A Phylogenetic Foundation for Comparative Mammalian Genomics

Peter J. Waddell[1] waddell@biol.sc.edu

Hirohisa Kishino[2] kishino@wheat.ab.a.u-tokyo.ac.jp

Rissa Ota[3] r.ota@massey.ac.nz

[1] Biological Sciences, University of South Carolina, SC 29208, USA
[2] Graduate School of Agriculture and Life Sciences, University of Tokyo, 1-1-1 Yayoi Bunkyo-ku, Tokyo 113-8657, Japan
[3] Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand

Abstract

A major effort is being undertaken to sequence an array of mammalian genomes. Coincidentally, the evolutionary relationships of the 18 presently recognized orders of placental mammals are only just being resolved. In this work we construct and analyse the largest alignments of amino acid sequence data to date. Our findings allow us to set up a series of superordinal groups (clades) to act as prior hypotheses for further testing. Important findings include strong evidence for a clade of Euarchonta+Glires (=Supraprimates) comprised of primates, flying lemurs, tree shrews, lagomorphs and rodents. In addition, there is good evidence for a clade of all placental mammals except Xenarthra and Afrotheria (=Boreotheria) and for the previously recognised clades Laurasiatheria, Scrotifera, Fereuungulata, Ferae, Afrotheria, Euarchonta, Glires, and Eulipotyphla. Accordingly, a revised classification of the placental mammals is put forward. Using this and molecular divergence-time methods, the ages of the superordinal splits are estimated. While results are strongly consistent with the earliest superordinal divergences all being > 65 mybp (Cretaceous period), they suffer from greater uncertainty than presently appreciated. The early primate split of tarsiers from the anthropoid lineage at $\sim$55 mybp is seen to be an especially informative fossil calibration point. A statistical framework for testing clades using SINE data is presented and reveals significant support for the tarsier/anthropoid clade, as well as the clades Cetruminantia and Whippomorpha. Results also underline our thesis that while sequence analysis can help set up hypothesised clades, SINEs obtainable from sequencing 1-2 MB regions of placental genomes are essential to testing them. In contrast, derivations suggest that empirical Bayesian methods for sequence data may not be robust estimators of clades. Our findings, including the study of genes such as TP53, make a good case for the tree shrew as a closer relative of primates than rodents, while also showing a slower rate of evolution in key cell cycle genes. Tree shrews are consequently high value experimental animals and a strong candidate for a genome sequencing initiative.

Keywords: phylogeny, mammalian evolution, placental classification, SINE, ancestral population size, BIC, likelihood ratio test

1 Introduction

At present there is great interest in which mammalian genomes should be sequenced next to join the nearly completed human and mouse genomes. This in turn has focused much of molecular biology on evolutionary questions, since the sequencing of multiple genomes implies comparative genomics, whereby the differences observed (and made use of) are due to evolution.

One of the key and fundamental evolutionary questions outstanding for placental mammals is what are the relationships of the 18 presently recognised orders [9, 24] and where and when did their evolution take place? Indeed, of the 18 orders, three have been identified and named (Cetartiodactyla, Afrosoricida, and Eulipotyphla) only in the last decade based on studies of molecular data. Whether the last of these three is monophyletic is still contentious [9]. Clarifying these questions will also lead to a stable phylogenetic (tree-like) classification of mammals and a framework on which all comparative mammalian genomics can be hung. This paper focuses upon these questions and then considers which genomes and experimental animals may be of greatest use for medical research.

Another vexing question is how old are the divergences within mammals? Can these be estimated reliably to give a framework for inferring the evolutionary rate of the many sorts of genes in the mammalian genome? More generally, these dates give clues to the ecological and environmental context within which early placental evolution occurred, e.g. were there still dinosaurs around, and on which continents did placentals first appear [14].

To answer such questions it is important to use the large amount of amino acid sequence data now available. Much of these data have not previously been analysed, yet include 8000 well-aligned positions sequenced in the vast majority of mammalian orders. These include all whole mtDNA genome sequences published to data, which constitute the longest sequences presently available [14, 23], plus a variety of nuclear data sets, including the longest sequences for a single nuclear gene, those of BRCA exon 11 [11]. This helps to address the rarity of amino acid based trees from nuclear sequences in recent studies. Amino acid sequences are potentially more informative than nucleotide sequences because there are more states and a slower rate of change, thus the possibility of less convergence and parallelism (long edge or branch attract effects). This is especially important since different groups of mammals show different nucleotide compositions indicating non-stationarity substitution, which is a major reason for estimated evolutionary trees being incorrect [20].

Even with amino acid sequences, as more data are concatenated, bootstrap support for clades rises to greater than (>) 90%, but different methods of tree reconstruction may then disagree strongly thus indicating error in at least one of them! Accordingly, the sequence data cannot be relied on to give an unambiguous interpretation of placental phylogeny. That is, they are useful for setting up hypotheses of relationships, but these will need to be tested with something other than sequence data. We also show herein that increasingly popular Bayesian approaches [27], which many hope will lead to more reliable estimates of phylogeny, may be unreliable guides to the significance of clades.

A type of data that is becoming increasingly popular for the analysis of evolutionary relationships is the random insertion of nuclear elements [6, 13]. Apparently, unique insertions of these genes are not precisely deleted or erased, so they should not suffer from long edge attraction problems. However, like all data they are subject to the effects of ancestral polymorphism or gene sorting. It is an important outstanding problem to understand how they may be statistically analysed [6].

Following analyses considering both congruence and concatenated data, an “expert opinion” (or EO) of placental phylogeny is presented. (An EO is a statistical term used to describe an informed summary of analyses, which may act as the foundation for such things as Bayesian priors). The revised classification of placental mammals may then serve as priors for the statistical testing of evolutionary hypotheses with SINE data. A statistical framework is developed to use SINE data to test clades. These tests should be both accurate and unbiased. The SINE test framework can also be extended to make estimates of historical features such as the population size of our distant ancestors, for example those at the time of our last common ancestor with mice more than 65 million years before present (mybp).

2 Methods and Results

Amino acid sequences, which were sequenced in the vast majority of placental orders, were retrieved, aligned, and then edited to remove regions of ambiguity. The major data sets were whole mtDNA sequences [23], BRCA1 exon 11 [11], IRBP, vWF and A2AB [11], fragments of 12 different nuclear genes [12] and a-crystallin [8]. The resulting data sets and lists of Genbank accession numbers are available upon request (see the GIW 2001 website for additional information).

