Molecular Phylogenetics and Evolution 84 (2015) 14–26

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A species-level phylogeny of all extant and late Quaternary extinct mammals using a novel heuristic-hierarchical Bayesian approach

Søren Faurby *, Jens-Christian Svenning

Section for Ecoinformatics & Biodiversity, Department of Bioscience, Aarhus University, Ny Munkegade 114, DK-8000 Aarhus C, Denmark

ARTICLE INFO

Article history:
Received 12 March 2014
Revised 30 October 2014
Accepted 2 November 2014
Available online 24 December 2014

Keywords:
Mammalia
Phylogenetic asymmetry
Species-level phylogeny
Supermatrix
Supertree

ABSTRACT

Across large clades, two problems are generally encountered in the estimation of species-level phylogenies: (a) the number of taxa involved is generally so high that computation-intensive approaches cannot readily be utilized and (b) even for clades that have received intense study (e.g., mammals), attention has been centered on relatively few selected species, and most taxa must therefore be positioned on the basis of very limited genetic data. Here, we describe a new heuristic-hierarchical Bayesian approach and use it to construct a species-level phylogeny for all extant and late Quaternary extinct mammals. In this approach, species with large quantities of genetic data are placed nearly freely in the mammalian phylogeny according to these data, whereas the placement of species with lower quantities of data is performed with steadily stricter restrictions for decreasing data quantities. The advantages of the proposed method include (a) an improved ability to incorporate phylogenetic uncertainty in downstream analyses based on the resulting phylogeny, (b) a reduced potential for long-branch attraction or other types of errors that place low-data taxa far from their true position, while maintaining minimal restrictions for better-studied taxa, and (c) likely improved placement of low-data taxa due to the use of closer outgroups.

1. Introduction

Many macroevolutionary, biogeographical, and ecological studies require phylogenetic information (e.g., Faurby et al., 2012; Kissling et al., 2012; Eiserhardt et al., 2013). However, although a few analytical methods can incorporate missing species under certain restrictions when analyzing phylogenetic patterns (e.g., Fitzjohn et al., 2009), most methods assume that all species are included in the phylogeny and often also assume that the phylogeny is without error. This constitutes a major drawback, as neither of these assumptions is generally justifiable.

These problems are well exemplified by the mammalian clade, which is the focus of this study, and for which a fairly recent, somewhat complete, and often-used species-level supertree is available (Bininda-Emonds et al., 2007, hereafter the BE tree). However, a recent phylogenetic study of mammals (Meredith et al., 2011) noted that the BE tree is incorrect for many family-level nodes. Such errors are expected for nodes with limited data, but family-level nodes have generally received more attention than lower-level nodes, and the latter are therefore likely to be even less reliable. Thus, for phylogenetic information to be meaningfully incorporated into evolutionary and ecological studies it is vital to construct trees using methods that (a) are better at resolving the topology, even with little genetic information, and (b) provide a meaningful way of incorporating the varying uncertainty associated with different nodes into the tree. Here, we describe a heuristic-hierarchical Bayesian approach that satisfies both requirements.

Nearly all tests based on a phylogeny are really testing the probability of a given hypothesis assuming a given tree (i.e., P(A |Most likely Tree), where P(A) is the probability of any hypothesis of interest). This probability is assumed to be identical to the probability given the true phylogeny (i.e., P(A |True phylogeny)). However, the probability that the most likely tree is identical to the true phylogeny is exceedingly small (even 100 bifurcating branches with a posterior probability of 0.9 gives $0.9^{100} = 2.7 \times 10^{-5}$ for the topology to be identical, assuming that each node is considered to be independent). Hence, it is an extreme stretch to assume the two phylogenies to be identical or even strongly related, and the incorporation of phylogenetic uncertainty is therefore vital for such *P*-values to have any real meaning. This uncertainty can be incorporated by analyzing a large number of trees from the posterior distribution through a Bayesian analysis instead of the single highest probability tree (e.g., Huelsenbeck et al., 2000; Faurby et al., 2012). This approach will likely provide superior

* Corresponding author.
E-mail address: soren.faurby@bios.au.dk (S. Faurby).

http://dx.doi.org/10.1016/j.ympev.2014.11.001
1055-7903/© 2014 Elsevier Inc. All rights reserved.

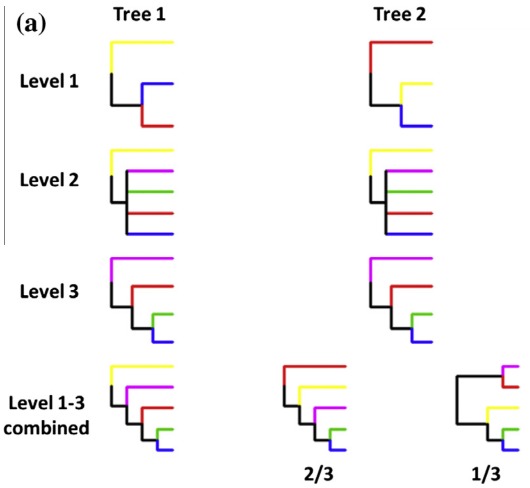


Fig. 1a. An illustration of the merging of phylogenies. In this figure source illustrates the merging of two trees from posterior distributions from two levels (1 and 3) while incorporating taxonomy (level 2). The color codes indicate different species. In tree 1, there is no conflict, while tree 2 shows conflict. The green species is sister to the blue species and can readily be added, but the purple species is sister to the clade of the red, green and blue species, which is not monophyletic in level 1. Because the magenta species is not included in level 1, its placement is restricted based on level 2, and it therefore has to be sister to a clade of one or more of species indicated in red, blue and green. The clade of the blue and green species has two times as many species as the clade comprised on the red species and placement as sister to the blue and green species is therefore twice as likely.

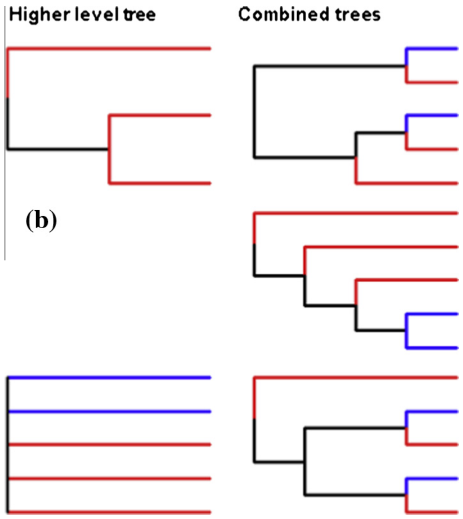


Fig. 1b. An illustration of the merging of phylogenies. In this figure source shows a resolved higher-level tree and a lower level polytomy where the species sampled in the higher-level phylogeny is indicated in red, and the unsampled species is indicated in blue. Only some of the potential trees are shown.

unbiased estimates of P(A) for any measure where P(A| Most likely Tree) is an unbiased estimate of P(A|True phylogeny) but may increase biases if such biases are already present when the most likely tree is used because this tree will be closer to the true phylogeny than an average tree from the posterior distribution (see for instance Symonds, 2002 for a discussion of potential effects of phylogenetic errors).

Another problem is that Bayesian analyses require convergence to the global optimum. Such convergence may not be possible to obtain for trees with thousands of taxa, but our approach uses a heuristic-hierarchical breakdown with phylogenetic analyses of subsets of clades with fewer species, which greatly improves the probability of reaching the global optimum.

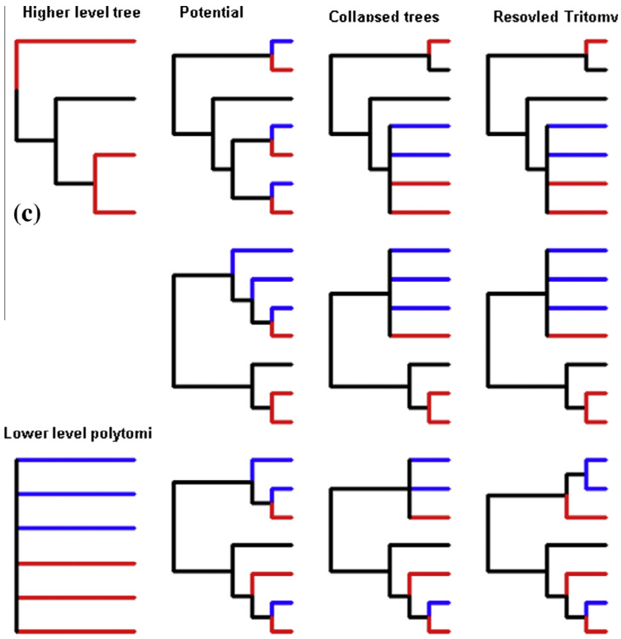


Fig. 1c. An illustration of the merging of phylogenies. In this figure source shows a resolved higher-level tree and a lower-level polytomy where the species sampled in the higher-level phylogeny are indicated in red and the unsampled species are indicated in blue. Only three of the potential outcomes when the two unsampled trees are added are shown. Note that in this example, the clade that is a polytomy in the lower-level clade is not monophyletic in the higher-level tree, and the black species represents a species nested in the clade in the higher-level tree. Because at least half of the species in the polytomy are missing genetic data (the ones indicated in blue), we collapse any clade for which at least half of the species lack genetic data. Each of the top and middle examples contains a clade with more than three species of which at least half do not have any genetic data. The lower example contains a clade with three species of which only one has genetic data. This is collapsed and then randomly resolved.

