to change tree lengths are updates to branches one at a time. From our experience, for large and complex datasets, this makes the analysis and convergence extremely difficult and time-consuming. We therefore used a special version of MrBayes 3.2 (rev. 390) with a modification from Jeremy Brown, who implemented a new scaling move that mixes much better across different tree lengths (see Brown et al. 2009 for more information). This move has now been implemented in the newer MrBayes 3.2 revisions as well. We used the species *Equus caballus* (horse) as outgroup. We ran a Bayesian analysis with four runs and six

chains in each run. We used reversible jump MCMC (RJ-MCMC) to allow MrBayes 3.2 to move across different schemes as part of its MCMC sampling. As reversible jumping is not currently set up for different models of rate variation across sites, it is still necessary to specify if a proportion of invariable sites and a gamma-shaped rate variation across sites should be used for each gene. For this, we identified the best-suited substitution model using JModelTest (Posada 2008) and Phyml (Guindon and Gascuel 2003) and chose the rate variation accordingly (Table 18). If both a proportion of invariable sites and a gamma-shaped rate variation across sites was selected by JModelTest, we changed the prior for the gamma-shaped rate variation (*shapepr = uniform(1.01,50.0)* instead of the default *uniform(1.01,50.0)*). The reason for this adjustment is that is has been shown that the two heterogeneity parameters (α and θ) are not genuinely independent, and it is extremely difficult to distinguish the effects from these parameters. Thus, several combinations of these two parameters appear to be almost equally probable, which may also cause convergence problems. To address this issue, we followed a recommendation of Gangolf Jobb and bound α to values bigger than 1. Essentially, this "avoids a situation where both parameters have nearly the same effect on the distribution shape and, as a consequence, 'fight' to explain the data. On the other hand, this constrained I+Gamma has still the advantages it was made for: It can produce two-peaked rate distributions as well as one-peaked ones, ranging from homogeneity to extremely L-shaped and anything between" (see http://evol.mcmaster.ca/~brian/evoldir/Answers/ GammaI.model.answers for a discussion on that issue). The analysis was run for 50 million generations, with trees sampled every 4,000 generations. To accommodate for the long-tree problem[6] (Marshall 2009), we changed the prior for branch length mean to *Unconstrained:Exponential(100)*, which is 1/10 of the default value[7]. We also assessed the heating (changed to 0.02), the number of swaps tried for each swapping generation of the chain (*Nswaps* value of 2), and unlinked the model parameters across partitions.

---

[6] The long-tree problem can be summarized as follows. Bayesian analyses may become trapped in regions of parameter space that are characterized by unrealistically long trees and distorted partition rate multipliers. Fortunately, however, this does typically not affect topological relationships.

[7] Various users reported in the internet that this modification was sufficient to solve the problem. The overall tree length of all four independent runs was very similar, which indicates that the analysis does not show the long-tree problem.

Table 18. Rate variation for each partition as selected by JModelTest.

Gene Name (abbr.)	**Rate variation**
12S rRNA	+I+G
16S rRNA	+I+G
ADORA3	+G
APOB	+G
ATPASE6	+I+G
ATPASE8	+G
BDNF	+I+G
BRCA1	+G
CHRNA1	none
COX1	+I+G
COX2	+I+G
COX3	+I+G
CYTB	+I+G
GHR	+G
IRBP	+I+G
ND1	+I+G
ND2	+I+G
ND3	+I+G
ND4	+I+G
ND4L	+I+G
ND5	+I+G
PNOC	none
RAG1	+I+G
RAG2	+G
RHO	+G
SRY	+I
TMEM20	+G
TRANSTHYRETIN	+G
WILLEBRAND	+I+G

Notes: The gene names are abbreviated; see Table 17 for full names.