Trees were estimated using a wide variety of techniques [20, 23]. From past and present experience, the most generally useful results were obtained using ML tree search. In this case the ProtML program [1] with options f plus m or j as appropriate, and to avoid being trapped in local optima (which were common with all these data sets) the searches were seeded with best trees found by at least 5 methods that gave topologically distinct trees. These included parsimony, minimum evolution, and least squares trees found using a TBR search in PAUP* [19, 20]. Support on these trees was estimated using local bootstrap proportions or LBP [1]. PAUP* was also used to calculate the consistency index (CI) and retention index (RI) of the data, with higher RI values being the more general indicator of the quality of the data. Note, that in comparing trees, the comparisons are to the tree of Waddell *et al.* [24], this being the first superordinal classification to show substantially the newly emerging consensus of opinion [9].

2.1 The Phylogenetic Tree of Placental Mammals

A significant problem is the tendency to either ignore the mtDNA data or to place too much emphasis upon it. Apart from the hedgehog sequence, trees based on mtDNA sequences are most confused by a handful of taxa that are evolving at a high rate and with transition probabilities that are not typical of other placentals. Examples identified so far are the hedgehog [18] and murids [23]. Using the methods in [23] we identify tenrec and elephant as probable additions. Accordingly, a mtDNA amino acid tree (not shown) of all mammal sequences in Genbank by July 2001 is typically incongruent and bizarre. The murids and the hedgehog/gymnure are strongly attracted to the root. By removing the outgroups and the murids, but retaining the hedgehog and gymnure sequences, we obtain a ProtML tree with $>$ 99% LBP for both Laurasiatheria and Supraprimates, assuming the root is closer to Afrotheria or Xenarthra (for the definitions of superordinal groups, see Figure 3 and [24]; to associate ordinal names such as Xenarthra with common representatives, see Figure 1(B)). Without the outgroups and murids, the hedgehog and gymnure move into close proximity to the mole and shrew sequences, and in doing so, also restore monophyly to the rodents. This newfound congruence with [24] is highly unlikely unless the hedgehog and gymnure mtDNA were badly misplaced [18, 23].

It is desirable to go further with the mtDNA data by removing additional taxa which show significant uncertainty in their positions, based on either the mtDNA data or a strong clash with other well established sources of data, including nuclear DNA sequences and SINE data. (Also removed are congeneric sequences to reduce run time). The resulting ProtML tree is shown in Figure 1(A). This tree agrees very well with the tree in Waddell *et al.* [24]. Features of note include the Afrotheria (represented by aardvark) as the first diverging placental, followed by the Xenarthra (armadillo), and then the two major groups of Laurasiatheria and Supraprimates. Such groups have not previously been found all together with the mtDNA data. The congruence of this result with other data sets and the generally high retention index of the data, reinforce the hypothesis that it is a handful of taxa with high and/or atypical substitution rates that throws the mtDNA tree into chaos.

The tree of the amino acid sequences for BRCA is shown in Figure 1(B). This tree too is highly congruent with the trees of [9, 11, 12, 24] and that of Figure 1(A). It serves as an independent test of these trees, in particular that of [24], which alone qualifies as a prior hypothesis. The apparent reason for disagreement with the prior hypotheses is the single distant outgroup being attracted to the murids. In fact, the outgroup wanders around this tree considerably and on the second best local optima found, implies Afrotheria sister to all other mammals. Note that the RI on the nuclear and the mtDNA data sets are very similar, suggesting that at the amino acid level all are approximately equally valuable (and similarly prone to errors). Note the substantial support within Afrotheria for a Tubulidentata (aardvark), Macroscelidea (elephant shrew), and Afrosoricida (golden mole, tenrec) clade.

Figure 1(C) shows the tree based on the concatenated amino acid fragments [12]. Notable are the strong support for Laurasiatheria, Afrotheria and Supraprimates, and for a sister relationship of