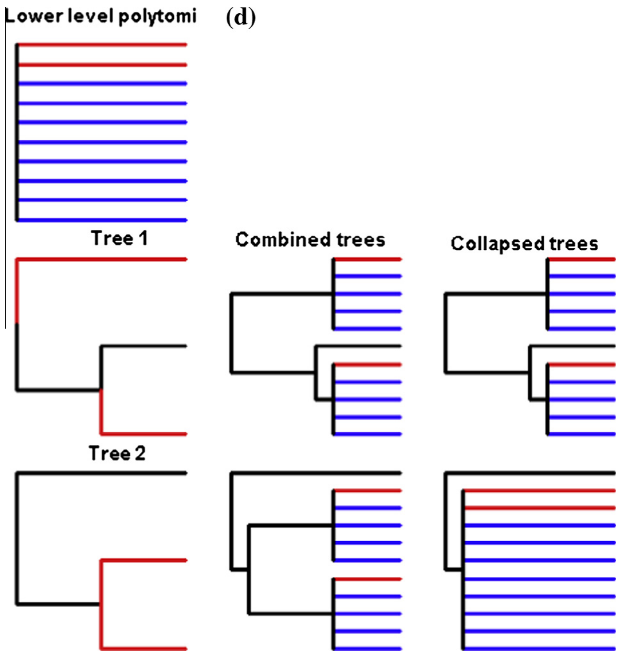


Fig. 1d. An illustration of the merging of phylogenies. In this figure source shows a resolved higher-level tree and a lower level polytomy where the species sampled in the higher-level phylogeny is indicated in red, and the unsampled species is indicated in blue. Because the polytomy is for at least 10 species, only two of which have genetic information, half of the species are placed as sisters to each of the sampled species. Because the combined tree 2 contains a clade where at least half of the species lack genetic information, this clade is collapsed.

Our heuristic-hierarchical approach increases topological precision by using morphological or taxonomic information to constrain the placement of species for which little or no genetic data exist. Genetic data do not exist in a vacuum because taxonomic research can be used to guide the placement of species with limited genetic data. Morphologically based phylogenies are often incorrect, and groupings such as Insectivora have been effectively dismantled by genetic studies (e.g., Meredith et al., 2011), but highly untrustworthy groupings can also arise from limited genetic data (e.g., the "Guinea pig is not a rodent" discussion of the 1990s; D'Erchia et al., 1996). As a trade-off, we suggest the use of taxonomy as a partial topological constraint so that species with limited genetic information are forced to be placed near species in the same taxonomic groupings, while species with more genetic information can be placed more freely. We thereby avoid placing species outside their taxonomically recognized clades, except where this is well supported by substantial genetic data.

We handle the various taxonomy-based constraints by creating phylogenies both across taxonomic groupings for the species with most genetic data and within taxonomic groupings for species with less data. Then, we merge trees from the posterior distributions of each phylogenetic analysis one by one, while in all cases trusting the topology of the least constrained higher-level phylogeny in cases where a difference exists (Fig. 1a, see details in Section 2.1, "The heuristic-hierarchical Bayesian approach").

The most important benefit by this approach is that it enables us to use a closely related outgroup for species with little genetic data because such outgroups generally increase topological precision (Rosenfeld et al., 2012). Another benefit is that it may also help mitigate some of the problems of data quality that arise from using GenBank data. An unknown but likely fairly large fraction of the specimens in GenBank represent misidentifications, contaminations, pseudogenes or chimeras between different sequences (Ashelford et al., 2005). Such errors are sometimes readily identifiable, but pseudogene sequences at almost any distance from the sequences of the desired genes are known (Bertheau et al., 2011). A larger fraction of the sequences for species with just a single marker in GenBank may represent unpublished studies or studies conducted by a non-taxonomist, which may contain a larger proportion of identification errors. We enforce all species that have genetic data for only a single marker as having a congeneric sister and generally use them only for within-genus phylogenies (see the description of the model), thereby greatly reducing the potential errors that result from incorporating such sequences.

The added benefits of topological precision come at the cost of reducing branch length precision. It is fairly easy to combine the topology of two or more trees if one of them is always preferred in cases of disagreement, but branch length information cannot be retained in such cases of disagreement. Imagine a simple example: In the phylogeny of three genera, A, B and C, 70% of the trees are given as ((A1:2, B1:2):2, C1:4) and 30% of the trees as ((A1:1, C1:1):3, B1:4), and the phylogeny of genus A rooted with and dated by the split between A and B is (((A1:0.5, A2:0.5):0.1, A3:1.5):0.5, B1:2). The combined topology would exhibit 70% of the trees as ((((A1, A2),A3), B1), C1) and 30% of the trees as ((((A1, A2),A3), C1), B1). The branch lengths for 70% of the trees are ((((A1:0.5, A2:0.5):0.1, A3:1.5):0.5, B1:2):2, C1:4), but the branch lengths of the genus and family tree cannot be combined in the remaining 30% of the trees. The genus-level phylogeny suggests that the genetic distance between A1 and A3 is 0.75 times the distance between A1 and B1, which would be 3, but the distance between A1 and C1 is only 2, and genus A is assumed to be monophyletic. The branch lengths are therefore unidentifiable in this case. In this particular example, the problem could be addressed by choosing a more distant outgroup in the genus-level phylogeny of A, but that would generally decrease the topological precision.

Another problem is that branch lengths will depend on the number of species in a group when using a tree-prior such as a Yule or Birth–Death prior. Imagine a case with four species, A1, A2, A3 and A4, with available genetic data and one species, A5, without data. In such a scenario, the prior of the length of the branch separating the root node from A2:5 in the topology (A1, (A2, (A3, (A4, A5)))) would be different from the prior of the length of the branch separating the root node from A2:4 in the topology (A1, (A2, (A3, A4))), and the length of this branch would therefore be estimated to differ, even though exactly the same number of mutations would be inferred on the branch in the two scenarios. Because we do not include all species in all trees, our estimates of branch lengths will be biased in regions without all species included, with the pattern being most evident for groups with limited genetic data, where the priors have higher influence.

Because our approach means that branch lengths cannot be estimated in the analysis, we are forced to simulate branch lengths for any branch that is not dated in other studies. This leads to a reduction in branch length information, but as we will discuss in the next section, the magnitude of this reduction may be fairly small, as the amount of information on branch lengths in real datasets will often be limited for most nodes. Nevertheless, because the branch length information and topology cannot both be maximized, difficult choices are required regarding the weight that will be placed on each of these items. Our approach places the greatest weight on the topology, which means that analyses using the resulting phylogeny should focus on the topology rather than on branch lengths.

2. Materials and methods

2.1. The heuristic-hierarchical Bayesian approach

Our approach is based on a heuristic incorporation of additional species into a thereby successively growing phylogeny, where the addition of species into the phylogeny never changes the topology of the species already present in the phylogeny. Our approach starts with the species with the best quantity and quality of data available for their placement, assuming hierarchical trustworthiness of the available information, so that in any case where a placement conflict is observed, the results from the source perceived to be most trustworthy will be used. One built-in assumption of the method is that the more information is available for a given species, the looser the restrictions for the placement of the species will be. Even at the highest level, however, few restrictions are enforced, and taxa are therefore required to belong to overall groups, which has never been questioned. These groups are chosen so that their enforcement is guaranteed to not influence the topology but is used only to provide closer outgroups for the remaining nodes.

Next, several (up to four in our example) levels of separate Bayesian analyses are run, and in each case, 1000 random trees (sampled with replacement) from each posterior distribution are retained. At higher levels, species from a wide range of taxa (in our case, all placental mammals, or all marsupials) with large amounts of data are included. At lower levels, a higher frequency of species from increasingly smaller taxonomic groups (i.e., subclass then order then family then genus) are included, but with fewer markers included to make the dataset more complete, and the lowest level is generally based on only a single marker and includes only species from a single genus. The trees are merged so that each tree from a lower level is merged to one tree from a higher level, where the topology from the higher level is chosen in all cases with conflicts. In cases where only a single representative is present in the higher-level phylogeny, the branch leading to this single representative is simply replaced by the lower-level phylogeny.

When a taxon or clade (T1) is the sister to a group of taxa (T2) in the lower level but T2 is not monophyletic in the higher level, T1 is placed as the sister to one of the monophyletic components of T2 in the combined tree, with the likelihood of each placement being proportional to the number of species in each monophyletic group (Fig. 1a). The main reason for doing this is based on the assumption of taxonomic monophyly whenever it is not in conflict with the higher-order phylogeny. This means that if in the higher-order phylogeny, the genus is in two separate clades, each of which is more closely related to other genera, the merged phylogeny should also have only two such clades. If we instead place T1 as the sister to the most recent common ancestor of T2 and the two species of T2 belong to the two separate clades, there could suddenly be three such clades. Our choice of placement is also related to a belief that the species selected for phylogenetic studies are generally selected to represent the most genetically divergent species in a genus or a phylogeny (further discussed later).

Subsequently, the species without genetic data are added based on morphological trees or taxonomy. Fully resolved parts of these trees are handled in exactly the same way as genetic trees, but polytomies are addressed separately. When genetic data are available for more than half of the taxa in the polytomy, the missing species are placed in random order as sisters to either species with genetic data or already placed species (Fig. 1b). The same procedure is followed when more than half of the species lack genetic data, but in this case, an additional step is added, in which any clade with more than two species where more than half of the species do not have any genetic data is collapsed, and any resulting trichotomies are randomly resolved (Fig. 1c). The collapse and resolution of trichotomies ensures that all topologies for the three species are equally likely. For technical reasons, we chose to address polytomies including at least twelve species where less than half have genetic data in a different way. In such cases, we assumed the same number of missing species to be assigned as sisters to each species with genetic data. Again, any clade in which 50% or more of the species lack genetic data is collapsed (Fig. 1d). We note that this alternative approach was generally used only for large genera without any resolution, and in any case where the genus was monophyletic, it was collapsed into a large polytomy, just as it would have been if the same method was used for polytomies of all sizes.

