The 10kTrees Website: A New Online Resource for Primate Phylogeny

The Harvard community has made this article openly available. *Please share* how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1002/evan.20251</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:4318029">http://nrs.harvard.edu/urn-3:HUL.InstRepos:4318029</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP</a></td>
</tr>
</tbody>
</table>
The 10kTrees Website: A New Online Resource for Primate Phylogeny

Christian Arnold\textsuperscript{1,2}, Luke J. Matthews\textsuperscript{1}, and Charles L. Nunn\textsuperscript{1} *

\textsuperscript{1} Harvard University, Department of Human Evolutionary Biology, Peabody Museum, 11 Divinity Avenue, Cambridge MA 02138

\textsuperscript{2} Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany

* to whom correspondence should be addressed (cnunn@oeb.harvard.edu, tel. 617-495-4710, fax 617-496-8041)

17 pages, 1 figure, 0 tables

Key words: primate phylogeny, comparative method, Bayesian phylogenetics, phylogenetic uncertainty
About the authors

Christian Arnold received his Master’s in Bioinformatics in June 2008. He then worked for one year at Harvard University in the Department of Human Evolutionary Biology. Since October 2009, he is a Ph.D. candidate at University of Leipzig, Germany. His bioinformatics research focuses on epigenetics, in particular non-coding RNAs and chromatin. Generally, he is interested in finding creative solutions for a wide variety of bioinformatic-related problems in phylogenetics, phylogenetic comparative methods, and epigenetics.

Luke Matthews is an anthropologist and primatologist whose interests include the evolution of primate behavior, primate traditions, and human cultural evolution. He has conducted primatological fieldwork in Ecuador and Argentina. He employs phylogenetics, network analysis, and cluster analysis to study inheritance systems that range from DNA sequences to primate foraging traditions and human material culture. Luke received his Ph.D. from New York University in Physical Anthropology in 2008. Since September 2008, he has been a Postdoctoral Fellow at Harvard University in Charles Nunn’s laboratory.

Charles Nunn is an evolutionary anthropologist with interests in primate disease ecology, behavior, and cultural evolution. His research uses phylogenetic methods and mathematical modeling, and he has conducted fieldwork in Costa Rica and Madagascar. He is the author of Infectious Diseases of Primates: Behavior, Ecology and Evolution.
(with Sonia Altizer) and a forthcoming book on *Phylogenetic Comparative Methods in Evolutionary Anthropology and Biology*. Charlie received his Ph.D. from Duke University in 1999 and has held positions at the University of California Berkeley and Davis, the Max Planck Institute, and University of Virginia. Currently, he is an Associate Professor at Harvard University.
The comparative method plays a central role in efforts to uncover the adaptive basis for primate behaviors, morphological traits and cognitive abilities. The comparative method has been used, for example, to infer that living in a larger group selects for a larger neocortex, that primate territoriality favors a longer day range relative to home range size, and that sperm competition can account for the evolution of primate testes size. Comparison is fundamental for reconstructing behavioral traits in the fossil record, for example in studies of locomotion and diet. Recent advances in comparative methods require phylogenetic information, but our knowledge of phylogenetic information is imperfect. In the face of uncertainty about evolutionary relationships, which phylogeny should one use? Here we provide a new resource for comparative studies of primates that enables users to run comparative analyses on multiple primate phylogenies. Importantly, the 10,000 trees that we provide are not random, but instead use recent systematic methods to create a plausible set of topologies that reflect our certainty about some nodes on the tree and uncertainty about other nodes given the dataset. The trees also reflect uncertainty about branch lengths. The comparative method has undergone a revolution in the past 20 years. Specifically, new phylogenetic methods provide a way to incorporate evolutionary history directly into comparative research. Phylogeny is essential to comparative research because related species tend to resemble one another, resulting in non-independent data points. Phylogenetic comparative methods can be used to investigate whether two traits change in tandem through time, while also providing the historical scaffolding to identify independent evolutionary origins of the traits of interest. More recently,
phylogenetic methods have provided a toolkit to investigate the tempo and mode of evolution,¹⁹-²⁰ to quantify phylogenetic signal in comparative data,²¹-²² and to study the factors that influence diversification rates.²³-²⁴ Computer simulations have revealed that it is usually preferable to conduct comparative tests with some form of phylogenetic method because this reduces false positives (Type I errors) and increases statistical power.¹⁷-¹⁸,²⁵-²⁶ This latter point is often under-appreciated, but it is a logical outcome of phylogenetic comparative analyses that reduce error associated with the estimation of statistics and thus enhance the probability of detecting real effects.²⁶