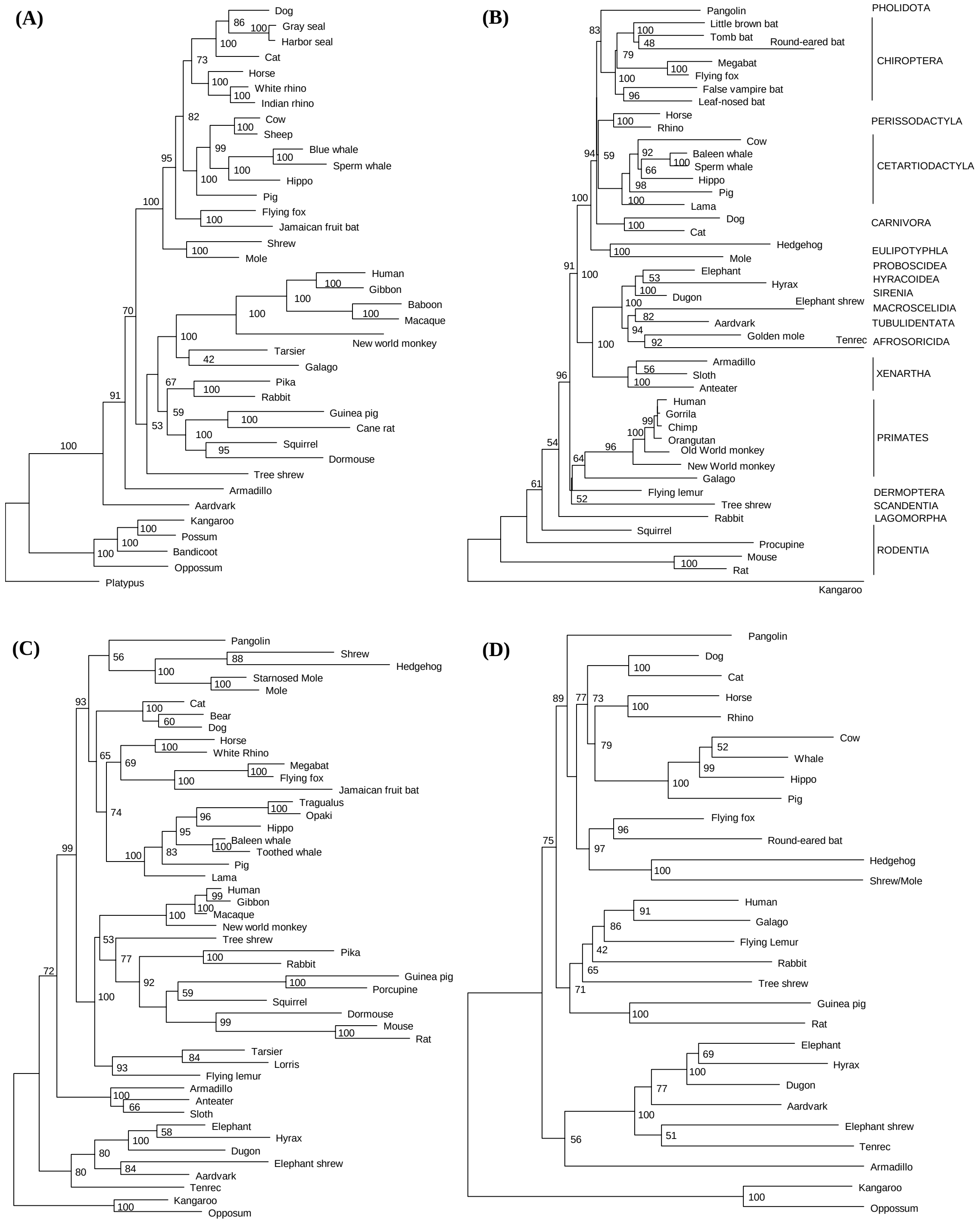

Figure 1: ProtML trees for 4 major amino acid datasets (numbers are local bootstrap proportions for edges): (A) mtDNA encoded proteins (3516 residues, CI=0.43, RI=0.47); (B) BRCA (925 residues, CI=0.54, RI=0.47); (C) Multiple concatenated fragments (2020 residues, CI=0.49, RI=0.50); (D) Concatenated IRBP, vWF, A2AB (1222 residues, CI=0.55, RI=0.44).

the last two superorders mentioned (Boreotheria). Like the BRCA data, there is little doubt that hedgehog, shrews and moles form a monophyletic group, supporting the order Eulipotyphla.

The tree of Figure 1(D) is derived from the concatenated proteins IRBP, vWF and A2AB. The tree again shows strong congruence to [24] from which it is not wholly independent, since IRBP and vWF were datasets used to infer that classification. Not shown is the tree from α-crystallin with orders constrained to be monophyletic. This tree contains major features consistent with Boreotheria, Afrotheria, and Laurasiatheria. It too was used to infer the classification of [24].

The tree of all concatenated nuclear sequences is shown in Figure 2(A). This data set is largely independent of the tree of [24] (and by exclusion of α-crystalline, IRBP and vWF comprising $<$ 20% of the data, totally independent). It gives strong support to Laurasiatheria, Supraprimates and Boreotheria, while the relationships within these groups are totally congruent with [24] excepting the association of Perissodactyla and Cetartiodactyla (identified in [24] as a local alternative). This last grouping is increasingly being found with the mtDNA data also, and we suspect that it is indeed correct, not least since it requires only one origin and no major loss within the Laurasiatheria of a host of ungulate-like features. The relationships within the Afrotheria remain contentious except for the strong support for Paenungulata. The association of aardvark, elephant shrew and tenrec is not inconsistent with the mtDNA data. It does, however, imply a separate origin of the ungulate-like features that aardvarks have (implying at least three such derivations from apparently shrew-like ancestors).

Figure 2(B) shows the ML tree for 7999 well aligned nuclear and mtDNA encoded sites. The tree is again nearly totally congruent with that of [24]. Not only is it congruent, but bootstrap support is 90% or more for nearly all the previously named clades. Note again the strong support for Perissodactyla plus Cetartiodactyla. One of the few groups to drop support is Ferae, perhaps due to pangolin alone having only a partial mtDNA sequence; something to be rectified. Within Supraprimates, there is a lack of resolution regarding exactly where the tree shrew goes. Of note also is the resolution appearing within Rodentia, with the squirrel diverging before the hystricognath/murid group. The mtDNA data would further imply that the dormice are sisters to the squirrels and not murids as commonly assumed. Within the primates, the traditional hypothesis of tarsier being sister to the galago and lemurs is supported, in contrast to the prevailing view that places tarsier closer to the human (anthropoid) lineage [3, 16, 28]. We are near certain the former is an error in the amino acid tree (which also occurs when analysing DNA sequences), as is explained below after analyzing SINE data. At the very root of the placentals, Xenarthra are implied to diverge first with moderate support. However, this may be a taxon-sampling problem. At present there is only one Xenarthran mtDNA sequence, resulting in a longer single edge than is present with the Afrotheria. One might suspect long edge attraction to the outgroup. Retaining just the slowest evolving of the afrotherians, the aardvark, tests this possibility. Rerunning the analysis, moderate support is found for the Afrotheria as sister to all other placentals (Figure 2(C)).

Where the larger data sets are ambiguous, smaller datasets enhance the resolution. One example is the TP53 gene tree shown in Figure 2(D). The outstanding result here is the strong evidence for the association of tree shrew with primates (consistent with Euarchonta) and sister to this the Glires clade (with ML the tree shrew is weakly supported as closest to humans, with parsimony, the human goes with monkeys). This dataset is very clean; that is, overall it appears to be highly reliable with few convergent or parallel amino acid changes as seen by the RI of 0.76. There are 12 unambiguous substitutions on the most parsimonious tree consistent with Euarchonta, but 0 substitutions on the edge grouping tree shrew and rabbit on the most parsimonious tree consistent with this constraint. This is in contrast particularly to the mtDNA data where there are many characters supporting each of these two mutually exclusive groups. If one accepts the general scientific argument that statistical measures of phylogenetic support, such as the bootstrap, need to be weighted by quality and not just quantity of data, then TP53 shifts the posterior probability strongly in favour of Euarchonta.

In summary, the amino acid trees for concatenated mtDNA sequences, nuclear sequences, and