After tree inference, we chose a burn-in of 10,001 trees (approximately 40 million generations). Thus, in all four runs, a total of 10,000 trees contributed to the Bayesian tree block. Although the analysis clearly seemed to converge before 40 million generations, we chose this value so that we had exactly 10,000 trees left (note that this somewhat arbitrary decision is not an issue, since convergence was *before* this value). We determined the burn-in with the program Tracer (available at http://tree.bio.ed.ac.uk/software/tracer/). Furthermore, we verified that our Bayesian analysis reached (apparent) convergence with the online tool AWTY (http://ceb.csit.fsu.edu/awty/) (Nylander et al. 2008), Tracer, and the convergence diagnostics from MrBayes (for example, the "potential scale reduction factor"). The convergence diagnostics statistics from AWTY are available upon request. We summarized these topologies by constructing a 50% majority rule consensus tree. Branch lengths were calculated as the mean branch length from all trees in the posterior distribution in which this branch was present.

4.3.5. Dating the trees

For the dated tree, we inferred node ages using the mean molecular branch lengths (nucleotide substitutions per site) from the Bayesian search and 16 fossil calibration points, which we extracted from the *Paleobiology Database* (http://paleodb.org) (Table 19). For more methodological details, see section 3.3.5.

Table 19. Fossil calibration ranges used to date the consensus molecular phylogeny.

MRCA node	Min. Age (ma)	Max. Age (ma)	Source
Urocyon cinereoargenteus – Urocyon littoralis	1.8	4.9	http://paleodb.org
Panthera leo - Panthera tigris	4.2	4.9	http://paleodb.org
Atilax paludinosus - Suricata suricatta	5.3	7.2	http://paleodb.org
Phoca largha - Phoca vitulina	11.6	12.7	http://paleodb.org
Acinonyx jubatus – Prionailurus rubiginosa	11.6	13.6	http://paleodb.org
Genetta angolensis - Genetta johnstoni	11.6	13.7	http://paleodb.org
Conepatus chinga - Mydaus marchei	13.6	16	http://paleodb.org
Mustela africana - Mustela strigidorsa	16	20.4	http://paleodb.org
Crocuta crocuta - Proteles cristatus	16	16.9	http://paleodb.org
Gulo gulo - Martes pennanti	20	22.4	http://paleodb.org
Bassaricyon alleni - Potos flavus	23	24.8	http://paleodb.org

Arctocephalus australis - Monachus schauinslandi	28.4	33.9	http://paleodb.org
Ailuropoda melanoleuca - Melursus ursinus	33.9	37.2	http://paleodb.org
Atelocynus microtis – Vulpes macrotis	40.4	46.2	http://paleodb.org
Acinonyx jubatus - Nandina binotata	61.7	63.3	http://paleodb.org
Ailuropoda melanoleuca – Monachus schauinslandi	164.7	175.6	http://paleodb.org

Notes: MRCA stands for most recent common ancestor.

5. Methodological Details for the Cetartiodactyla Part of the Website

In what follows, we provide details on each of the versions, beginning with the most recent version.

5.1. Version 1

5.1.1. Data Collection

For the first version of the dataset, we collected data for 14 mitochondrial and six autosomal genes that were generally available in GenBank across even-toed ungulates and cetaceans species and the outgroup species *Equus caballus* (horse). During data collection, we only included a gene if sequences were available for at least 55 different species. For more details, see section 4.1.1.

During data collection, we identified species names synonyms (e.g., due to genus name changes) and merged duplicate species. Also, some artiodactyls species are domesticated (e.g., the pig), and we included the wild and domesticated version only if the source of the sequences reliably indicated that they come from the wild or domesticated species. Thus, we merged the species *Bos frontalis* (gayal, domesticated) and *Bos gaurus* (gaur, wild), as the GenBank sequences did not clearly indicate the origin of the sequences. The same was true for *Bubalus carabanensis* and *Bubalus bubalis* (water buffalo).

After tree inference, we pruned the species *Hyemoschus aquaticus* from all trees due to an odd topological placement caused by the limited sequence availability for this species and/or potential sequence issues for the cytochrome B gene.

Table 20. Summary of the data collection for Version 1.