The missing species are placed as sisters to sampled species rather than at random locations in the clade (i.e., allowing both apical and more basal locations) because taxonomists generally include species to maximize genetic variation. Therefore, incomplete phylogenies are generally less symmetrical than complete ones (Mooers, 1995), and most missing species are, thus, close relatives of sampled species. Hence, by restricting the placement of these species such that they will be sisters to sampled trees, our method will produce complete trees that are more symmetrical than the incomplete trees and will thereby counteract the bias in species selections during phylogenies illustrated by Mooers (1995). We note that this strategy will not by itself guarantee that the apical placement is correct, but similarity in any summary statistic (such as tree symmetry) generally makes similarity of the underlying data more likely. In some situations, an alternative solution could have been to place the species without genetic data in a basal polytomy in the genus (or unsampled genera in a basal polytomy in the family), but that would work only if (a) there is more than one unsampled species (because it would otherwise appear to be a totally basal placement) and (b) there is no additional resolution in the genus or family. This latter problem is, for instance, illustrated by the relationship of the six species of Old World monkeys in the *Cercopithecus cephus* species group. These species are placed in a six-way polytomy by Bininda-Emonds et al. (2007), and we have genetic data for only four of the species. Our approach produces a well-defined placement of the two missing species (*Cercopithecus erythrotis* and *Cercopithecus sclateri*), each as a sister to one of the sampled species (or combined as sisters to one of the sampled species), with the only enforcement of the other four species (*Cercopithecus ascanius*, *Cercopithecus cephus*, *Cercopithecus erythrogaster*, *Cercopithecus petaurista*) being the trivial assumption that they belong to a monophyletic family of Old World monkeys, whereas their placement in a polytomy would require the enforcement of either genus- or species-group monophyly depending on how large one would wish to make the polytomy.

Most phylogenetic studies will include all species for which data are available, but the effort that is put into retrieving data for the remaining species will depend heavily on their perceived importance because much more effort is likely to be put into retrieving an enigmatic potentially basal taxon than one that is thought to be related to already sequenced species. This situation is illustrated by the fact that several recent papers have used the inclusion of such enigmatic species as a major selling point for the importance of their studies (Jansa et al., 2009; Sato et al., 2012). The lack of genetic data can therefore be interpreted as weak evidence that experts on the group believe that the phylogenetic placement of a given species without genetic data is likely to be apical rather than basal. We acknowledge that we only have anecdotal evidence justifying the apical placement of these missing species, but we note that the alternative of making all placements equally likely exhibits exactly the same problems of missing evidence.

We chose to collapse polytomies in which half or more of the species are missing genetic data because the importance of the data in determining the shape of the constructed trees diminishes as more species without genetic data are added. However, we note that the decrease in data content during the determination of tree shape is a continuous phenomenon and that thresholds other than our chosen 50% threshold could have been used as well.

Finally, after the topology is complete, branch lengths are estimated. The divergence dates between higher taxonomic groups and the lengths of the branches separating them are manually incorporated from other sources, while the remaining branch lengths are simulated based on the age of the clade and its assumed evolution according to Yule or BD models (Yule, 1924; Raup, 1985). The appropriate relative extinction rate for each family is found by running the MEDUSA algorithm (Alfaro et al., 2009) on the overall family tree (setting the model parameter to “mixed” and thereby allowing partitions with either a BD or a Yule model) and recovering the estimated extinction rate for each family from this analysis.

Although this strategy means that there will be a reduction in branch length information, this reduction is likely to be fairly small. This can be observed most readily if we look only at the external branches and make a comparison between a hypothetical true phylogeny, branch lengths estimated with our method and branch lengths estimated using a Yule prior, which is frequently employed for both MrBayes and BEAST analyses (Ronquist and Huelsenbeck, 2003; Drummond and Rambaut, 2007). More than 41% of all external branches are estimated without any genetic data (Appendix B) and will therefore necessarily have simulated branch lengths. Approximately 0.6% of the branches are dated from other sources (Table A.1) and their precision will not be affected. Approximately 22% of all external branches are estimated with only a single marker, and their ages are therefore heavily influenced by the priors. With Yule or BD priors, the branch lengths of all (external) branches are heavily dependent on the topology, and if our approach leads to a sufficient increase in topological precision (and the Yule or BD models are an adequate description of the evolution of the group), our simulated branch lengths may actually be closer to the true lengths than branch lengths inferred using Yule or BD priors for some of these species. Any real loss in branch-length precision is therefore restricted to the 36% of the external branches that are analyzed using several markers and are not dated from other sources.

The precision of these procedures could potentially be increased through a two-step heuristic-hierarchical approach where family-level monophyly is enforced and the divergence time between the family and an outgroup is used to date the root of the family tree, similar to the approach employed for the species-level tree for all birds (Jetz et al., 2012), but it is unknown how large the increase in precision would be. It has repeatedly been shown that precise estimation of branch lengths requires multiple dating times at various ages in the phylogeny (e.g., Lukoschek et al., 2011; Meredith et al., 2011), and several additional dating sites in each family would therefore be required to obtain precise branch length information. However, even for the families where sufficient fossil dating points are available, this will reduce the ability to incorporate phylogenetic uncertainty because the constraint of a large number of nodes greatly restricts the part of the tree parameter space that is investigated (for instance, if 10 nodes with a mean posterior probability of 0.90 are constrained, any test of P(A) is actually P(A|B) with $P(B) = 0.90^{10}$–0.34), and reliably incorporating phylogenetic uncertainty therefore quickly becomes problematic.

2.2. Simulations

To test the performance of our method, we generated a large number of sequences based on Rosenfeld et al. (2012) modification of the yeast sequences from the study by Rokas et al. (2003), which is a dataset for which the topology is known to be difficult to infer and for which it has previously been shown that the application of distant outgroups decreases phylogenetic precision (Rosenfeld et al., 2012). However, the topology for the species in question does converge to a single tree as more data are added, which is identical to the tree constructed for the full genomes, and we will assume that this single tree is the true phylogeny. We considered the subtree (((((*Saccharomyces cerevisae*, *Saccharomyces paradoxus*), *Saccharomyces mikatae*), *Saccharomyces kudriavzevii*), *Candida glabrata*), *Candida albicans*) and simulated a scenario where three or five markers were sequenced for the two *Candida* species, *Saccharomyces cerevisae* and one of the other *Saccharomyces* species, whereas only one marker was sequenced for the remaining two *Saccharomyces* species.

Our analysis focused on the probability of generating the true phylogeny for the six-species dataset with three or five markers (two of the species had only one marker), with *Saccharomyces* constrained to be monophyletic, and compared it with the probability of obtaining the true topology of the four *Saccharomyces* species rooted with *Candida glabrata* using only a single marker. This analysis therefore compares the potential gain obtained from using a closer outgroup for the species with smaller amounts of data (in this case, two species of *Saccharomyces*) with the potential gain in information about the branch lengths for the species with more than one marker.

We generated 100 datasets for each of multiple scenarios. The first contained the four *Saccharomyces* species rooted with *Candida glabrata* with all species only sequenced for a single marker. The remaining contained the four species of *Saccharomyces* (*Saccharomyces paradoxus*, *Saccharomyces mikatae* and *Saccharomyces kudriavzevii*), *Candida glabrata* and *Candida albicans*, with the two *Candida* and two of the four *Saccharomyces* species (all six combinations were tried in separate scenarios) sequenced for either three or five markers and the last two species of *Saccharomyces* sequenced for a single marker. For all analyses, we randomly selected 700 base pairs from randomly selected markers (of at least this length). All phylogenetic analyses were performed using the same settings as in our case study on the phylogeny of mammals, which we discuss in the "Genetic Analyses" section. These simulations therefore test the likely most controversial element in our analysis, which is the separation of the genus from the family-level phylogenies for species with limited data. Our analysis includes analyses performed on three separate levels, but the first separation, of inter-family versus intra-family, is much more widespread and is, for instance, very similar to the approach used in a recent species-level phylogeny of all birds (Jetz et al., 2012).

We analyzed the trees through generalized linear modeling (GLM) analyses (Table 1). In this analysis, we compared the precision (the probability of the true topology in the posterior distribution) of the 1-marker scenario (which mimics our heuristic-hierarchical approach) with each of the six alternatives individually as well as with the scenarios grouped together by the number of markers. The comparison showed that the precision is extremely low when using three or five markers under one scenario (when *Saccharomyces cerevisae* and *Saccharomyces mikatae* have several markers) and that the precision may be slightly, albeit non-significantly, higher under the other two scenarios (when *Saccharomyces cerevisae* and *Saccharomyces kudriavzevii* have several markers). When the analyses were grouped by marker, the one-marker scenario outperformed both the three- and the five-marker scenarios, but the differences between three and five markers were very small. We cannot be sure of the underlying reason for the poor performance of some scenarios caused by an apparent tendency of MrBayes to separate the species with large amounts of data as much as possible in our simulation scenarios. Irrespective of the reason that MrBayes had problems generating the true topology, we can state that our method led to a clear improvement in precision under our simulation conditions (i.e., that the gain achieved by using a closer outgroup was much more important than the loss from not using the information for all available markers for some of the species).

2.3. Using the heuristic-hierarchical Bayesian method to construct a phylogeny of mammals

2.3.1. Overall considerations for using the method on mammals

Here, we will use this method to construct the first ever species-level phylogeny for mammals that includes not only all extant species but also all known extinct Holocene and Late Pleistocene species and includes not only a single most likely tree but an estimate of the full tree space. The causes of the Late Pleistocene mammalian extinctions are still not fully understood, but it is becoming increasingly clear that many of the extinctions were at least partially the consequence of human activities (Barnosky et al., 2004; Sandom et al., 2014), and the human causation is even clearer for most Holocene extinctions (Turvey, 2009). We believe that any study interested in natural processes should therefore include all human-caused extinctions if the goal is to understand basic evolutionary patterns. Thus, we believe that the incorporation of such species into our phylogeny will improve the reliability of any downstream analyses using this phylogeny. A few of the Late Pleistocene extinctions may prove to be independent of humans, but the inclusion of all species with confirmed Late Pleistocene records would nevertheless provide a scenario closer to the natural conditions than including only extant or recently extinct species, as is often done (see also Dalerum et al., 2009 for a similar argument for incorporating Pleistocene species).