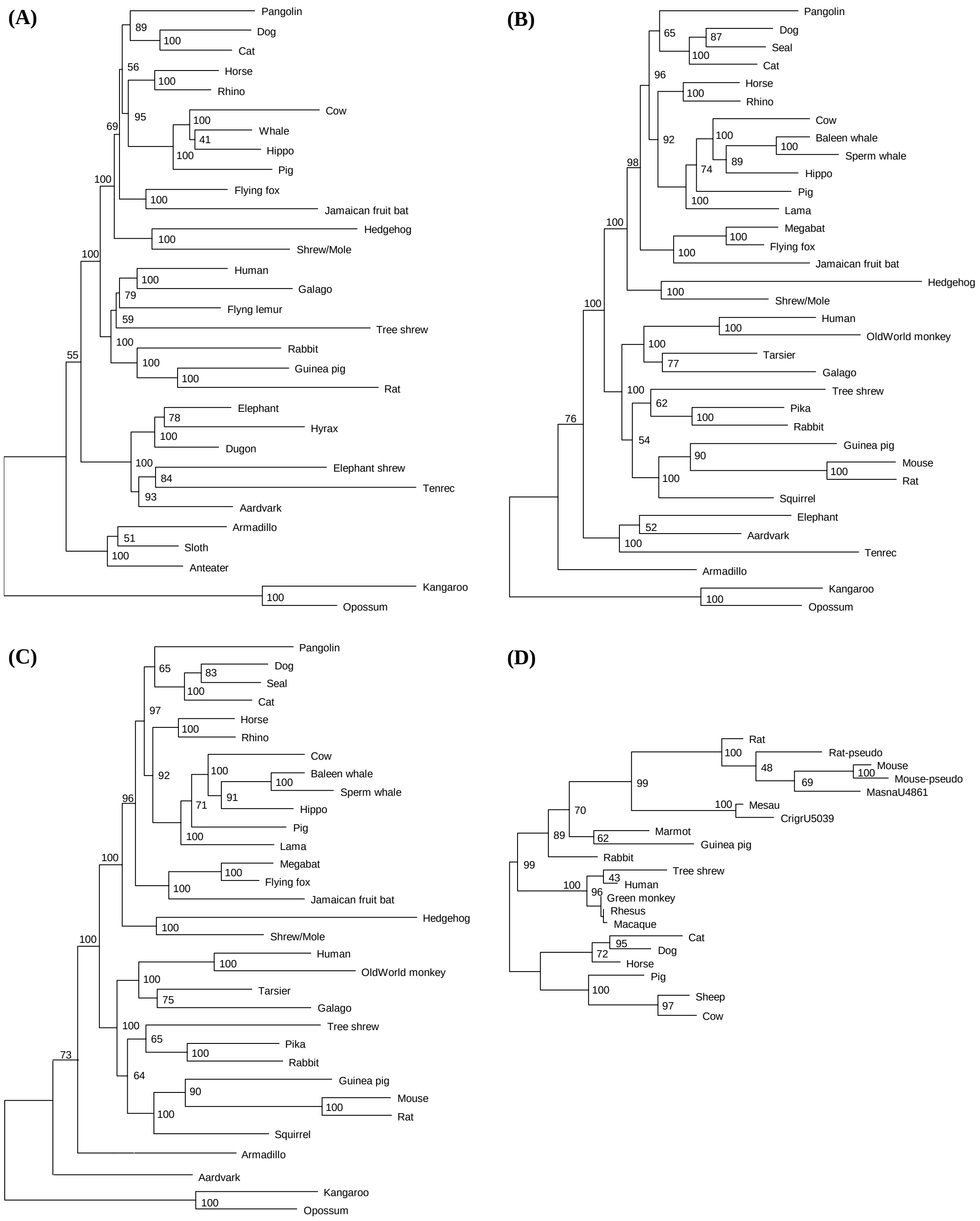

Figure 2: ProtML trees for concatenated data and TP53: (A) Concatenated nuclear encoded proteins (4310 residues, CI=0.56, RI=0.39); (B) Concatenated nuclear plus mitochondrial proteins (7999 residues, CI=0.49, RI=0.38); (C) As previously except minus the elephant and tenrec sequence (7999 residues, CI=0.51, RI=0.39); (D) Tree of TP53 (381 residues, CI=0.76, RI=0.76), where "pseudo" indicates pseudogene sequences.

both combined give good resolution plus strongly congruent results with each other and the prior work of [24]. Such results are also generally consistent with DNA based analyses [11, 12]. However, no matter how strong the support for clades is, sequenced-based analyses of this type of data seem incapable of giving a final and definitive solution to the phylogeny of the placentals, as the example of the tarsier/galago association will show. This amino acid analysis, through congruence particularly, gets us to the point of setting up a probable and restricted set of superordinal clades.

2.2 Why Empirical Bayesian Methods May Falter

It is often assumed that an edge in a tree with a high level of support is a suitable basis for a classification. In contrast, our own approach in [24] was to put forward a classification based on EO, where values such as bootstrap proportions and empirical Bayesian posterior probabilities are only guides. This is because statistical support measures may be contradicted on different data sets or taxon sampling and are always subject to the validity of the model (which is always erroneous to some degree). This is somewhat appreciated for bootstrap values, but not so much for posterior probabilities.

Here we show the close relationship between likelihood ratio values and posterior probabilities for these types of phylogenetic problem. BIC was developed to help choose amongst competing models [17], here selecting the tree with the highest BIC score. The BIC score for a tree T_i is given by

$$\mathrm{BIC} \quad = \quad \ell(\hat{\theta}_i|T_i,\mathbf{X}) - \frac{m}{2}\ln(n) \tag{1}$$

where $\ell(\hat{\theta}_i|T_i,\mathbf{X}) = \ln L(\hat{\theta}_i|T_i,\mathbf{X})$ is log of the maximum likelihood function, $\mathbf{X}$ the data matrix (of dimension $n \times r$), n is the amount of data, r is the number of sequences, m is the number of parameters in the tree, and θ the vector of all parameters. The usual priors of the Bayesian method do not appear explicitly in its formula, since the prior is considered equal for all topologies.

Using the Laplace approximation, and assuming that the determinant of the information matrix is constant across trees, we obtain

$$\int L(\theta_i|T_i,\mathbf{X})\pi_i d\theta_i \quad \propto \quad P(T_i,\mathbf{X}) \tag{2}$$

where p_i is the prior of the tree T_i, $P(T_i|\mathbf{X})$ is the posterior probability. To be explicit, in this use of BIC the posterior probability is estimated from the likelihood by

$$p_i \quad = \quad \frac{\exp(\ell(\hat{\theta}_i|T_i,\mathbf{X}) - b_n)}{\sum_{j=1,\ldots,k}\exp(\ell(\hat{\theta}_j|T_j,\mathbf{X}) - b_n)} = \frac{L(\hat{\theta}_i|T_i,\mathbf{X})}{\sum_{j=1,\ldots,k} L(\hat{\theta}_j|T_j,\mathbf{X})}, \tag{3}$$

for all $i = 1,\ldots,k$, where $b_n = m\ln(n)/2$. Here, the penalty b_n is the same for all topologies (or histories, see below, if they are the basic unit), because we are considering only resolved trees (or histories). Accordingly, a difference of as little as $2\ln L$ units between the best tree and all others may result in the clades on the best tree having posterior probabilities over 90%.

In calculating the posterior probabilities of tree topologies, we need to consider how we want to treat histories (tree topology plus information on the order of branching events). A fully asymmetric tree, such as (((a, b), c), d), can be generated by only one coalescent history, where a and b coalesce first, then c, and finally d. In contrast, balanced trees like ((a, b), (c, d)) can be generated by two coalescent patterns. Since the probability of each history is equal under the coalescent model, then balanced 4 taxon trees will have a prior probability twice that of those which are asymmetric. However, if we condition on the history (and so measure the maximum likelihood of the data for each distinct history), then the priors will once again be equal, while the posteriors for rooted tree topologies will be the sum of the posteriors of the histories they are associated with. In Table 1 we compare BIC methods with the Bayesian results reported in [27]. The BIC variants are much more similar to the Bayesian MAP than the RELL bootstrap results, despite sequences being only a few thousand positions long.