Gene Name (abbr.)	Full name	Genomic position	Number of species for which seq. are available
12S rRNA	12S ribosomal rRNA	MIT	233
16S rRNA	16S ribosomal rRNA	MIT	213
ATPASE6	ATPase 6	MIT	102
ATPASE8	ATPase 8	MIT	56
COX1	Cytochrome c oxidase subunit I	MIT	161
COX2	Cytochrome c oxidase subunit II	MIT	139
COX3	Cytochrome c oxidase subunit III	MIT	87
CSN3	kappa-casein	CHR	86
CYTB	Cytochrome B	MIT	294
MC1R	melanocortin-1 receptor	CHR	86
ND1	NADH dehydrogenase subunit 1	MIT	117
ND2	NADH dehydrogenase subunit 12	MIT	104
ND3	NADH dehydrogenase subunit 3	MIT	108
ND4	NADH dehydrogenase subunit 4	MIT	109
ND4L	NADH dehydrogenase subunit 4L	MIT	109
ND5	NADH dehydrogenase subunit 5	MIT	102
PRKCI	protein kinase C iota	CHR	90
PRP	prion protein	CHR	62
SPTBN1	B-spectrin nonerythrocytic 1	CHR	74
SRY	sex-determining Region Y	Y-CHR	68

Notes: MIT stands for mitochondrial, CHR for chromosome in general, while Y-CHR stands for Y-chromosome.

The following list summarizes the data collection for Version 1:

Number of species: 301 (only 299 are listed on the website, as we pruned the outgroup from the trees after the tree inference and the species *Hyemoschus aquaticus*, see text)

Total number of available sequences: 2400 (out of 301*20 = 6,020 total)

Percentage of missing data: 60.0% (61.3% if missing data within genes are also counted)

5.1.2. Multiple sequence alignments

See section 4.1.2 for details

5.1.3. Phylogenetic constraints

In Version 1, we did not include any phylogenetic constraints.

5.1.4. Tree inference

For the full methodological details, see section 4.1.4. We here describe only the differences to what we describe in section 4.1.4.

The analysis was run for 80 million generations, and we assessed the heating (changed to 0.015) and the number of swaps tried for each swapping generation of the chain (*Nswaps* value of 3).

Table 21. Rate variation for each partition as selected by JModelTest.

Gene Name (abbr.)	**Rate variation**
12S rRNA	+G
16S rRNA	+I+G
ATPASE6	+I+G
ATPASE8	+G
COX1	+I+G
COX2	+I+G
COX3	+I+G
CSN3	+G
CYTB	+I+G
MC1R	+G
ND1	+I+G
ND2	+I+G
ND3	+I+G
ND4	+I+G
ND4L	+G
ND5	+I+G
PRKCI	+G
PRP	+I+G
SPTBN1	+G
SRY	+G

Notes: The gene names are abbreviated; see Table 20 for full names.

5.3.5. Dating the trees

We do not provide dated trees yet, but we will in the near future.

6. Using the Website for Downloading Trees

6.1. Requirements for the *10kTrees Website*

We recommend using a modern web browser (we tested the website with Mozilla Firefox and Safari). Also, we strongly recommend enabling JavaScript, as we implemented a set of features that enhance usability and user-friendliness that require JavaScript. For example, users with disabled JavaScript will be unable to use the help system on the website (see below); neither will they be able to take advantage of the intuitive species selection feature and the progress bar that indicates computational progress when downloading trees.

6.2. Using the help system on the website

We implemented a help system on the website. If a ? symbol is displayed left to a link or to text, a help popup will open if you move with the mouse over the text right to the ? symbol or the ? symbol itself. The small help window provides explanations, additional information or other general instructions relevant to the particular feature.

6.3. Educational tools

In the "How To Use" section, we now provide four tutorials how to actually use the *10kTrees* website for your research, and what to do with so many tree. For example, we provide instructions on downloading trees and viewing them, as well as running analyses across a tree block. For more details, see http://10ktrees.fas.harvard.edu/howToUse.html.

6.4. Downloading trees

In the "Download Trees" section of the website, users are able to download the trees produced by our Bayesian tree search. Here, we provide some instructions for downloading the trees and describe the options that the user has. The website also provides an intuitive help system.

Only five steps are needed to download the trees:

1. Select the version of the dataset; we recommend using the most recent version. This selection is currently only available for the order Primates.
2. Select a taxonomy.
3. Specify the number of trees and whether to include a consensus tree.

4. Select if the trees should be dated (a chronogram). This selection is not yet available for the Cetartiodactyla part of the website.
5. Select the species that should be included in the trees, and choose among several display options.