We attempted to include all extinct species with dated records from the Late Pleistocene (defined as the last 130,000 years) in the

Table 1
Simulations of the effects of data quantities and outgroups on the obtained topology. The precision column lists the posterior probability of the true topology. All significance tests compare the difference between this scenario and the 1-marker scenario (which mimics our heuristic-hierarchical approach). Significance is calculated using Wald tests.

(a) Comparison between the seven scenarios (see the "Simulations" section for details). The standard deviation of the precision estimate is given in parentheses. The presented p-value estimates the likelihood that the precision is identical to the precision for one marker		
Number of markers	Multi-marker species in Saccharomyces	Precision
1	–	0.400 (0.036)
		Difference in precision in relation to 1 marker
3	*S. cerevisae, S. paradoxus*	0.000 (0.048)^{NS}
3	*S. cerevisae, S. mikatae*	−0.384 (0.048)***
3	*S. cerevisae, S. kudriavzevii*	0.059 (0.048)^{NS}
5	*S. cerevisae, S. paradoxus*	0.006 (0.048)^{NS}
5	*S. cerevisae, S. mikatae*	−0.398 (0.048)***
5	*S. cerevisae, S. kudriavzevii*	0.024 (0.048)^{NS}
(b) Comparison between the three numbers of markers (see the "Simulations" section for details)		
Number of markers		Precision
1		0.400 (0.036)
		Difference in precision in relation to 1 marker
3		−0.108 (0.042)*
5		−0.122 (0.042)**

^{NS} $p > 0.05$.
* $0.05 > p > 0.01$.
** $0.001 > p > 0.01$.
*** $p < 0.001$.

analysis. We have not included any suggested chronospecies (other than the final form), such as the many named forms of *Bison*, because our goal is to estimate potentially temporarily co-occurring species. For extant and recently extinct species, we generally followed the IUCN taxonomy (Schipper et al., 2008), although we removed six highly dubious species. We performed a complete literature review for earlier extinctions not covered by the IUCN. For island species and smaller continental species, we generally followed Turvey and Fritz (2011), although with several exceptions. For large continental species we followed (Sandom et al., 2014). A full list of all deviations from the IUCN taxonomy and species list can be found in Appendix B.

Our phylogenetic estimate was based on the following hierarchy (a few exceptions will be mentioned later). (1) Infraclass and subclass assignment (monotremes (marsupials, eutherians)); (2) a recent 35 kbp study of nearly all mammalian families (Meredith et al., 2011); (3) monophyly of all families as defined by the IUCN (Schipper et al., 2008); (4) monophyly of rodents, to place the two murid rodent families (Calomyscidae and Platacanthomyidae) that were not included in the report of Meredith et al. (2011); (5) molecular phylogenies at the family level using species with more than one independent marker; (6) detailed family-level morphological phylogenies; (7) genus monophyly following the taxonomy of the IUCN (Schipper et al., 2008); (8) genus-level phylogenies based on a single marker; and (9) relationships between species following the most recent supertree of all mammals (Bininda-Emonds et al., 2007). The workflow is illustrated in Fig. 2.

For level five (the family level), we included all taxa with more than one independent marker. If no species in the genus had information for more than one marker, we included species with information for just a single marker, but in these cases, we constrained the genera in question to be monophyletic.

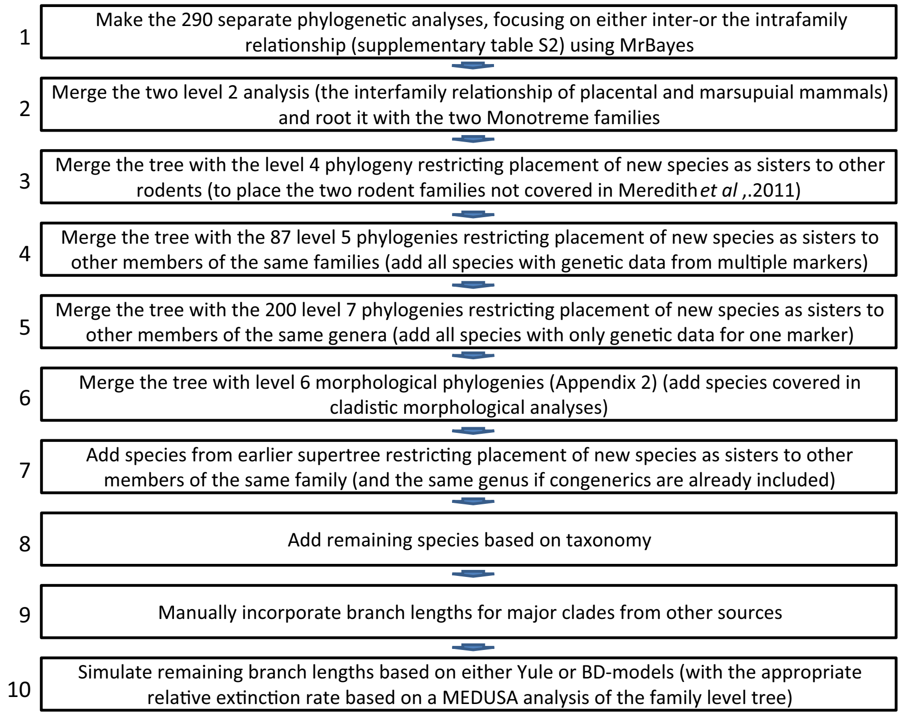
Family monophyly and family assignment in the New World monkeys are unsettled, although the monophyly of the New World monkey superfamily is settled beyond doubt (Perelman et al., 2011). We therefore treated this superfamily as a "family" throughout this paper. The family-level assignment of some taxa to Muridae or Cricetidae, particularly the maned rat *Lophiomys imhausi,* is also problematic. We considered this species to belong to Muridae based on Jansa and Weksler (2004), differing from the IUCN, classifies it within Cricetidae (Schipper et al., 2008). Ideally, the same approach applied for the New World monkeys could be used to analyze Muridae and Cricetidae simultaneously. However, the very large number of species in these two families makes it computationally unfeasible to employ this approach for these two families, and we therefore reluctantly chose to enforce family monophyly for them instead.

The recently extinct Thylacinidae and Chaeropodidae were placed through analysis at the order level, instead of the family level for Dasyuromorphia and Peramelemorphia, while enforcing family monophyly. No study has suggested non-monophyly of either order or family, which therefore follows our stated idea of using taxonomy as a partial topological constraint, so that taxa with limited genetic information are forced to be placed near species in the same taxonomic grouping, while species with more genetic information can be placed more freely. Lepilemuridae was placed as sister to Cheirogaleidae following Perelman et al. (2011).

After addressing the genetic data, we merged the data with taxonomic and morphological data and with the earlier supertree of Bininda-Emonds et al. (2007). This included placement of the 19 extinct families, which were placed based on phylogenetic or taxonomic information (Appendix B). For these extinct groups, the taxonomic restrictions of the hierarchy were relaxed, and paraphyletic families or genera were accepted because non-monophyletic groups have not always led to taxonomic changes in paleontology (e.g., in Cingulata, both Glyptodontidae and Pampatheriidae appear to be nested within Dasypodidae; Porpino et al., 2009).

For a few Late Pleistocene or Holocene extinctions, where only a few hundred base pairs of ancient DNA are available and where the genetic data are congruent with morphological hypotheses of relationships (e.g., saber-toothed cats following Barnett et al., 2005), we chose to place the groups with certainty rather than performing phylogenetic analyses, which would incorporate uncertainty. We made this decision because ancient DNA is more prone to errors than contemporary DNA (Briggs et al., 2010) and a large proportion of apparent genetic uncertainty could therefore be caused by a lower quality of the ancient DNA.

After the topology was settled, as discussed above, we incorporated branch lengths into the phylogeny. Most nodes at the family level and occasionally the first branching within the families were manually incorporated from other sources (generally Meredith et al., 2011, see Table A.1). We assumed that the age of these nodes was known with certainty, which we acknowledge to be untrue, but they were generally fairly precise (the ages of most nodes from



1 Make the 290 separate phylogenetic analyses, focusing on either inter-or the intrafamily relationship (supplementary table S2) using MrBayes

2 Merge the two level 2 analysis (the interfamily relationship of placental and marsupuial mammals) and root it with the two Monotreme families

3 Merge the tree with the level 4 phylogeny restricting placement of new species as sisters to other rodents (to place the two rodent families not covered in Meredith *et al* ,.2011)

4 Merge the tree with the 87 level 5 phylogenies restricting placement of new species as sisters to other members of the same families (add all species with genetic data from multiple markers)

5 Merge the tree with the 200 level 7 phylogenies restricting placement of new species as sisters to other members of the same genera (add all species with only genetic data for one marker)

6 Merge the tree with level 6 morphological phylogenies (Appendix 2) (add species covered in cladistic morphological analyses)

7 Add species from earlier supertree restricting placement of new species as sisters to other members of the same family (and the same genus if congenerics are already included)

8 Add remaining species based on taxonomy

9 Manually incorporate branch lengths for major clades from other sources

10 Simulate remaining branch lengths based on either Yule or BD-models (with the appropriate relative extinction rate based on a MEDUSA analysis of the family level tree)

Fig. 2. An illustration of the workflow to construct the phylogeny.

Meredith et al., 2011 showed 95% confidence intervals shorter than 10% of their mean value), and the dependence of the ages of individual nodes means that the uncertainty of the ages of individual nodes cannot easily be implemented. A few interfamily nodes did not exhibit 100% posterior probability, in which case all potential trees were manually modified separately. All other nodes were estimated assuming that evolution followed the Yule or BD models.