Our main point here is to note that Bayesian posteriors can be linked to a simple likelihood difference. However, if the data do not fit the model, then, as with edge length tests of the significance of an edge, there is a tacit assumption that conflicting edges on competing trees are of length near zero. Failure of this assumption may lead to far too much confidence in these posterior values. Accordingly, application of such methods to different datasets for mammals can lead to strongly contradictory results. Bayesian methods may therefore be less robust to model violation than bootstrap methods.

Table 1: Comparison of posterior probabilities calculated by BIC and MCMC with RELL bootstrap values for the data set used in [27] (RELL bootstrap values were close to full bootstrap values).

tree	clock ML	BIC posteriors		MAP posteriors [27]		RELL
topology[a]	(log likelihood)	BIC_1	BIC_2	HBA[b]	EBA[c]	
T_1	−5250.37	0.881	0.899	0.955	0.957	0.690
T_4	−5253.08	0.059	0.060	0.032	0.029	0.084
T_7	−5256.36	0.002	0.001	0.002	0.002	0.221
T_9	−5253.08	0.059	0.040	0.011	0.011	0.005

[a] Index of the tree topologies follow [27], with topology posteriors being the sum of related history posteriors for BIC_2(e.g. T_1 is a sum of the posteriors for histories 1 to 3) or just the best history for BIC_1. [b] HBA, hierarchical Bayesian analysis; [c] EBA, empirical Bayesian analysis.

2.3 A Revised Phylogenetic Classification of Mammals

Figure 3 shows our EO of the phylogeny of extant placental groups along with our uncertainty in each clade (c.f. [24]). These are intended to reflect our Bayesian priors for these groups in expectation of further data being gathered. All superordinal taxa in Figure 3 are defined as crown based groups. Divergence times, as in [24], are EO's based upon reanalysis of the majority of the published data with applicable methods (see below). The newly named groups are: Supraprimates (meaning beyond primates). Boreotheria (or therian mammals of the northern areas) for the clade composed of Supraprimates plus Laurasiatheria. Euungulata (true ungulates) for Cetartiodactyla plus Perissodactyla. Exafroplacentalia (placentals excluding those in Africa) for all extant placental orders except those within Afrotheria. The clade composed of elephant shrews and tenrecs is named the Afroinsectivora (African insectivores), while this clade plus the aardvark is named the Afroinsectiphillia (African insect eaters/lovers).

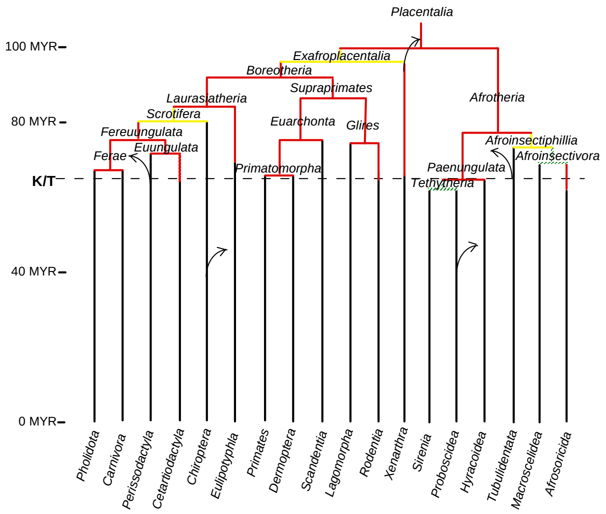

Figure 3: Our current best estimate (EO) of the placental tree. Dashed lines indicate that a clade is preferred, but has less than 50% probability of being correct, light grey lines have a 50 to 70% probability, and dark grey indicates > 70% probability. Arrows indicate local rearrangements that might yet be correct.

As detailed in [24] there are some strong local alternatives to parts of the classification. These represent alternative and independently distinguishable statistical hypotheses, three of which remain unnamed. In order to give them prior status for future testing, they too should have names. The first is Insectiphillia (insect lovers) for the crown group of Eulipotyphla (core insectivores) and Chiroptera (bats). The second is the association of elephants and hyraxes that we name the Hyracoproboscidea (the fusion of the ordinal names). The statistical advantage of naming the alternative clades in unambiguous terms is that they too may then form specific *a priori* hypotheses when more data are gathered.

2.4 The Age of Mammalian Groups and the Pace of Molecular Evolution

Another area of intense interest at present is when did the first extant placental orders originate and how much earlier did the superorders appear. To make estimates of times, the ProtML tree of the combined sequences is modified to fit our best estimate of the placental phylogeny in Figure 3. Edge lengths are then reestimated using the PAML program [26] allowing for a G distribution of site rates to allow for more accurate divergence time estimates [20, 25]. Next, the non-parametric method of Sanderson is used to infer the relative divergence times using the program TreeEdit [15]. Options were to infer the rates on each side of the root independently and infer the rate at the root as the average of the two descendant lineages. The resulting values give the relative times listed in Table 2.

In contrast to the usual assumption that fossil dates are minima, fossil dates behave more like sampled-points from a continuous distribution. If the fossil shows true synapomorphies with a lineage, then it is a minimum date (subject still to over estimation of its geological age). A subjective judgement needs to be made regarding how long it took to accumulate such synapomorphies. It can become an overestimate if it does not contain true synapomorphies, but rather homoplasies. The following fossil divergence times (e.g. [10]) are amongst the oldest calibration points in Placentalia and should be best for dating the earliest splits:

(1) The horse/rhino split. Close to 55 mybp (million years before present) followed soon after by the rhino/tapir split. This divergence sets the standard for fossil calibration data close to the age of placental ordinal divergences [22] and an approximate 95% confidence interval would be 52-58 mybp.

(2) The tarsier/anthropoid: There is now good evidence that these are sister taxa, supporting the case in [3] that these taxa split close to the Eocene/Paleocene boundary at 55 mybp. An approximate confidence interval would be 50-60 mybp.

(3) The whale/hippo split. Whale fossils date back to ∼52 mybp, but there is no clear cap on the maximum age. A confidence interval of 49-61 mybp would seem reasonable [21].

(4) The rabbit/pika split. Suggested to be middle Eocene (∼40-45 mybp), with a wide confidence interval of ∼36 to 55 mybp due to a patchy fossil record and questions regarding ancestral states.

Other calibration points are less developed. Calibration points within rodents are worth considerably more investigation, not least since distinct rodent teeth should be identifiable from the late Cretaceous if suborders of rodents were established then. There is increasing molecular evidence for a monophyletic Histricognatha sister to murids (excluding dormice), and Sciurida diverging beforehand. Possible members of suborder Sciuromorpha date back to ∼60 mybp, but rely on the uncertain placement of taxa such as beavers closer to squirrels than mice. The oldest members of Sciurida appear to be 50-55 mybp, of murids ∼50 mybp, of dormice ∼60 mybp and of hystricognaths ∼50 mybp. The proboscidian lineage is minimally 55 mybp and possibly ∼58-60 mybp. An armadillo scute at 61 mybp suggests a minimum age for Xenarthra. It is known that there are bat fossils around at 55 mybp, and possibly as old as late Paleocene (∼60 mybp), but exactly which extant lineage they are closest to and what is the phylogeny within bats are still contentious. There are many more fossil calibrations in the late Eocene and Oligocene (after 37 mybp), but the later in time the less attractive for our purpose.

