ELSEVIER

Gene 294 (2002) 119–129

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Pika and vole mitochondrial genomes increase support for both rodent monophyly and glires

Yu-Hsin Lin*, Peter J. Waddell[1], David Penny

Allan Wilson Centre for Molecular Ecology and Evolution, Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand

Received 4 December 2001; received in revised form 28 March 2002; accepted 14 May 2002
Received by T. Gojobori

Abstract

Complete mitochondrial genomes are reported for a pika (*Ochotona collaris*) and a vole (*Volemys kikuchii*) then analysed together with 35 other mitochondrial genomes from mammals. With standard phylogenetic methods the pika joins with the other lagomorph (rabbit) and the vole with the other murid rodents (rat and mouse). In addition, with hedgehog excluded, the seven rodent genomes consistently form a homogeneous group in the unrooted placental tree. Except for uncertainty of the position of tree shrew, the clade Glires (monophyletic rodents plus lagomorphs) is consistently found. The unrooted tree obtained by ProtML (Protein Maximum Likelihood, a program in MOLPHY) is compatible with a reclassification of mammals [Syst. Biol. 48, 1–5 (1999)] which is also supported by other recent studies. However, when this tree is rooted with marsupials plus platypus, the outgroup often joins the lineage leading to the three murid rodents, so the rodents are no longer monophyletic. Apart from misplacing the root, the presence of the outgroups also distorts other parts of the unrooted tree. Either constraining the tree to maintain rodents monophyletic, or omitting murids, maintains the ingroup tree and sees the outgroup join on the edge to Xenarthra, to Afrotheria, or to these two groups together. This emphasises the importance of carrying out both an unrooted and a rooted analysis. It is known from cancer research that murid rodents have reduced activity in some DNA repair mechanisms and this alters their substitution pattern – this may be the case for mitochnodrial DNA as well. Comparing nucleotide compositions may identify taxa that differ in aspects of their DNA repair mechanisms.

Keywords: Mammal evolution; Murid rodents; Rodent DNA repair; Pika; Vole

1. Introduction

The superordinal tree of placental mammals is rapidly being resolved (Madsen et al., 2001; Murphy et al., 2001; Waddell et al., 2001) but several important questions remain. These include the question of the monophyly of rodents, the position of rodents within placentals, and whether rodents plus lagomorphs (rabbits plus pikas) form a monophyletic group (Glires). All have been controversial. For example, D'Erchia et al. (1996) and Reyes et al. (2000) did not get rodent monophyly, others did (for example, Penny et al., 1999; Waddell et al., 1999b). In addition, the position of the lagomorph (rabbit) with respect to rodents has been variable. Although historically the position of lagomorphs within placentals has been uncertain (see Wood, 1957; van Valen, 1964), in recent morphological analyses (Shoshani and McKenna, 1998; Liu and Miyamoto, 1999) strong support has been found for Glires (a strictly monophyletic group of rodents plus lagomorphs). Nuclear data, with some mitochondrial ribosome sequences (Madsen et al., 2001; Murphy et al., 2001), or without them (Waddell et al., 2001), has recently been grouping rodents and lagomorphs. (For simplicity, we refer to the Murphy and Madsen datasets as 'nuclear' because, although they have some mitochondrial ribosome sequences, similar results are found without the mitochondrial sequences, see Waddell et al., 2001). The only evidence of Glires with mitochondrial DNA (mtDNA) proteins (Waddell et al., 1999b) remains ambivalent. It is thus desirable to include a mitochondrial genome from the lagomorph family Ochotonidae (pikas, mouse hares and conies) in addition to the rabbit (Leporidae, rabbits and hares) since this is the deepest divergence within Lagomorpha.

In addition to the above more classical groups, other super-ordinal groupings involving rabbits have recently emerged. Tree shrew has come with rabbit (e.g. Schmitz et al., 2000), rather than with Euarchonta (paraprimates –

Abbreviations: mtDNA, mitochnodrial DNA

* Corresponding author. Tel.: +64-6-350-5033; fax: +64-6-350-5688.
E-mail address: y.lin@massey.ac.nz (Y.-H. Lin).
[1] Present address: Department of Statistics, University of South Carolina, Columbus, SC 29208, USA.

0378-1119/02/$ - see front matter
PII: S0378-1119(02)00695-9

tree shrews, flying lemurs and primates, Waddell et al. 1999a). Euarchonta has rapidly growing support, as does its sister relationship with Glires (Madsen et al., 2001; Murphy et al., 2001; Waddell et al., 1999a, 2001). The rooting of the mtDNA tree has been questioned (Waddell et al. 1999b) based on different trees from the tRNA genes and the non-stationarity of amino acid frequencies in taxa near the root. Our findings suggest that unrooted placental tree for the mtDNA proteins is consistent with a tree very similar to that of Madsen et al. (2001), Murphy et al. (2001) and Waddell et al. (1999a, 2001). In contrast, the rooted trees of mtDNA proteins differ from the nuclear data.