1. Selecting the version of the dataset

First, users have to decide which version of the dataset they want to use for downloading the trees. By default, the latest version is selected. To change to previous versions, simply check the appropriate box in the *"Which version do you want to use?"* section. If you change the version of the dataset, however, your current selection of species will be lost, unless you saved your selected set of species earlier (see below).

2. Selecting a taxonomy

We also provide a taxonomic translation tool. Readers are able to select species based on their names from GenBank, or from lists of names in which the original species designations are translated to commonly used taxonomies, such as the taxonomies by Groves in Wilson and Reeder (2005) and Corbet and Hill (1991). The latter is currently only available for primates and perissodactyles. To change the taxonomy, simply click on the appropriate link. Note, however, that changing the taxonomy will reset the current selection of species! Thus, select the taxonomy first, and then select the species that should be included in the trees.

The total number of species in the different taxonomies may be different because two or more distinct species from the GenBank (GB) taxonomy may translate into the same species in the Corbet and Hill (CH) or Wilson and Reeder (WR) taxonomy. This issue is mainly relevant for primates (particularly for the Corbet and Hill translation), but to a smaller degree also for the Carnivora and Cetartiodactyla part of the website. In such a case, the species with the most available sequence data available is selected and other species that translate into the same name are deleted from the list of taxa on the *10kTrees Website* and also automatically pruned from all subsequent trees.

If the WR or CH taxonomy is selected, some species may be displayed in gray followed by a " ? " after the species name. For species highlighted in gray, we did not find a direct translation into the selected taxonomy. This was particularly a problem with the CH taxonomy.

While there are a few newly discovered species (e.g. *Rungwecebus kipunji*) most new names are examples of taxonomic revision. It quickly becomes quite subjective to decide what Corbet and Hill would have called a species name in GenBank. Thus, we prefer to leave this level of judgment up to the user of our site and do not automatically prune the species from the trees. We instead deselect them by default when the user first selects the taxonomy, but provide the option for the user to reselect them if desired.

To summarize, differences in the three available taxonomies are due to the following reasons:

a) Alternative taxonomies may have less extensive documentation of synonyms than the GB taxonomy (e.g., CH).
b) Alternative taxonomies may recognize fewer subspecies than the GB taxonomy (e.g., WR and especially CH)
c) Due to taxonomic revisions, some species have recently been recognized as full species, renamed or were discovered after the latest release of alternative taxonomies (e.g., *Rungwecebus kipunji*).

3. <u>Specifying the number of trees</u>

For downloading tress, users have the following three choices:

a) Download a consensus tree
b) Download a tree block
c) Download consensus tree and tree block

If you select b) or c), users can specify how many trees from the tree block they want to download. All trees are sampled evenly across the whole tree block; thus, if users entered the number 10, trees are sampled every 1,000 phylogenies from the full tree block, starting from the first tree, rather than simply taking the first ten trees. Users must enter valid numbers between 10 and 10,000.

4. <u>Selecting if the trees should be dated</u>

In addition to providing the trees with branch lengths proportional to genetic change (phylogram), we provide dated trees (branches that reflect the time since two species last shared a common ancestor) based on fossil calibration points (chronogram) (see section 2.1.5). Note

that dated trees are, by definition, ultrametric (except when extinct species are included in the trees, as in Version 3 of *10kTrees* Primates).

5. Selecting the species that should be included in the trees

Lastly, users can select the species to include in the trees. Species that are not selected will be pruned from all trees. For enhanced usability, we provide two different options how species can be listed:

a) We organized species into major clades and provide the possibility to select / deselect all species in a clade at once. If species are listed alphabetically (see below), you may click "List species organized taxonomically" to change the organization.
b) If species are listed taxonomically (see above), you may click "List species in alphabetical order" to change the displaying.

In both cases, we also provide common names for the species for researchers who are not familiar with the Latin names.

Load previously selected set of species

If a user plans to use the same set of species multiple times, we implemented a feature that saves the current set of selected species into a file that can be downloaded onto your computer. The file is stored encrypted, and users must not modify the file after the download (or the server will not accept the file). To restore a previous set of selected species, click the "*Load previously selected set of species"* link and select the file that you previously downloaded. The correct version of the dataset will also be restored after you uploaded the file to the server, even if a newer version became available after you first downloaded the file. Simply follow the instructions on the website.