Researchers generally want to include as many species as possible in a comparative analysis. To incorporate phylogeny in comparative studies of primates, previous researchers have used either published primate-wide “supertrees” such as the Purvis phylogeny,²⁷ or they compiled smaller trees from the literature, often patching these together from among existing phylogenies based on morphology or genetics.²⁸-²⁹ More recently, Bininda-Emonds et al.³⁰-³¹ produced a new supertree of mammals, and researchers have begun to use the primate portion of this tree in comparative studies of primates.³²-³⁴

The actual tree topology and timing of speciation events is, however, never known with certainty. In addition, phylogenetic relationships should be continually reassessed as new data become available, which recommends against the continued use of older phylogenies such as Purvis’,²⁷ 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. Indeed, different trees can produce different results in a comparative analysis, and thus it is unwise to condition comparative
analyses on a single hypothesis of evolutionary relationships when that hypothesis is
legitimately uncertain.\textsuperscript{35}

Some evolutionary anthropologists have accounted for phylogenetic uncertainty
by conducting multiple analyses using more than one tree.\textsuperscript{28,36-38} But this raises an
important question: how should we systematically decide on the trees to use? A number
of authors have proposed that Bayesian phylogenetic approaches provide a way to
systematically incorporate phylogenetic uncertainty into comparative research.\textsuperscript{35,39-40} In
particular, Bayesian methods allow the user to obtain a set of trees that are sampled in
proportion to their posterior probability (see Box 1). The set of trees obtained reflects
uncertainty in the phylogeny given the substitution model and data; more certain nodes
are found across a greater proportion of the sample of trees, while less certain nodes are
found less often. The user can create as many trees as he or she wishes – hundreds, even
thousands of phylogenies, all fully bifurcating and with branch lengths, and not simply
random permutations of the species in the study.\textsuperscript{41-42} It is possible to then run
comparative analyses on this sample of trees, and in this way the results of a comparative
study are not conditioned on a particular phylogeny or set of branch lengths.\textsuperscript{40}

Here we describe a new online resource for comparative studies of primates,
which we call \textit{10kTrees} and make available at http://10kTrees.fas.harvard.edu. The
\textit{10kTrees} website provides a way for users to download up to 10,000 primate phylogenies
with branch lengths obtained from a Bayesian phylogenetic analysis. The trees in Version
1 include 189 primate species that are commonly used in comparative research on
primates. The data are provided in a standard format\textsuperscript{43} that can be read by a wide variety
of comparative methods programs,\textsuperscript{44-45} including phylogenetics packages for \textit{R}.\textsuperscript{46} We
request that people using this resource cite two papers: the current paper, which describes the 10kTrees website and Version 1 of the trees, and another paper that applies Version 1 to study primate diversification in relation to body mass.\textsuperscript{47}

**A Bayesian inference of primate phylogeny.** Bayesian phylogenetic methods provide a way to sample a set of trees in proportion to their posterior probabilities using Metropolis-coupled Markov chain Monte Carlo (MCMC, or MC\textsuperscript{3}) algorithms (see Box 1).\textsuperscript{39,48-50} Nodes that the data strongly support are identical or nearly so across most of the “tree block” (i.e., the sample of trees obtained from the analysis). On the consensus tree of the 10kTrees block, these nodes are thus depicted with high clade credibility values. Some nodes are not well supported, which indicates that alternative arrangements produce similar likelihoods, and these nodes vary across the tree block in proportion to their posterior probabilities. **By running comparative analyses across this set of trees rather than using a single tree, the results are no longer conditioned on a single tree being correct.**

We expect this resource to be especially important for primate phylogenetic comparative studies because it provides a statistically rigorous and principled way to control for uncertainty at various nodes in primate phylogeny.\textsuperscript{51-55} Importantly, our goal for this project is not to produce the definitive primate phylogeny; that goal will be best achieved with more focused studies of gene insertions, whole genomes, and standardized data collection, and will involve longer-term concerted effort by experts in primate phylogenetics (e.g., ref. 56). Instead, **the goal is to produce a set of phylogenetic trees from available data that is appropriate for comparative research on primates.** We will,
however, regularly update the dataset to accommodate ever-increasing availability of sequence data and advances in tree inference methods.