2.3.2. Genetic analyses

We performed a total of 290 separate phylogenetic analyses (two infraclass phylogenies/hierarchical level 2, one additional family assignment/hierarchical level 3, 87 family-level phylogenies/hierarchical level 5 and 200 genus-level phylogenies/hierarchical level 7; Table A.2, Appendix C). We initially searched for phylogenetic studies within each taxon with a broad taxonomic coverage, and whenever possible, we attempted to restrict our analysis to sequences from a single or a few known sources as much as possible (Table A.2), as we suspect misidentifications to be more frequent in Genbank sequences from unpublished studies or studies without a taxon-specific phylogenetic focus because they will rely on non-expert identifications more often. However, to maximize taxonomic coverage, most analyses were supplemented with Genbank sequences, and many analyses did contain sequences from many sources. We used Phylota 1.5 (Sanderson et al., 2008) to supplement taxonomic studies, or as a start-up, and added additional information from Genbank to include sequences released since the last update of Phylota, or sequences not included in Phylota, such as data from fully sequenced mitochondria. For each group, we performed individual assessments of which clusters to include, trading off including as much information as possible against using as few sources as possible. We generally chose not to include markers for which only a few species have been sampled because they will rarely be identified by experts and will therefore exhibit a higher change of being misidentified.

We generally restricted our analysis to include only a single mitochondrial marker to reduce the likelihood that the mitochondrial signal may overwrite a signal for multiple nuclear markers. We chose the marker with the widest coverage of the taxonomic group in question, and if several markers presented the same taxonomic coverage, we preferred the longest protein-coding gene over 16 s or 12 s. For two analyses (the lemur genera *Eulemur* and *Avahi*) where the widest taxonomic coverage was found for a large multiprotein region, we restricted the analysis to the first 1800 bp of the sequence excluding indels. This maximum of 1800 bp was chosen because it is slightly larger than the largest single markers (16 s), and it was therefore necessary to apply this restriction only in rare cases. This omission of data could seem odd, but this decision was made because there is often a poor overlap between gene and species trees as a consequence of incomplete lineage sorting and/or hybridization (see for instance Scally et al., 2012). Regardless of how many mitochondrial markers are analyzed, they still come from a single gene tree and therefore potentially differ from the species tree, whereas there is not a similar problem for long nuclear markers because recombination means that the evolutionary history of different parts of the markers may be different. If many mitochondrial markers were to be included in the analysis, along with a few nuclear genes, the resulting tree would essentially be identical to the mitochondrial gene tree, whereas our treatment means that we downweight the mitochondrial gene tree. If only mitochondrial data are available, the inclusion of more markers would at least provide a better estimate of the gene tree, which would be advantageous if we were specifically interested in the most likely tree; however, our goal is to estimate the overall phylogeny including precision. Deviations between species and gene trees are most likely for short branches (Hein et al., 2005), and these cases also normally exhibit the lowest support values. By restricting our analysis to a single marker, the trees from the posterior distribution of our analysis may, thus, provide a better estimate of the phylogenetic uncertainty, even though we decrease the likelihood of obtaining the true gene tree by decreasing the support of the most dubious clades.

We allowed small amounts of non-overlap where not all taxa were sampled for the same markers. If we assumed a constant relative substitution rate between markers throughout the phylogeny, non-overlapping species could be placed with some certainty because ((A1:1, A2:1):3, A4:4) and (A1:2, A3:2):2, A4:4) can logically be merged to (((A1:1, A2:1):1, A3:2):2, A4:4). However, we consider the assumption of a constant relative substitution rate to be unlikely and therefore acknowledge that such non-overlap is likely to be problematic for real-world datasets. Nevertheless, we believe this assumption to be the least problematic solution to a suboptimal situation as long as the non-overlap is small. In some cases where such small non-overlap was present, our analyses included a couple of mitochondrial markers because we judged increased species coverage to be more important than restricting our analyses to a single mitochondrial marker.

All analyses were rooted with the closest available outgroup. Generally, a few outgroups were included whenever possible, but phylogenetic relatedness was prioritized more highly than including several outgroups, and many analyses therefore included only a single outgroup.

For all markers, we aligned the sequences using ClustalW (Larkin et al., 2007), after which we removed all sites that were not present in at least 50% of the species with data on that particular marker, and we checked all alignments manually and occasionally made minor modifications by hand. Species that were not sampled for a particular marker were coded as missing data. The best model for each partition was estimated using jModeltest (Posada, 2008), and the phylogeny of each taxon was estimated with MrBayes 3.1 (Ronquist and Huelsenbeck, 2003), using different partitions for each marker, with the parameter shape, pinvar, statefreq, revmat and tratio unlinked between partitions, but otherwise employing the standard parameters. The analyses were generally run until the standard deviation of the split frequencies of all partitions with a frequency higher than 0.05 was less than 0.010, although the analyses were stopped at 10 or 20 million generations if the split frequencies were below 0.030 at either of these points. All analyses were performed with two runs and four chains, except for the Eutherian tree, which was analyzed with only a single cold chain for computational reasons but was run until the standard deviation of the split frequencies of all partitions with a frequency higher than 0.05 was less than 0.010.

Except when otherwise mentioned, all analyses were performed in R 2.13.1, 2.14.1 or 2.15.1 (R Development Core Team, 2011) using functions from APE, Biostrings, caper, Geiger, phangorn, phylobase, phyloch and seqinr (Pages et al., 2003; Paradis et al., 2004; Charif and Lobry, 2007; Harmon et al., 2009; Heibl, 2008; R Hackathon, 2011; Schliep, 2011; Orme et al., 2012). New codes were written to merge phylogenies and to simulate branch lengths (Appendix D).

3. Results and discussion

3.1. The produced phylogeny

The overall phylogeny is very similar to the phylogeny produced by Meredith et al. (2011), and the most likely topology was identical to their nucleotide and/or amino acid topology for all clades. Because our raw data were the same, our results cannot be viewed as independent evidence, but the similarity of the results demonstrates at least moderate robustness of the results of methodology, and our analysis therefore strengthens their conclusions.

Almost all the genetically placed family-level nodes were recovered with 100% posterior support (here and throughout the paper, all mentions of posterior support refer to the post burn-in value). The only four exceptions were the exact placement of Calomyscidae within Muroidea (posterior support = 0.929 for (((Cricetidae, Muridae), Nesomyidae), Calomyscidae), posterior support = 0.071 for ((Cricetidae, Muridae), (Calomyscidae, Nesomyidae))); the placement of hyraxes within Afrotheria (posterior support = 0.670 for (Procavidae, Proboscidea), Sirenia), posterior support = 0.330 for ((Proboscidea, Sirenia), Procavidae)); the internal relationship between the four Australian marsupial orders (posterior support = 0.994 for (((Dasyuromorphia, Peramelemorphia), Notoryctemorphia), Diprotodontia), posterior support = 0.006 for (((Dasyuromorphia, Diprotodontia), Notoryctemorphia), Peramelemorphia)); and the topology between three overall clades within Yangochiroptera (superfamily assignment follows Meredith et al., 2011), (posterior support = 0.550 for ((Emballonuroidea, Noctilionoidea), Vespertilionoidea + Myzopodidae), posterior support = 0.411 for ((Emballonuroidea, Vespertiliono idea + Myzopo didae), Noctilionoidea), posterior support = 0.039 for ((Noctilionoidea, Vespertilionoidea + Myzopodidae), Emballonuroidea)).

The way in which we handled missing species means that the results for the 1000 generated trees cannot meaningfully be combined into a single tree, but for illustration purposes, a single representative tree is shown on Fig. 3. All 1000 trees are available in Appendix D (the results will be updated on an irregular basis based on new published data, and the newest version can be downloaded from our homepage at all times (http://bios.au.dk/om-instituttet/organisation/oekoinformatik-biodiversitet/data/)). Individual researchers can therefore quickly download the data and identify likely sisterhood relationships within any specific group of interest. The figure shows that while species with different amounts of genetic information are spread throughout the phylogeny, some parts of the tree contain a larger proportion of species without genetic data (Fig. 3c) or a larger proportion of extinct species not covered by the IUCN (Fig. 3d). Hence, restricting the species considered to extant and historically extinct species or, even further, to extant species with genetic data would mean that the resulting analysis would be performed on a biased subset of the species and would therefore likely be biased itself. To illustrate topological uncertainty, we have plotted the topology of all 1000 trees of four selected clades: two smaller clades (non-saber-toothed (modern) cats (Felidae excluding Machairodontinae), with 40 species; and weasels and relatives (Mustelidae) with 59 species) and two larger clades (carnivores (Carnivora), with 304 species; and shrews (Soricidae), with 378 species) (Fig. 4). The examples illustrate two ends of a continuum, and most clades will be intermediate between these examples. The first three examples show clades with a well-known topology, whereas the last clade is one of the least-studied and therefore most uncertain clades, where only a few branches are well supported.

Some end users of the phylogeny may disagree with the way in which we handle species with uncertain placement or with our handling of polytomies, as described in Figs. 1b–1d. We have therefore added a third tree of 4125 species, which is the largest possible tree of species with genetic data plus other species with unambiguous placement in the tree, initially constructed based on the species with genetic data. Any bifurcating clade in morphological trees, taxonomically based trees or the BE tree where only one species has genetic data would have such an unambiguous placement and we therefore replaced the single species with the entire bifurcating clade for such clades. In addition, we added any species or clades that were sisters to multiple species with genetic data in the morphological trees, taxonomically based trees or the BE tree when the posterior support for the monophyly of the potential sister clades was equal to 1. While doing this, we removed species from polytomies to generate the largest possible strictly bifurcating tree.