Table 2 shows the estimated ages of various groups using the four best calibration points. Note that these give quite different estimated ages, particularly for clades distant to the calibration point in both

Table 2: Ages of placental crown groups implied using the HR (horse/rhino split at 55 mybp), WH (whale/hippo split at 52 mybp), TH (tarsier/human split at 55mybp) and RP (rabbit/pika split at 42 mybp) calibration points. In bold are nodes with direct fossil ages, while underlined are independent estimates of these ages using molecular methods.

Clade	HR	WH	TH	RP	Clade	HR	WH	TH	RP
Anthropodea	49	61	34	34					
tarsier/human	79	99	**55**	56	Perissodactyla	**55**	69	38	39
Primates	83	103	57	59	Cetartiodactyla	60	75	42	43
Euarchonta	92	114	64	65	Artiofabula	55	68	38	49
Glires	87	109	61	62	Cetruminantia	47	59	33	33
Rodentia	78	98	54	56	Whippomorpha	42	**52**	29	30
hystricognath/murid	68	85	47	48	Cetacea	20	25	14	14
Lagomorpha	59	74	41	**42**	Ferae	76	95	53	54
Supraprimates	94	117	65	67	Carnivora	55	68	38	39
Laurasiatheria	91	114	63	65	Caniformia	45	56	31	32
Eulipotyphla	79	99	55	56	Boreotheria	101	126	70	72
Scrotifera	86	107	59	61	Exafricomammalia	110	137	76	78
Chiroptera	68	85	47	48	Placentalia	115	143	80	82
Fereuungulata	81	101	56	57	Afrotheria	98	123	68	70
Euungulata	77	96	53	55	Afroinsectiphilla	91	114	63	65

time and place on the tree (the local relative ages agree quite well with [4, 22, 24]). Shifting the fossils to the extremes of their probable ages cannot eliminate these differences. The calibration point involving tarsier is most conservative, and that involving whales the most liberal. The implied age of Placentalia is vastly different as the calibration point is changed. What is clear is that estimating the age of the earliest splits within Placentalia is fraught with uncertainty until: (a) Fossil calibration points become more accurate. (b) Some fossils can be placed accurately amongst the superordinal splits (e.g. [2]). (c) Models for converting sequence data to relative times on a clock-like tree improve. We are skeptical of how accurately the latter can be achieved. This is because the underlying rate of synonymous substitution will be strongly correlated across genes for a taxon and strongly autocorrelated with the lineage. However, the rate of non-synonymous substitution will be dominated by selective pressures, which we suspect will be generally uncorrelated between most genes and weakly autocorrelated with lineage effects. Neither condition suits the use of an average for multiple genes to make a more accurate estimate, which ideally involves uncorrelated rates between genes, yet each showing good autocorrelation through time.

That there were splits between orders prior to the KT boundary seems indisputable, but exactly how old they are is very much contentious. Note that the calibration points giving the older dates occur along lineages where there was a early increase to very large body size plus long generation times. If the evolutionary model, including Sanderson's method, is underestimating the extent of rate slowdown in the lineages leading to whales and horses, then this could explain the much older dates these calibration points are implying. Given that the calibration points in Primates and Lagomorpha are moderately consistent, it would seem probable that Primates and Rodentia originated close to the KT boundary. Any suggestion that crown groups such as these are much older than 75 million years [14] would appear problematic, although probable dormice relatives at ~60 mybp (and dormice not being the deepest split in rodents) suggest that Rodentia could be 65-70 mybp.

2.5 Statistical Testing with SINE Data

A proper statistical analysis of SINEs offers the best hope for accurately testing the validity of any proposed superordinal clade. Fortunately, sequencing approximately 1-4 MB of (non-rearranged)

DNA from many orders should generate enough SINEs and LINEs for this purpose. The former are especially interesting since they have no known mechanism of exact deletion or excision.

The test described next can be applied to any rooted subtree of three taxa. Assume a standard Wright-Fisher coalescent model, with panmictic mating, non-overlapping generations, and a constant population size. Then the equations in [7] imply that with a trichotomy, all three resolutions will have equal support. It follows that in order to resolve a trifurcation, the hypothesis that the SINE frequencies are consistent with an equal number of SINEs being shared by all three pairs of lineages must be rejected. Let $T^{(s)}$ denote the species tree. Let there also be an unambiguous *a priori* preference for just one of the three resolved species trees, here taken to be T_1, which corresponds to the SINE pattern 110, for taxa A, B, C, respectively (so A and B are sister taxa). The first hypotheses to test are: (H_0): $T^{(s)} = T_0$ (the star tree), (H_1): $T^{(s)} = T_1$. It is also necessary to specify the probability of a Type I error, α, that is, the probability of the test rejecting H_0 in favour of H_1 when H_0 is indeed true.

The above hypotheses can be respecified in terms of the relative frequencies of SINE patterns S_1 to S_3, that SINEs support the trees T_1 to T_3 respectively. That is, H_0: $S_1 = S_2 = S_3$, H_1: $S_1 > S_2 = S_3$, where the pattern 110 supports T_1, the pattern 101 supports T_2, and the pattern 011 supports T_3. The test procedure proposed is: (1) Set α (e.g. at 0.05). (2) Estimate the maximum likelihood of the data under: (a) H_0 (b) H_1 (3) Evaluate the following log likelihood ratio for the observed data: i.e. $\ln L(\mathbf{X}|H_1) - \ln L(\mathbf{X}|H_0)$ (Where, $\ln L(\mathbf{X}|Y)$ is the maximum log likelihood of the observed data, $\mathbf{X}$, under hypothesis Y) (4) Evaluate the distribution of 3(a) under H_0, and 3(b) under H_1. (4) Reject H_0 at level α if $P < \alpha$, where P equals the proportion of times observed statistic 3(a) is larger than the distribution of 3(a) under H_0.