Load default species for this taxonomy

By clicking on the *"Load default species for this taxonomy"* link, you can load the default set of species for the selected taxonomy. That is, only the species for which we found a direct translation for the GenBank name into the selected taxonomy are selected (see *"Selecting a taxonomy"* above for more information).

6.5. Archive

We also provide an archive with all previous versions of the dataset (if any exist), for which users will still be able to download the same set of files that is provided for the latest version. Currently, Version 1 and Version 2 of the dataset are available in the archive on the Primates part of the website.

6.6. Feedback system and mailing list

Here, you can subscribe to the *10kTrees* mailing list. For more details, go to the website and click the "Feedback / Mailing List" link.

Furthermore, we established a feedback system. It is easy and quick, and vital for the continuous success of this site. You can use the feedback system to provide feedback of any kind (for example, if you are missing a species in the trees for which you have comparative data). We value all the feedback that we receive and will try to reply in a timely manner.

7. Importing the Trees into other Programs

For both tree block and consensus tree, we provide files in the NEXUS format. The following phylogenetic programs have been tested and can read the files produced by the *10kTrees Website* without errors:

1. Mesquite (Maddison and Maddison 2006)
2. FigTree (available at http://tree.bio.ed.ac.uk/software/figtree/)
3. BayesTraits (Pagel and Meade 2007)
4. R (R Development Core Team 2008)

Other phylogenetic program that can read NEXUS files will most likely also be able to read the files produced by the *10kTrees Website*; however, we cannot guarantee full compatibility, and users may in some cases have to alter the text files. If you encounter any problems with other programs, feel free to contact me, Christian Arnold, and I will be happy to work with you on a solution.

To use the trees in the program R, you may use the following code:

```
#make sure you installed the APE library -> install.packages("ape")
library(ape)
#read trees from downloaded file
treeBlock <- read.nexus("TreeBlock_10kTrees.nex")
#extract individual trees
tree_1 <- treeBlock[[1]] #IMPORTANT: NOT [1], as treeBlock is a list
#examine internal structure of object
str(tree_1)
#edge lengths of first tree
tree_1$edge.length
```

For more details on how to make use of this resource in terms of downloading a bunch of trees, viewing them, and modifying them, see the "How To Use" section of the website.

8. Upcoming and Recently Added Features

The *10kTrees Website* is a work in progress, and we will implement additional features in the near future that provide more tools for primate comparative biology. We are currently discussing what features we want to add in the near future.

The *10kTrees Website* now also contains different sections that correspond to different mammalian orders for which we provide trees. We already finished producing *10kTrees* Version 1 for odd-toed ungulates (order *Perissodactyla*), *10kTrees* Version 1 for carnivorans (order *Carnivora*), and *10kTrees* Version 1 for even-toed ungulates and cetaceans (clade Cetartiodactyla). Thus, currently, four sections are available on the website: *10kTrees* Primates, *10kTrees* Perissodactyla, *10kTrees* Carnivora, and *10kTrees* Cetartiodactyla. We may provide Bayesian tree blocks for additional mammalian orders in the near future (let us know which groups interest you!)

For the 1) carnivores, 2) artiodactyles and cetaceans and 3) primates Version 3 trees, we used MrBayes 3.2, which is a substantial improvement as compared to MrBayes 3.1.2. For example, MrBayes 3.2 implements sampling across the entire time-reversible substitution model space as an alternative to a priori model testing (RJ-MCMC), new tree moves that improve convergence, automatic tuning of proposal tuning parameters, a wider range of convergence diagnostics, richer summaries of tree samples (see the consensus trees in the respective Dataset sections).

With Version 2 for the Primates part of the website, we added some of the features that we announced with Version 1, such as a larger and more complete dataset, a taxonomic translation tool, and the possibility to download dates trees based on fossil calibration points. With Version 3, we added the possibility to download a consensus tree that also contains clade credibility values. The user can choose if he or she wants to download the consensus tree with or without (as before) clade credibility values. Version 3 now also includes two extinct species (*Homo sapiens neanderthalensis* and *Archaeolemur majori*). For example, inclusion of these extinct species may be useful for comparative tests based on morphology for which the data would be also available for the extinct species.