3.2. Comparisons of the branch lengths with previous analyses

To estimate the precision of our inferred branch lengths, we compared the median age of the MRCA (most recent common

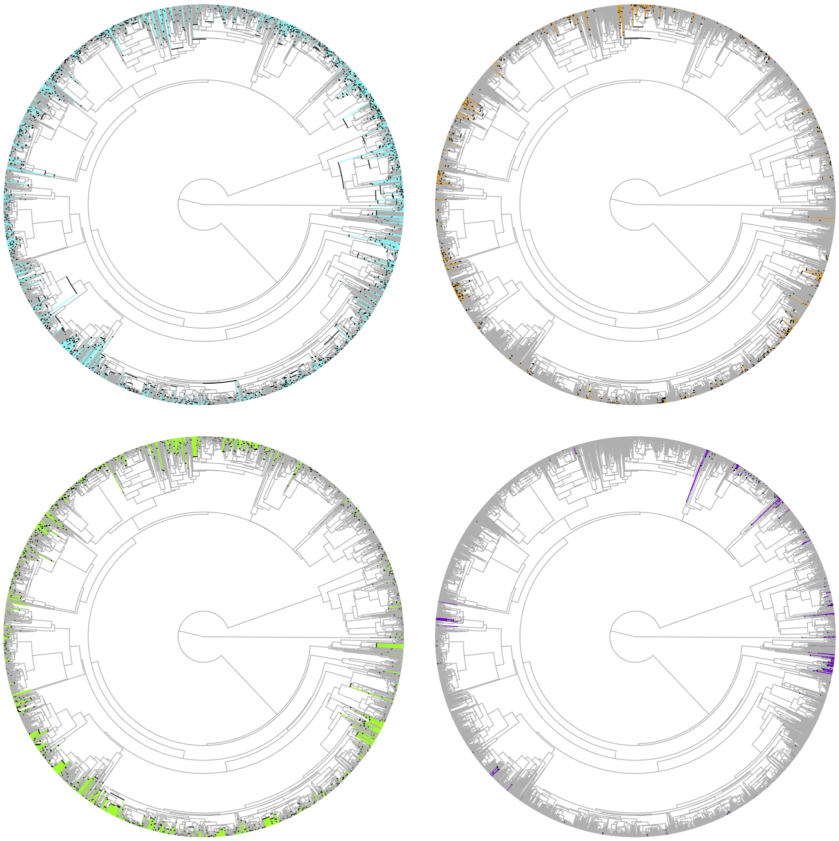


Fig. 3. A representative tree from the posterior distribution. In the upper left corner, species with genetic data from more than one independent marker are highlighted in light blue. In the upper right corner, species with genetic data from a single marker are indicated in orange. In the lower left corner, species accepted by the IUCN without genetic data are indicated in light green. In the lower right corner, prehistoric extinct species not covered by the IUCN and without genetic data are indicated in purple. In all cases, the internal branches as well as the remainder of the species are colored in black.

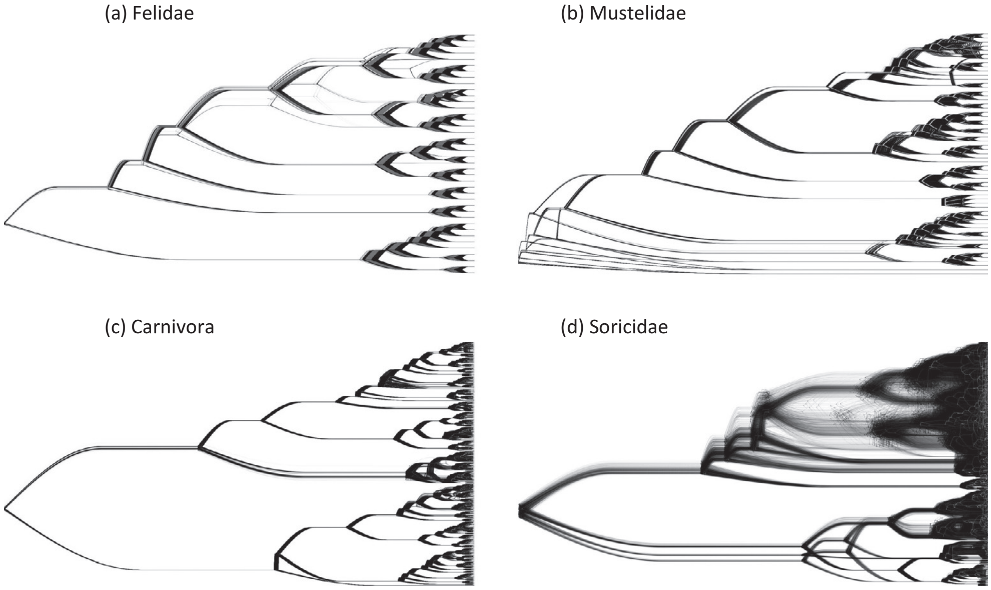


Fig. 4. Topological uncertainty in four selected clades. Topological uncertainty according to the presented heuristic-hierarchical Bayesian model of four selected clades, presented using the program DensiTree (Bouckaert, 2010): (a) Felidae excluding Machairodontinae (non-saber-tooth cats): 40 species (Fig. 3a); (b) Mustelidae (weasels and relatives): 59 species; (c) Carnivora (carnivores): 304 species; and (d) Soricidae (shrews): 378 species. The branch lengths were standardized for illustrative purposes, so that trees with similar topologies exhibit the most similar branch lengths but do not represent our estimated branch lengths. The width of the branches showing different topologies represents the relative probability of each possible topology.

ancestor) of 34 clades within the family Felidae excluding Machairodontinae (non-saber-toothed cats, hereafter Felidae) and 34 clades within the family Mustelidae from our analysis with the results from recent phylogenies (Johnson et al., 2006; Sato et al., 2012). The absolute ages of the clades were somewhat greater in our analysis (mean ratio of MRCA ages between the two analyses of 1.37, with a range of 0.15–2.76 for Felidae; mean of 0.82, with a range of 0.13–3.56 for Mustelidae). Only 13 of the median values for Felidae and 4 values for Mustelidae from our analysis were within the 95% confidence interval from the previous analysis, but 33 of the point-values for Felidae and 28 values for Mustelidae from the previous analyses were within the 95% HPD (highest posterior density) interval of our analysis. The 95% HPD interval in our analysis was substantially larger than the 95% confidence levels in the previous analysis (mean ratio of 2.09, range of 0.77–4.59 for Felidae, mean of 3.19, range of 0.87–20.13 for Mustelidae). The fact that the values from the previous analyses were outside the 95% HPD interval from our analysis for a total of five out of 66 comparisons indicates that our analysis provides a fair, albeit imprecise estimate of branching times within the families. However, there will be branches that are poorly estimated by our approach because our dating strategy assumes constant speciation and extinction rates. Hence, any clade experiencing large temporal variation in diversification rates will exhibit poorly estimated branch lengths.

We analyzed our topology using the rooted Robinson-Foulds distance (Robinson and Foulds, 1981) (hereafter RF distance), which compares two trees by counting the number of clusters found in only one of the two trees and varies from 0 for identical trees to $2n–4$ (where n is the number of tips) for trees with no clades in common. We calculated the RF distances between each of our 1000 individual trees and 1000 trees each based on three previous phylogenies: (a) the BE tree (using the modified version by Fritz et al., 2009), (b) a recent supertree of all carnivores generated by Nyakatura and Bininda-Emonds (2012), and (c) the carnivore subtree from the BE tree (Table 2). We removed all species that were present in only one of the comparisons and subsequently resolved all polytomies in all the trees assuming a Yule model.

The comparisons consistently showed fairly high RF distances, many of which were approximately 50% of the maximum distances. The RF distances between trees constructed from the new carnivore supertree (Nyakatura and Bininda-Emonds, 2012) were low as a consequence of the small number of polytomies in this tree and were only approximately half of the RF distances from our trees constructed with the same species list, whereas the earlier BE tree presented substantially more polytomies and therefore much higher RF distances. The RF distances between our tree and the new carnivore supertree (Nyakatura and Bininda-Emonds, 2012) were substantially lower than the RF distances between our tree and the carnivore subtree from the BE tree. This observation suggests that the carnivore supertree and our tree would converge with increasing data, which would be expected because most phylogenetic methods produce consistent results (i.e., converge towards the true phylogeny with increasing data). However, the difference between the two trees was still fairly large, and because the underlying datasets are comparable, although not identical, methodological differences are likely a major reason for the relatively large RF distances. We do not know the true phylogeny and therefore cannot state with certainty which tree is closest to the true phylogeny, but we believe that our treatment may be more suitable because the errors caused by limited information have been reduced. For instance, Nyakatura and Bininda-Emonds (2012) found that their analysis produced *Vulpes ferrilata* and *Dusicyon australis* as sister species and noted this result as a clear artifact, whereas we forced *Vulpes ferrilata* to be placed within *Vulpes* because it was placed without genetic data and nearly everybody believes this placement to be correct (Clark et al., 2008; Nyakatura and Bininda-Emonds, 2012).

3.3. Model strengths and limitations

Our approach produced a new and most likely improved estimate of the mammalian phylogeny and one that includes all known Late Pleistocene extinct species for the first time, thus better representing natural phylogenetic patterns in mammals given the many anthropogenic and probable anthropogenic extinctions. This means that our tree enables direct tests of the effects on these extinctions on various evolutionary patterns for the first time, allowing the estimation of biases introduced by considering only extant species.

Our new method we used to produce the phylogeny is expected to produce more accurate topologies, as we demonstrated via simulations in Section 2.2. We have restricted ourselves to a few simple analyses of the phylogeny because a full exploration of which patterns change when using this phylogeny as opposed to the BE tree is beyond the scope of this study. We suggest that it could be worth rerunning some of the previous analyses based on earlier mammalian phylogenies to verify that their conclusions are robust.

Our overall heuristic-hierarchical approach clearly employs what has been called an “inappropriate appeal to authority” (Gatesy et al., 2002), meaning that our tree contains groupings that are classically assumed to be monophyletic, even if the data suggest otherwise, and our tree may therefore be viewed as less philosophically appropriate than some recent treatments of smaller groups, such as the Nyakatura and Bininda-Emonds (2012) Carnivora supertree. However, we believe that our treatment may be more suitable in many cases because we reduce the errors caused by limited information, as observed for the already discussed placement of *Vulpes ferrilata*. It is clear that the existence and frequency of such artifacts may provide information on the overall precision of the tree, but we believe that for “down-stream” analyses, it is best to remove as many of these artifacts as possible.