As an illustrative example, consider the case of data, collected to test the hypothesis that T_1 Whippomorpha (the association of whales and hippos [24]) is the species tree. From [13] this results in 3 SINEs supporting T_1, 0 supporting T_2, and 0 supporting T_3, so the total number, n, is 3. For samples of less than 5 SINEs, it is necessary to calculate likelihood ratios directly, with large numbers of SINEs a $1/2\chi^2$ distribution with d.f.=1 may be used to assess the significance of the likelihood ratio. Numerical enumeration of the distributions of these statistics is made using standard combinatorial formulae and the results for up to 5 SINEs are presented in Table 3. For the example data of [3 0 0] (Table 3), $P = 0.0370$ so we reject H_0 at $\alpha = 0.05$ (see Table 3, index 4). Similar testing of the taxon Cetruminantia (or ruminants plus whippomorphs exclusive of all other cetartiodactyls) is also accepted at the 0.05 level as the SINE counts are [3 0 0] also. However, the clade Artiofabula (or pigs and relatives plus cetrumminants) has support of [1 0 0] and the probability of this by chance is 0.33. Accordingly, more data is needed to test this clade further. In the above examples, the priors were established by [5].

As another example of the test, consider the question of tarsier affinities. The sequence data tend to support the old prosimian hypothesis of tarsiers being closer to lemurs and galagos than to humans. However, there is now evidence of 3 SINEs shared between humans and tarsier that are not present in prosimians (and none that are contradictory [16, 28]). Accordingly, the new test rejects the alternatives to the tarsier/anthropoid grouping with 95% confidence. This example is a paradigm of why we feel that sequence analysis is, of itself, insufficient to critically test hypotheses of placental relationships; sequences can set up hypotheses, but SINEs and an appropriate statistical framework are required to test them.

The explicit likelihood test statistics introduced above are ideal for the new era of genomics. Using SINEs it is possible to marry phylogenetics and population genetics together using the framework described in [21]. As an example, assume that the tree is ((human, mouse):85,cow):90 with the numbers specifying the divergence times in mybp. Each mammalian genome contains ∼1,000,000 recognizable SINEs. Given that SINEs are identifiable back about 250 million years on average, then, $\sim 1,00,000/250 = 4000$ SINEs are fixed in the genome per million years. Thus, the common Human-Mouse ancestral lineage, exclusive of the cow, would have about 5myr×4,000/myr=20000 identifiable

SINE events. Thus, the sequencing of the first 1/2 of 1% of the cow genome, or 30 million base pairs, should yield $> 14000 \times 0.005 = 70$ informative SINEs (assuming 30% are unidentifiable due to subsequent evolution). Hypothetically, let the resulting data be: Mouse-Human, 50 SINEs, Human-Cow 8 SINEs, and Mouse-Cow 6 SINEs. By the test described above, the data would confirm the tree as mouse human. The 14 SINEs not fitting the species tree are no significantly different in frequency in the two other configurations and are therefore expected to be due to ancestral polymorphism, and the observed proportion, $\hat{P}$, of these is 14/64.

Table 3: Cumulative P values for H_0 vs H_1, regarding patterns of SINE frequencies, for $n = 1$ to 5. Here, T_1 is specified *a priori* and if ever $T_1 < T_2$, T_1 is not expected to be the species tree so P values are not shown.

n	Index	T_1^*	T_2	T_3	P	n	Index	T_1^*	T_2	T_3	P
1	1	1	0	0	0.3333	4 cont.	9	2	2	0	0.2840
2	2	2	0	0	0.1111		10	2	1	1	0.2840
	3	1	1	0	0.3333	5	11	5	0	0	0.0041
3	4	3	0	0	0.0370		12	4	1	0	0.0247
	5	2	1	0	0.1481		13	3	2	0	0.1481
	6	1	1	1	0.3704		14	3	1	1	0.1481
4	7	4	0	0	0.0123		15	2	2	1	0.2716
	8	3	1	0	0.0617						

Analysing sequence data will also inform us that the internal edge of our rooted mouse/human/cow tree is ∼5 myr (call this t_{div}). Since the common ancestor of these mammals was most likely a shrew-like insectivore (based on fossil evidence and ancestral reconstructions on the tree of [24](Figure 3), and since most such animals generally have one or two generations per year (so $t_{gen} = 0.5$ or 1 year), the internal edge represents about 5-10 million generations. From [7] we have the equation,

$$P = \frac{2e^{-t}}{3}\left\{\frac{1}{t + e^{-t}}\right\} \qquad (4)$$

where t, the sole variable, is measured in units of $2N$ generations when dealing with an autosomal locus in a diploid (N being the ancestral population size). t can, in turn, be expressed as $t_{div}/(t_{gen}2N)$. Since $\hat{P} = 14/64 \sim 0.219$ (the proportion of SINEs which do not follow the species tree), we can then solve for $\hat{t}$ numerically giving, $\hat{t} = 0.863$. Since $\hat{t} = 0.863 = \hat{t}_{div}/(\hat{t}_{gen}2N)$, then, $\hat{N}_{ancest} = \hat{t}_{div}/(2\hat{t}_{gen})$ which, yields $\hat{N}_{ancest} = 5 \times 10^6/(2 \times 0.863 \times 1) = 2.90 \times 10^6$ (or 5.79×10^6 assuming the shorter generation time). (Implicit assumptions are that SINEs do not insert at the same site in independent lineages and that SINEs decay at similar rates in each lineage. Violation of the last condition may distort ratios of patterns detected, which can be countered if necessary with modeling of the evolution of SINEs). Thus, SINE data and genomic sequencing will allow us to estimate the population size of ancestors all over the mammalian tree, including our own.

3 Discussion

3.1 Implications for Medical Genomics

Having a reliable classification of the Placentalia is important for genomics and for medical genetics. One of the points established by recent work [4, 5, 9, 11, 12, 24] has been the dissolution of many mistaken hypotheses of early placental evolution, such as the belief primates are closer to cows than they are to rodents, or that primates are closer to bats than just about anything else except tree shrews. The ideal animal model for medical genetics should combine a close genetic similarity to humans with practical advantages such as small size, high fecundity and short generation time. Mice meet the last of the above criteria admirably and their phylogenetic relatedness to primates bodes well. However, they do not meet the genetic similarity criteria well due to having a very high rate of mutation/substitution in murids (which includes rats and hamsters also, e.g. Figure 2(D)). Accordingly genes in the cell cycle, which is central to studying cancer genetics, are often more unlike humans than in just about any other placental and this is especially true in the key cancer suppressor gene TP53.

Rodents other than murids apparently do not suffer the same problems as murids. That leaves the guinea pig as a possibility and rabbits also, although their size is a concern. The TP53 gene not only confirms that tree shrews are closer to primates than to rabbits, but it also suggests that there was both a burst of positive selection on this gene in early euarchontan evolution followed by a general slow down in evolutionary rate. Accordingly, the tree shrew may be an ideal animal for cancer genetics, given its generally good handling characteristics, which are far better than those of any primate. Thus, sequencing its genome concordant with its expansion as an experimental animal should be given serious consideration.