9. References

Arnold, C. 2011. The FAST pipeline: A bioinformatics pipeline for automated re-trieval, processing, and dataset construction for sequence data to infer phylogenetic trees. in prep.

Benefit, B. R., and M. L. McCrossin. 2002. The Victoriapithecidae, Cercopithecoidea, Pages 241-253 *in* W. C. Hartwig, ed. The Primate Fossil Record. Cambridge, Cambridge University Press.

Brunet, M., F. Guy, D. Pilbeam, H. Mackaye, A. Likius, D. Ahounta, A. Beauvilain et al. 2002. A new hominid from the Upper Miocene of Chad, Central Africa. Nature 418:145-151.

Castresana, J. 2002. GBLOCKS: selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, Version 0.91 b. Copyrighted by J. Castresana, EMBL.

Corbet, G. B., and J. E. Hill. 1991, A world list of mammalian species. Oxford, Oxford University Press.

Disotell, T. R. 2008. Primate Phylogenetics Encyclopedia of Life Sciences. Chinchester, John Wiley and Sons, Ltd.

Godinot, M. 2006. Lemuriform origins as viewed from the fossil record. Folia Primatologica 77:446-464.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52:696-704.

Haile-Selassie, Y. 2001. Late Miocene hominids from the middle Awash, Ethiopia. Nature 412:178-181.

Hartwig, W. C., and D. J. Meldrum. 2002. Miocene platyrrhines of the northern Neotropics, Pages 175-188 *in* W. Hartwig, ed. The Primate Fossil Record. Cambridge, Cambridge University Press.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22:160-174.

Hodgson, J. A., K. N. Sterner, L. J. Matthews, A. S. Burrell, R. A. Jani, R. L. Raaum, C. B. Stewart et al. 2009. Successive radiations, not stasis, in the South American primate

fauna. Proceedings of the National Academy of Sciences of the United States of America 106:5534-5539.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-2314.

Janecka, J. E., W. Miller, T. H. Pringle, F. Wiens, A. Zitzmann, K. M. Helgen, M. S. Springer et al. 2007. Molecular and genomic data identify the closest living relative of primates. Science 318:792-794.

Kelley, J. 2002. The hominoid radiation in Asia, Pages 369-384 *in* W. C. Hartwig, ed. The Primate Fossil Record. Cambridge, Cambridge University Press.

Kjer, K. M. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. Molecular Phylogenetics and Evolution 4:314-330.

Leakey, M. G. 1993. Evolution of Theropithecus in the Turkana Basin, Pages 85-123 *in* N. G. Jablonski, ed. Theropithecus: The rise and Fall of a Primate Genus. Cambridge, Cambridge University Press.

Lutzoni, F., M. Pagel, and V. Reeb. 2001. Major fungal lineages are derived from lichen symbiotic ancestors. Nature 411:937-940.

Maddison, W. P., and D. R. Maddison. 2006.Mesquite: a modular system for evolutionary analysis, version 2.5.http://mesquiteproject.org.

Marshall, D. 2009. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. Systematic Biology.

Mittermeier, R. A., I. Tattersall, W. R. Konstant, R. B. Mast, F. Hawkins, and D. M. Meyers. 1994. Chapter 4: The Extinct Lemurs, Pages 33–48 *in* R. A. Mittermeier, I. Tattersall, W. R. Konstant, R. B. Mast, F. Hawkins, and D. M. Meyers, eds. Lemurs of Madagascar, Conservation International.

Morrison, D. A., and J. T. Ellis. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. Molecular Biology and Evolution 14:428-441.

Nylander, J., J. Wilgenbusch, D. Warren, and D. Swofford. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581.

Ogden, T. H., and M. S. Rosenberg. 2006. Multiple sequence alignment accuracy and phylogenetic inference. Systematic Biology 55:314-328.

Pagel, M., and F. Lutzoni. 2002. Accounting for phylogenetic uncertainty in comparative studies of evolution and adaptation, Pages 148-161 *in* M. Lässig, and A. Valleriani, eds. Biological Evolution and Statistical Physics. Berlin, Springer-Verlag.