Table 2

Comparisons of Robinson–Foulds distances. The median and range of values are given for the Robinson–Foulds RF distances (RF distance) from sets of 1000 trees. The column RF distance$_{\text{Ours}}$ shows the RF distances between 1000 trees from our heuristic-hierarchical Bayesian phylogeny of all mammals for different groups of species. The column RF distance$_{\text{Old}}$ shows the RF distances between 1000 trees constructed by randomly resolving multifurcations in the trees we use for comparison. The column RF distance$_{\text{Ours/Old}}$ shows the RF distances between the 1000 trees from our heuristic-hierarchical Bayesian phylogeny and the 1000 trees constructed by randomly resolving multifurcations in the trees we use for comparison.

Comparison	Species	RF distance$_{\text{Ours}}$	RF distance$_{\text{Old}}$	RF distance$_{\text{Ours/Old}}$
Our heuristic-hierarchical phylogeny	5747	5226 (4957–5487)	–	–
The Bininda-Emonds supertree (Bininda-Emonds et al., 2007)	4908	3922 (3672–4142)	4460 (4331–4584)	6408 (6266–6536)
The new carnivore supertree (Nyakatura and Bininda-Emonds, 2012)	280	66 (26–110)	34 (14–40)	206 (178–224)
The Bininda-Emonds supertree (Carnivores)	277	66 (26–110)	88 (54–112)	312 (280–338)

The specific choices we made for the heuristic hierarchy in our example of mammal evolution may also be somewhat subjective. However, we are convinced that this strategy provides the best way to address the low quantities of data available for most mammalian species. The study conducted by Meredith et al. (2011) was very thorough and included nearly all extant families. These authors explicitly attempted to include all old lineages, and all families with more than one representative in their study showed 100% posterior probability for monophyly in our re-analysis of their data. If one or more of the remaining families appeared to be non-monophyletic in another phylogeny, it would therefore be more likely to indicate the existence of phylogenetic problems, such as long-branch attraction, or a technical problem, such as phylogenies constructed based on non-homologous sequences.

The assignment of species to a genus is more problematic than assignment to a family because it is clear that mammalian genera are not always monophyletic, and the assignment of species to individual genera is debatable at times. Our approach will, however, recover the polyphyly of a genus as long as each of the non-related species groups has representatives with genetic data for more than one marker. Any reasonably robustly resolved polyphyly will therefore be included in the analysis.

The morphological phylogenies were placed low in the hierarchy. This is not meant as a critique of the morphological phylogenies themselves. However, this decision was made because we wanted to estimate phylogenetic uncertainty as well as the most likely topology, which requires a clear probabilistic theory of evolution, and such theories are difficult to apply for morphological data, especially for ordered characters (Nyakatura and Bininda-Emonds, 2012).

Our choice of hierarchy has two peculiar results. (1) Because the topological restriction is lower when more information is available and because there is generally a large overlap between species with an intensely studied taxonomy and species with large amounts of sequence data, our approach causes us to rely more on taxonomy in the groups where such reliance is most problematic; and (2) because we enforce genus monophyly when no or only one marker is available but impose no restriction when two or more markers are available, the apparent topological resolution is oddly shaped, showing minimum resolution for intermediate quantities of data. Both of these factors could potentially be mitigated by following taxonomy for only a set proportion of the trees while using non-genus (or non-family) sisters for the remainder. This would, however, require precise knowledge of the probability that genera or families are non-monophyletic, which is not available at the moment.

It should also be noted that while our approach ensures that phylogenetic uncertainty can be easily taken into account in downstream analyses, the posterior distribution of the trees represents only the uncertainty given our chosen model. Uncertainty driven by different analytical approaches providing complete support for different topologies, as observed for some of the most difficult intra-ordinal branches between approaches such as ours as opposed to maximum information parsimony or multispecies coalescence analysis (see Meredith et al., 2011; Song et al., 2012; O'Leary et al., 2013), are therefore not incorporated.

3.4. Model applications

We believe that our heuristic-hierarchical Bayesian approach may prove to be a good method for constructing species-level phylogenies for large clades with a variable density of genetic information between species. In these circumstances, this strategy may provide a superior alternative to the often-applied supertree approaches, such as that of Baum (1992), or the Bayesian version discussed by Ronquist et al. (2004). The low precision of branch length information in our method is of course suboptimal, and for groups with denser sampling of species and markers, alternative methods, such as the one used by Jetz et al. (2012), may be superior. For mammals however, it is striking that the groups that have received most attention, such as cats or bears (Johnson et al., 2006; Pages et al., 2008), are also some of the groups that suffered the highest Late Pleistocene and Holocene extinction rates (see electronic Appendix B), while groups showing much lower extinction rates, such as shrews, have received much less attention. We therefore believe that it may take years before alternative approaches with higher weights on branch lengths may prove optimal for mammalian species-level phylogenies for most families.

The suggested approach can easily be extended to other taxonomic groups. Because the approach relies on taxonomic information, it will be useful mainly for clades with a fairly well-understood taxonomy, e.g., most vertebrate or vascular plant groups, whereas it would not be effective for less well-studied groups. This strategy would be especially useful for groups exhibiting substantial variation in phylogenetic information between species, as groups with similar information levels can be analyzed using other simpler methods that retain branch-level information, such as the method of Jetz et al. (2012). An important advantage of our approach may lie in the placement of missing species because our method enables species to be placed near taxonomically close species without enforcing taxonomic monophyly for the species with genetic data, which would be very difficult under standard methods. Researchers attempting to construct species-level phylogenies of large groups ultimately need to decide if their main priority is to obtain as accurate a topology as possible, in which case we believe our approach to be the best choice, or as accurate branch lengths as possible, in which case other approaches, such as that of Jetz et al. (2012), may be more effective.

Acknowledgements

This work was funded by the Danish Natural Science Research Council (#11-115750 to S.F.) and the European Research Council for financial support (ERC Starting Grant #310886 "HISTFUNC" to J.C.S.). We thank the associate editor and two anonymous reviewers for comments and suggestions that have improved this manuscript.

Appendices A–D. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2014.11.001.

References

Alfaro, M.A., Santinia, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D.L., Carnevale, G., Harmon, L.J., 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc. Natl. Acad. Sci. 106, 13410–13414.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J., Weightman, A.J., 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl. Environ. Microbiol. 71, 7724–7736.

Barnosky, A.D., Koch, P.L., Feranec, R.S., Wing, S.L., Shabel, A.B., 2004. Assessing the causes of late Pleistocene extinctions on the continents. Science 306, 70–75.

Barnett, R., Barnes, I., Phillips, M.J., Martin, L.D., Harrington, C.R., Leonard, J.A., Cooper, A., 2005. Evolution of the extinct sabretooths and the American cheetahlike cat. Curr. Biol. 15, R589–R590.

Baum, B.R., 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41, 3–10.

Bertheau, C., Schuler, H., Krumbock, S., Arthofer, W., Schuler, C., 2011. Hit or miss in phylogeographic analyses: the case of the cryptic NUMTs. Mol. Ecol. Resour. 11, 1056–1059.

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L., Purvis, A., 2007. The delayed rise of present-day mammals. Nature 446, 507–512.

Bouckaert, R.R., 2010. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics 26, 1372–1373.

Briggs, A.W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., Pääbo, S., 2010. Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. Nuleid Acids Res. 38, e87.

Charif, D., Lobry, J.R., 2007. Seqin R 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M. (Eds.), Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Springer-Verlag, New York, pp. 207–232.

Clark, H.O., Newman, D.P., Murdoch, J.D., Tseng, J., Wang, Z.H., Harrid, R.B., 2008. *Vulpes ferrilata* (Carnivora: Canidae). Mamm. Species 821, 1–6.

D'Erchia, A.M., Gissi, C., Pesole, G., Saccone, C., Arnason, i.U., 1996. The guinea-pig is not a rodent. Nature 381, 597–600.

Dalerum, F., Cameron, E.Z., Kunkel, K., Somers, M.J., 2009. Diversity and depletions in continental carnivore guilds: implications for prioritizing global carnivore conservation. Biol. Lett. 5, 35–38.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.

Eiserhardt, W.L., Svenning, J.C., Borchsenius, F., Kristiansen, T., Balslev, H., 2013. Separating environmental and geographical determinants of phylogenetic community structure in Amazonian palms (Arecaceae). Bot. J. Linn. Soc. 171, 244–259.

Faurby, S., Jøgensen, A., Kristensen, R.M., Funch, P., 2012. Distribution and speciation in marine intertidal tardigrades: testing the roles of climatic and geographical isolation. J. Biogeogr. 39, 1596–1607.

Fitzjohn, R.G., Maddison, W.P., Otto, S.P., 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol. 58, 595–611.

Fritz, S.A., Bininda-Emonds, O.R.P., Purvis, A., 2009. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. Ecol. Lett. 12, 538–549.

Gatesy, J., Matthee, C., DeSalle, R., Hayashi, C., 2002. Resolution of a supertree/supermatrix paradox. Syst. Biol. 51, 652–664.

Harmon, L., Weir, J., Brock, C., Glor, R., Challenger, W., Hunt, G., 2009. Geiger: Analysis of Evolutionary Diversification. R Package Version 1.3-1. <http://CRAN.R-project.org/package=geiger>.

Heibl, C. 2008. PHYLOCH: R Language Tree Plotting Tools and Interfaces to Diverse Phylogenetic Software Packages. <http://www.christophheibl.de/Rpackages.html>.

Hein, J., Schierup, M.H., Wiuf, C., 2005. Gene Genealogies, Variation and Evolution. A Primer in Coalescent Theory. Oxford University Press.

Huelsenbeck, J.P., Rannala, B., Masly, J.P., 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science 288, 2349–2350.

Jansa, S.A., Weksler, M., 2004. Phylogeny of muroid rodents: relationships within and among major lineages as determined by IRBP gene sequences. Mol. Phylogenet. Evol. 31, 256–276.