For Version 1, we collected data on four mitochondrial genes and one autosomal gene from GenBank. To create the multiple sequence alignments (MSA), we used Muscle 3.7 with the default parameters. Because alignment quality can have a substantial impact on the inferred tree, we manually excluded poorly aligned sites or sites with a high percentage of missing data (especially at the beginning and end of the MSA). We constrained 29 major nodes if they were well characterized by at least three genomic Alu insertions. These constraints eliminate uncertainty at constrained nodes, which we think is reasonable because Alu insertion events are generally regarded as more reliable cladistic indicators that are less prone to homoplasy than DNA sequence data. However, we are likely to relax these constraints in future versions of 10kTrees that use additional autosomal loci (and thus users should refer to the website for details appropriate for the version they use).

For tree inference in Version 1, we used the program MrBayes 3.1.2. Galeopterus variegatus (Sunda flying lemur or colugo) was identified as the outgroup, as it has been shown that colugos are the closest living relatives to the order Primates. We ran a Bayesian analysis with two runs and eight chains (one cold chain and seven heated chains) in each run. We used a GTR+I+G substitution model for each of the five genes in a partitioned dataset, which was identified as the best substitution model in the program FindModel. The analysis for Version 1 was run for eight million generations, with trees sampled every 1,000 generations. We assessed the heating (changed to 0.02) and excluded the first three million generations as burn-in (see Box 1). We summarized these
topologies by constructing a 50% majority rule consensus tree, which we provide in various graphical formats 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. In future versions of 10kTrees, we will provide both molecular branch lengths and, by using fossil calibration points, branches that reflect the time since two species last shared a common ancestor.

**Applications to primate comparative biology.** An important goal of our project is to make the trees readily available for comparative research. To that end, users can download trees in NEXUS format. On the 10kTrees website, users can select the number of trees to download. These are sampled from the tree block such that they cover full range of variation in the analysis (i.e., they are sampled evenly along the stored chain of trees, rather than simply taking the first $n$ trees in the sample, where $n$ is the number of trees requested by the user). The consensus tree of the full sample is also available to download. In addition, the user has the option to select specific species of interest; the trees are then pruned to the selected species prior to download. In terms of the actual data used to generate the trees, users can obtain the original sequence data, the list of species that were studied, an availability matrix for the distribution of genetic data across species, and details on how the data were analyzed (i.e., the substitution model and parameters of the MC$^3$ analysis, such as sampling rate, number of chains, and number of sampled generations). The phylogenetic constraints based on Alu insertions are also downloadable from the website. By providing the raw data files, users can easily rerun the phylogenetic analysis in a different computer package or with different settings, including without the constraints.
The phylogenies available from 10kTrees provide a significant improvement over the two primate-wide phylogenies that are currently most commonly used in comparative research.\textsuperscript{27,30} The older of these phylogenies, published by Andy Purvis in 1995\textsuperscript{27}, was a ground-breaking contribution to comparative primatology when it was published. However, fewer genetic data were available in 1995, and many polytomies were present in the tree. In particular, the Asian colobines showed almost no resolution because phylogenetic information for this clade was generally unavailable when the tree was constructed. In addition to its lack of resolution, the Purvis phylogeny includes a number of topological misplacements even at the generic level as assessed by comparison to the preponderance of DNA sequence and Alu insertion data used to infer the 10kTrees. For example, the basal bifurcation of the platyrrhines is incorrect on the basis of Alu insertions, as are the positions of \textit{Aotus} and \textit{Callicebus}.\textsuperscript{65} Recent sequence-based studies and our research clearly contradict the topology given for other genera such as \textit{Callimico} and \textit{Lophocebus}.\textsuperscript{72-73} Given its lack of resolution and topological misplacements at the generic level, the continued use of the Purvis phylogeny\textsuperscript{27} cannot be recommended.