Pagel, M., and A. Meade. 2007.BayesTraits (www.evolution.rdg.ac.uk), version 1.0, Reading, UK.

Posada, D. 2008. jModelTest: Phylogenetic Model Averaging. Molecular Biology and Evolution 25:1253-1256.

R Development Core Team. 2008.R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ray, D. A., and M. A. Batzer. 2005. Tracking Alu evolution in New World primates. BMC Evolutionary Biology 5:51.

Ray, D. A., J. C. Xing, D. J. Hedges, M. A. Hall, M. E. Laborde, B. A. Anders, B. R. White et al. 2005. Alu insertion loci and platyrrhine primate phylogeny. Molecular Phylogenetics and Evolution 35:117-126.

Rodriguez, F., J. Oliver, A. Marin, and J. Medina. 1990. The general stochastic model of nucleotide substitution. Journal Theoretical Biology 142:485–501.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Roos, C., J. Schmitz, and H. Zischler. 2004. Primate jumping genes elucidate strepsirrhine phylogeny. Proceedings of the National Academy of Sciences 101:10650-10654.

Salem, A. H., D. A. Ray, J. Xing, P. A. Callinan, J. S. Myers, D. J. Hedges, R. K. Garber et al. 2003. Alu elements and hominid phylogenetics. Proceedings of the National Academy of Sciences of the United States of America 100:12787-12791.

Sanderson, M., D. Boss, D. Chen, K. Cranston, and A. Wehe. 2008. The PhyLoTA Browser: processing GenBank for molecular phylogenetics research. Systematic Biology 57:335-346.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Molecular Biology and Evolution 19:101-109.

Schmitz, J., M. Ohme, and H. Zischler. 2001. SINE insertions in cladistic analyses and the phylogenetic affiliations of Tarsius bancanus to other primates. Genetics 157:777-784.

Seiffert, E. R., E. L. Simons, and Y. Attia. 2003. Fossil evidence for an ancient divergence of lorises and galagos. Nature 422:421-424.

Senut, B., M. Pickford, D. Gommery, P. Mein, K. Cheboi, and Y. Coppens. 2001. First hominid from the Miocene (Lukeino formation, Kenya). Comptes Rendus de l'Academie des Sciences Series IIA Earth and Planetary Science 332:137-144.

Smythe, A. B., M. J. Sanderson, and S. A. Nadler. 2006. Nematode small subunit phylogeny correlates with alignment parameters. Systematic Biology 55:972-992.

Symonds, M. R. E. 2002. The effects of topological inaccuracy in evolutionary trees on the phylogenetic comparative method of independent contrasts. Systematic Biology 51:541-553.

Talavera, G., and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology 56:564-577.

Tao, N., R. Richardson, W. Bruno, and C. Kuiken. 2005.FindModel (http://hcv.lanl.gov/content/hcv–db/findmodel/findmodel.html).

Vignaud, P., P. Duringer, H. Mackaye, A. Likius, C. Blondel, J. Boisserie, L. De Bonis et al. 2002. Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. Nature 418:152-155.

Wilson, D. E., and D. M. Reeder. 2005, Mammal Species of the World, Johns Hopkins University Press.

Xing, J., H. Wang, K. D. Han, D. A. Ray, C. H. Huang, L. G. Chemnick, C. B. Stewart et al. 2005. A mobile element based phylogeny of Old World monkeys. Molecular Phylogenetics and Evolution 37:872-880.

Xing, J. C., D. J. Witherspoon, D. A. Ray, M. A. Batzer, and L. B. Jorde. 2007. Mobile DNA elements in primate and human evolution. American Journal of Physical Anthropology:2-19.

Yang, Z. H., and A. D. Yoder. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Systematic Biology 52:705-716.

Yoder, A. D., and Z. H. Yang. 2004. Divergence dates for Malagasy lemurs estimated from multiple gene loci: geological and evolutionary context. Molecular Ecology 13:757-773.

Young, N., and L. MacLatchy. 2004. The phylogenetic position of Morotopithecus. Journal of Human Evolution 46:163-184.

Zharkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. Journal Molecular Evolution 39:315–329.