Jansa, S.A., Giarla, T.C., Lim, B.K., 2009. The phylogenetic position of the rodent genus *Typhlomys* and the geographic origin of Muroidea. J. Mammal. 90, 1083–1094.

Jetz, W., Thomas, G.H., Boy, J.B., Hartmann, K., Mooers, A.O., 2012. The global diversity of birds in space and time. Nature 491, 444–448.

Johnson, W.E., Eizirik, E., Pecon-Slattery, J., Murphy, W.J., Antunes, A., Teeling, E., O'Brien, S.J., 2006. The late Miocene radiation of modern Felidae: a genetic assessment. Science 311, 73–77.

Kissling, W.D., Eiserhardt, W.L., Baker, W.J., Borchsenius, F., Couvreur, T.L.P., Balslev, H., Svenning, J.C., 2012. Cenozoic imprints on the phylogenetic structure of palm species assemblages worldwide. P. Natl. Acad. Sci. USA 109, 7379–7384.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. ClustalW and ClustalX version 2. Bioinformatics 23, 2947–2948.

Lukoschek, V., Keogh, J.S., Avise, J.C., 2011. Evaluating fossil calibrations for dating phylogenies in light of rates of molecular evolution: a comparison of three approaches. Syst. Biol. 61, 22–43.

Meredith, R.W., Janečka, J.E., Gatesy, J., Ryder, O.A., Fisher, C.A., Teeling, E.C., Goodbla, A., Eizirik, E., Simão, T.L.L., Stadler, T., Rabosky, D.L., Honeycutt, R.L., Flynn, J.L., Ingram, C.M., Cynthia, C., Williams, T.L., Robinson, T.J., Burk-Herrick, A., Westerman, M., Ayoub, N.A., Springer, M.S., Murphy, W.J., 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334, 521–524.

Mooers, A.O., 1995. Tree balance and tree completeness. Evolution 49, 379–384.

Nyakatura, K., Bininda-Emonds, O.R.P., 2012. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. BMC Biol. 10, 12.

O'Leary, M.A., Bloch, J.I., Flynn, J.J., Gaudin, T.J., Giallombardo, A., Giannini, N.P., Goldberg, S.L., Kraatz, B.P., Luo, Z., Meng, J., Ni, X., Novacek, M.J., Perini, F.A., Randall, Z.S., Rougier, G.W., Sargis, E.J., Silcox, M.T., Simmons, N.B., Spaulding, M., Velazco, P.M., Weksle, R.M., Wible, J.R., Cirranello, A.L., 2013. The placental mammal ancestor and the post–K-Pg radiation of placentals. Science 339, 662–667.

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., Pearse, W., 2012. Caper: Comparative analyses of phylogenetics and evolution in R. R package version 0.5.

Pages, H., Gentleman, P.R.A., DebRoy, R., 2003. Biostrings: String Objects Representing Biological Sequences, and Matching Algorithms. R Package Version 2.20.3.

Pages, M., Calvignac, S., Klein, C., Paris, M., Hughes, S., Hanni, C., 2008. Combined analysis of fourteen nuclear genes refines the Ursidae phylogeny. Mol. Phylogenet. Evol. 47, 73–83.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Perelman, P., Johnson, W.E., Roos, C., Seuánez, H.N., Horvath, J.E., Moreira, M.A.M., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M.P.C., Silva, A., O'Brien, S.J., Pecon-Slattery, J., 2011. A molecular phylogeny of living primates. PLOS Genet. 7, e1001342.

Porpino, K.O., Fernicola, J.C., Bergqvist, L.P., 2009. A new Cingulate (Mammalia: Xenarthra), *Pachyarmatherium brasiliense* sp. nov., from the late Pleistocene of northeastern Brazil. J. Vertebr. Paleontol. 29, 881–893.

Posada, D., 2008. JModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25, 1253–1256.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. <http://www.R-project.org/>.

R Hackathon, 2011. Phylobase: Base Package for Phylogenetic Structures and Comparative Data R Package Version 0.6.5. <http://CRAN.R-project.org/package=phylobase>.

Raup, D.M., 1985. Mathematical models of cladogenesis. Paleobiology 11, 42–52.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosc. 53, 131–147.

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.

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

Ronquist, F., Huelsenbeck, J.P., Britton, T., 2004. Bayesian supertrees. In: Bininda-Emonds, O.R.P. (Ed.), Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life, vol. 3. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 193–224, Chapter 9.

Rosenfeld, J.A., Payne, A., DeSalle, R., 2012. Random roots and lineage sorting. Mol. Phylogenet. Evol. 64, 12–20.

Sanderson, M.J., Boss, D., Chen, J., Cranston, K.A., Wehe, A., 2008. The PhyLoTA Browser: processing GenBank for molecular phylogenetics research. Syst. Biol. 57, 335–346.

Sandom, C., Faurby, S., Sandel, B., Svenning, J.C., 2014. Global late Quaternary megafauna extinctions linked to humans, not climate change. Proc. Roy. Soc. London Ser. B 281, 20133254.

Sato, J.J., Wolsan, M., Prevosti, F.J., D'Elía, G., Begg, C., Begg, K., Hosoda, T., Campbell, K.L., Hitoshi, S., 2012. Evolutionary and biogeographic history of weasel-like carnivorans (Musteloidea). Mol. Phylogenet. Evol. 63, 745–757.

Scally, A., Duthei, l.J.Y., Hillier, LW, Jordan, G.E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S.H., Schwalie, P.C., Tang, Y.A., Ward, M.C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L.N., Ayub, Q., Ball, E.V., Beal, K., Bradley, B.J., Chen, Y., Clee, C.M., Fitzgerald, S., Graves, T.A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G.K, Lunter, G., Meader, S., Mort, M., Mullikin, J.C., Munch, K., O'Connor, T.D., Phillips, A.D., Prado-Martinez, J., Rogers, A.S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J.T., Stenson, P.D., Turner, D.J., Vigilant, L., Vilella, A.J., Whitener, W., Zhu, B., Cooper, D.N., de Jong, P., Dermitzakis, E.T., Eichler, E.E., Flicek, P., Goldman, N., Mundy, N.I., Ning, Z., Odom, D.T., Ponting, C.P., Quail, M.A., Ryder, O.A., Searle, S.M., Warren, W.C., Wilson, R.K., Schierup, M.H., Rogers, J., Tyler-Smith, C., Durbin, R., 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483, 169–175.

Schipper, J., Chanson, J.S., Chiozza, F., Cox, N.A., Hoffmann, M., Katariya, V., Lamoreux, J., Rodrigues, A.S.L. Stuart, S.N., Temple, H.J., Baillie, J., Boitani, L., Lacher, T.E. Jr, Mittermeier, R.A., Smith, A.T., Absolon, D., Aguiar, J.M., Amori, G., Bakkour, N., Baldi, R., Berridge, R.J., Bielby, J., Black, P.A., Blanc, J.J., Brooks, T.M., Burton, J.A., Butynski, T.M., Catullo, G., Chapman, R., Cokeliss, Z., Collen, B., Conroy, J., Cooke, J.G., da Fonseca, G.A.B., Derocher, A.E., Dublin, H.T., Duckworth, J.W., Emmons, L., Emslie, R.H., Festa-Bianchet, M., Foster, M., Foster, S., Garshelis, D.L., Gates, C., Gimenez-Dixon, M., Gonzalez, S., Gonzalez-Maya, J.F., Good, T.J., Hammerson, G., Hammond, P.S., Happold, D., Happold, M., Hare, J., Harris, R.B., Hawkins, C.E., Haywood, M., Heaney, L.R., Hedges, S., Helgen, K.M., Hilton-Taylor, C., Ainul Hussain, S.A., Ishii, N., Jefferson, T.A., Jenkins, R.K.B., Johnston, C.H., Keith, M., Kingdon, J., Knox, D.H., Kovacs, K.M., Langhammer, P., Leus, K., Lewison, R., Lichtenstein, G., Lowry, L.F., Macavoy, Z., Mace, G.M., Mallon, D.P., Masi, M., McKnight, M.W., Medellín, R.A., Medici, P., Mills, G., Moehlman, P.D., Molur, S., Mora, A., Nowell, K., Oates, J.F., Olech, W., Oliver, W.R.L., Oprea, M., Patterson, B.D., Perrin, W.F., Polidoro, B.A., Pollock, P., Powel, A., Protas, Y., Racey, P., Ragle, R., Ramani, P., Rathbun, G., Reeves, R.R., Reilly, S.B., Reynolds, J.E.III, Rondinini, C., Rosell-Ambal, R.G., Rulli, M., Rylands, A.B., Savini, S., Schank, C.J., Sechrest, W., Self-Sullivan, C., Shoemaker, A., Sillero-Zubiri, C., De Silva, N., Smith, D.E., Srinivasulu, C., Stephenson, P.J., van Strien, N., Talukdar, B.K., Taylor, B.L., Timmins, R., Tirira, D.G., Tognelli, M.F., Tsytsulina, K., Veiga, L.M., Vié, J-C., Williamson, E.A., Wyatt, S.A., Xie, Y., Young, B.E., 2008. The status of the world's land and marine mammals: Diversity, threat, and Knowledge. Science 322, 225–230.

Schliep, K.P., 2011. Phangorn: phylogenetic analysis in R. Bioinformatics 27, 592–593.

Song, S., Liub, L., Edwards, S.V., Wu, S., 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. U.S.A. 109, 14942–14947.

Symonds, M.R.E., 2002. The effects of topological inaccuracy in evolutionary trees on the phylogenetic comparative method of independent contrasts. Syst. Biol. 51, 541–553.

Turvey, S.T., 2009. Holocene Extinctions. Oxford University Press, New York.

Turvey, S.T., Fritz, S.A., 2011. The ghosts of mammals past: biological and geographical patterns of global mammalian extinction across the Holocene. Phil. Trans. R. Soc. B 366, 2564–2576.

Yule, G.U., 1924. A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis. Philos. Trans. Roy. Soc. London Ser. B 213, 21–87.