The primate portion of the more recent Bininda-Emonds supertree\textsuperscript{30} is an improvement over Purvis’ phylogeny,\textsuperscript{27} but it still suffers from excessive polytomies within the guenons and the Asian colobines. The Bininda-Emonds tree contains an incorrect topology for the basal bifurcations among the platyrrhines given the Alu insertion data and whole mitochondrial genome evidence.\textsuperscript{65,72} It also includes apparent anomalies or errors, such as the paraphyletic placement of \textit{Callicebus personatus} as separate from its congeners and basal to Cebidae. While the Bininda-Emonds tree is
useful for mammal-wide comparative analyses, we recommend *10kTrees* for studies focused on primates.

We designed the website so that it can be easily updated as new genetic data become available to infer primate phylogeny. Thus, the website will provide multiple versions, including an “archive” where previous versions can be accessed. We are already working on Version 2 of the dataset, which will include over 230 species and more genes. We also expect that the **website itself will evolve to provide more tools for primate comparative biology**. In future versions, for example, we plan to provide a taxonomic translation tool. Thus, readers will be 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 Corbet and Hill and Groves in Wilson and Reeder. In addition to giving trees with branch lengths proportional to genetic change, a future version of the website will provide dated trees based on fossil calibration points. We will also make available scripts to link the trees to comparative data to produce a single NEXUS file for analysis, and we will provide advice for importing the trees into analysis programs, such as BayesTraits, R and Mesquite. Ultimately, we aim for a flexible, open-access and user-friendly platform that will enhance the use of phylogenetic approaches in primate evolution, and that will grow as new sequence data and methods become available.

**Acknowledgements**

We thank three anonymous reviewers, Richard Kay and John Fleagle for helpful comments. This research was supported by the National Science Foundation (BCS-0923791) and Harvard University.
References


Box 1. Schematic of Bayesian phylogenetics. Bayesian methods in phylogenetics typically use Metropolis-coupled Markov chain Monte Carlo algorithms (MCMC, or MC$^3$) to generate posterior probability distributions for a set of parameters; that is, the conditional distribution of the parameter given the data. Note that in Bayesian statistics, probability cannot be interpreted in its original meaning; rather, it is used to represent parameter uncertainty. In Bayesian phylogenetics, the parameters are comprised of a phylogenetic tree and a specific model of evolution, which is based on the individual priors for these parameters (a priori knowledge or beliefs about a parameter distribution) and the likelihood of the data.

Bayesian MC$^3$ methods start with a random tree and arbitrary initial values for branch lengths and model parameters. In each generation, either a new tree or a new model parameter is proposed. Typically, each proposed change is small; thus, the samples are not random because they are based on the parameter value of the previous generation. The proposal can then be accepted or rejected, depending on the ratio of the posterior densities of the new state to the old state (R). If R>1 (an “uphill” step, indicating a state with higher posterior probability) the proposed change is always accepted. If R<1 (a “downhill” step), the change is accepted with probability R. The smaller R, the smaller is the proposal acceptance probability. Thus, after a particular number of generations, the region of the parameter space with the highest posterior probability is reached and sampled most often. If the parameter space has multiple peaks that are separated by deep valleys, however, the algorithm may become stuck on a local peak. To rectify that
problem, additional chains (so-called hot or heated chains) are used in MC$^3$ that independently search the tree space, more readily accept proposals with a small $R$ value, and regularly swap states with the original chain (cold chain). Heated chains flatten the posterior probability distribution and thus have shallower valleys, which allows them to more easily cross those valleys.

During MC$^3$, a “chain” of trees is produced that reflects the accepted modifications in phylogenetic information and model parameters. At the beginning of the chain, the likelihood typically climbs quickly (which is called burn-in, see Figure) until the chain eventually reaches its equilibrium distribution (indicated by the putative plateau in the Figure). Every $k$ generations, the cold chain is sampled, which simply entails saving the tree topology, branch lengths and all model parameters, and further analyses generally make use of post-burn-in samples. If the posterior probability distribution for each parameter has been approximated adequately, their post-burn-in sampling reflects the true parameter uncertainty, and the quality of the sample generally improves as a function of the number of steps (generations). The sample can then be summarized using various statistics, such as histograms, means, or credible intervals. Topology and branch lengths can also be summarized by constructing a majority rule consensus tree with support values in the form of clade credibility values and mean branch lengths. One can also use the whole sample of trees (“tree block”) to incorporate topological and branch length uncertainty into comparative analyses that make use of the phylogenetic information. This is the approach that we advocate in this article and is also recommended by others.$^{39-40}$