The revised phylogeny of placentals in Figure 3 is expected to be an advance on previous work, and sets up a limited set of prior hypotheses for testing with SINE data. This phylogeny, like previous ones, [9, 11, 12, 24] emphasizes the occurrence of shrew-like forms across the tree. However, we predict that a number of apparent pre-Cretaceous superordinal groups, particularly Glires, Euarchonta, Supraprimates and Euungulata, should be directly recognizable from fossils. This is because they have common morphological features likely distinct from earlier ancestors. A central Asian fossil dating from ∼85 to 90 mybp claimed as a sister to Glires (D. Archibald pers comm.) may support this prediction. That the first placentals lived on Southern continents (e.g. Africa and South America) is feasible given mammals from northern areas (Boreotheria) are a compact group [24] and the most recent phylogenies [11, 12](Figure 3) suggest uniquely southern taxa (Xenarthra, Afrotheria) were the first to diverge.

Acknowledgements

Thanks to the Laboratory of DNA Information Analysis, IMS, U. of Tokyo, the Marsden Fund of New Zealand, and Monbusho Japan for funding in support of this research, to the staff of the Miyano lab, Debra Ernie, and the folks for making this work possible, to David Archibald, Jeff Thorne and Austin Hughes for comments, Chris Beard and Mary Dawson for discussions on dates, and Mike Sorenson for a number of the above plus help in preparing datasets and alignments.

References

[1] Adachi, J. and Hasegawa, M., MOLPHY: version 2.3: *Programs for Molecular Phylogenetics Based on Maximum Likelihood*, Comput. Sci. Monogr. 28, Institute of Statistical Mathematics, Tokyo, 1996.

[2] Archibald, J.D., Fossil evidence for a late Cretaceous origin of "hoofed" mammals, *Science*, 272:1150–1153, 1996.

[3] Beard K.C., Krishtalka L., and Stucky R.K., First skulls of the early Eocene primate *Shoshonius cooperi* and the anthropoid-tarsier dichotomy, *Nature*, 349:64–67, 1991.

[4] Cao, Y., Fujiwara, M., Nikaido, M., Okada, N., and Hasegawa, M., Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data, *Gene*, 259:149–158, 2000.

[5] Gatesy, J., Hayashi, C., Cronin, M.A., and Arctander, P., Evidence from milk casein genes that cetaceans are close relatives of hippopotamid artiodactyls, *Mol. Biol. Evol.*, 13:954–963, 1996.

[6] Hillis, D.M., SINEs of the perfect character, *Proc. Natl. Acad. Sci. USA.*, 96:9979–9981, 1999.

[7] Hudson, R.R., Gene trees, species trees and the segregation of ancestral alleles, *Genetics,* 131:509–512, 1992.

[8] de Jong, W.W., Leunissen, J.A.M., and Wistow, G.J., Eye lens crystallins and the phylogeny of placental orders: Evidence for a macroscelid-paenungulate clade? *Mammal Phylogeny: Placentals* (Szalay F. S., Novacek M. J., and McKenna M. C., eds.), Springer-Verlag, New York, 5–12, 1993.

[9] Liu, F.G., Miyamoto, M.M., Freire, N.P., Ong, P.Q., Tennant, M.R., Young, T.S., and Gugel, K.F., Molecular and morphological supertrees for eutherian (placental) mammals, *Science*, 291:1786–1789, 2001.

[10] McKenna, M.C. and Bell, S.K., *Classification of Mammals Above the Species Level*, Columbia University Press, N.Y., 1997.

[11] Madsen, O., Scally, M., Douady, C.J., Kao, D.J., DeBry, R.W, Adkins, R., Amrine, H.M., Stanhope, M.J., de Jong, W.W., and Springer, M.S., Parallel adaptive radiations in two major clades of placental mammals, *Nature*, 409:610–614, 2001.

[12] Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang. Y.P., Ryder, O.A., O'Brien, S.J., Molecular phylogenetics and the origins of placental mammals, *Nature*, 409:614–618, 2001.

[13] Nikaido, M., Rooney, A.J., and Okada, N., Phylogenetic relationships among Cetartiodactyls based on insertions of short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales, *Proc. Natl. Acad. Sci. USA*, 96:10261–10266, 1999.

[14] Penny, D. and Hasegawa, M., Molecular systematics. The platypus put in its place, *Nature*, 387:549–550, 1997.

[15] Rambaut, A. and Charleston, M., *Phylogenetic Tree Editor v1.0a8*, http://evolve.zoo.ox.ac.uk/software/, 2001.

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

[17] Schwarz, G., Estimating the dimension of a model, *Annals of Statistics*, 6:461–464, 1978.

[18] Sullivan, J. and Swofford, D.L., Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics, *J. Mammal. Evol.*, 4:77–86, 1997.

[19] Swofford, D.L., *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0b4.* Sinauer Associates, Sunderland, Massachusetts, 2000.

[20] Swofford, D.L., Olsen, G.J., Waddell, P.J., and Hillis, D.M., Phylogenetic inference, *Molecular Systematics (second edition)*, Hillis, D.M., Moritz, C., and Mable, B.K., eds., Sinauer Associates, Sunderland, Massachusetts, 407-514, 1996.

[21] Waddell, P.J., *Statistical methods of phylogenetic analysis: including Hadamard conjugations, LogDet transforms, and maximum likelihood*, Ph.D. Thesis, Massey University, Palmerston North, New Zealand, 1995.

[22] Waddell, P.J., Cao, Y., Hasegawa, M., and Mindell, D.P., Assessing the Cretaceous superordinal divergence times within birds and placental mammals using whole mitochondrial protein sequences and an extended statistical framework, *Syst. Biol.*, 48:119–137, 1999.

[23] Waddell, P. J., Cao, Y., Hauf, J., and Hasegawa, M., Using novel phylogenetic methods to evaluate mammalian mtDNA, including AA invariant sites-LogDet plus site stripping, to detect internal conflicts in the data, with special reference to the position of hedgehog, armadillo, and elephant, *Syst. Biol.* 48:31–53, 1999.

[24] Waddell, P.J., Okada, N., and Hasegawa, M., Progress in resolving the interordinal relationships of placental mammals, *Syst. Biol.*, 48:1–5, 1999.

[25] Waddell, P.J. and Penny, D., Evolutionary trees of apes and humans from DNA sequences, *Handbook of Human Symbolic Evolution* (A.J. Lock and C.R. Peters, eds.), Oxford Univ. Press, Oxford, England, 53–73, 1996.

[26] Yang, Z., PAML: a program package for phylogenetic analysis by maximum likelihood, *Comput. Appl. Biosci.*, 5:555–556, 1997.

[27] Yang, Z., and Rannala, B., Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method, *Mol. Biol. Evol.*, 14:717–724, 1997.

[28] Zietkiewicz E., Richer, C., and Labuda, D., Phylogenetic affinities of tarsier in the context of primate Alu repeats, *Mol. Phylogenet. Evol.*, 11:77–83, 1999.