Thus far, most work on mitochondrial genomes has focused on phylogeny, but as the tree becomes more stable, a wide range of other applications is possible (for example, Pollock et al., 2000). Recent statistical tests have emphasised the distinct amino acid composition of hedgehog, primates, murid rodents, and whales amongst placental mtDNA (Waddell et al., 1999b, Table 1). Thus, apart from representing very long branches (edges), it was possible that rat and mouse are evolving anomalously, while guinea pig has a more typical substitution process. It is well-categorised in DNA repair studies (Holmquist and Filinski 1994; Karlin and Mrázek 1997) that some murid rodents have a reduced effectiveness in their nuclear DNA repair. This results in a different mutational process and, as predicted by the neutral theory of molecular evolution (Kimura, 1983), leads to differences in sequence evolution. Addition of a more divergent murid rodent (vole *Volemys kikuchii*) is desirable to both break up this edge and hopefully help by showing a less divergent amino acid composition.

The simple mechanism for a change in amino acid composition is a change in the relative mutation rate between some pairs of nucleotides, such as for C → T interchanges. Karlin and Mrázek (1997) have detected a change in dinucleotide frequencies for murid rodents, relative to other placental mammals. Thus there is certainly prior evidence of a change in the nucleotide composition in nuclear genes relative to other mammals (Cortopassi and Wang, 1996; Holmquist and Filinski, 1994; Op het Veld et al., 1997). It will be interesting to see if mtDNA follows the same pattern, or suggests some differences in DNA repair mechanisms.

2. Materials and methods

DNA was extracted from liver or muscle of the collared pika *Ochotona collaris* and the Taiwan vole *Volemys kikuchii* using High Pure™ polymerase chain reaction (PCR) Template Purification Preparation Kit (Roche). In order to avoid amplifying nuclear copies long-range PCR was applied using the Expand™ Long template PCR kit (Roche). The mtDNA primers and their sequences for two ~9 kb fragments were:

Long 16S-For (AATTAGGGTTTACGACCTCGATGTTGGATCAGG) to
H11685-Rev (CCTAAGACCA ATGGATTACT TCTATCCT) and
L11012-For (AGCTCTATCTGCTTTCGTCAAACAG) to
Long16S-Rev (TGATTATGCTACCTTTGCACGGTCAGGATACC).

Long PCR DNA fragments were sequenced directly, and used as template for short-range PCR of 0.5 ~ 2 Kb. Sequencing reactions were according to manufacturer's protocols, run on a 377 ABI DNA sequencer, and sequenced in both directions. Because of the problem of different lengths in C or G homopolymers, and different copy numbers of tandem repeats in the control region, we could not always get clear sequences directly from PCR products. Where necessary, short-range PCR products were amplified and cloned into the vector pGem-T (Promega).

Complete mammalian mt-DNA sequences were obtained from Genbank for the following 30 taxa. Rodentia: mouse *Mus musculus* [NC_001569]; rat *Rattus norvegicus* [NC_001665]; guinea pig *Cavia porcellus* [NC_000884]; dormouse *Myoxus glis* [NC001892]; squirrel *Sciurus vulgaris* [NC_002369]; cane rat *Thryonomys swinderianus* [NC_002658]. Lagomorpha: rabbit *Oryctolagus cuniculus* [NC_001913]. Primates: human *Homo sapiens* [NC_001807]; gibbon *Hylobates lar* [NC_002082]; baboon *Papio hamadryas* [NC_001992]; Barbary ape *Macaca sylvanus* [NC002764]; Capuchin *Cebus albifrons* [NC_002763]; Loris *Nycticebus coucang* [NC_002765]. Scandentia: tree shrew *Tupaia belangeri* [NC_002521].

Table 1
Alternative trees for Figs. 2 and 3[a]

	1 + 2			RNA			(1 + 2) + RNA			AA		
	ML	MP	NJ	ML	MP	NJ	ML	MP	NJ	ML	MP	NJ
Fig. 2	C	C	B	A	A	A	A	A	B	Fig. 2	A	A
Fig. 3	3	3	1	1	1	2	1	1	1	1	1	1

[a] For Fig. 2 (the unrooted tree), alternative positions for tree shrew on different data sets and methods of analysis. (Fig. 2 is the ML tree on amino acids.) A, tree shrew joins to the rabbit/pika lineage; B, tree shrew is basal on the Supraprimate lineage; and C, tree shrew joins Armadillo. For Fig. 3 (the unconstrained rooted tree), the deepest branch for different data set and methods of analysis. 1, mouse/rat/vole; 2, Tenrec; and 3, Afrotheria. For both figures, 1 + 2: 1st and 2nd codon position of 12 amino acids genes, AA: amino acids sequences, ML: maximum likelihood, MP: maximum parsimony, NJ, neighbor joining (LogDet distances).

Tubulindentata: aardvark *Orycteropus afer* [NC_002078]. Proboscidia: elephant *Loxodonta africana* [NC_000934]. Afrosoricida: tenrec *Echinops telfairi* [NC_002631]. Xenarthra: armadillo *Dasypus novemcinctus* [NC_001821]. Chiroptera: fruit bat *Artibeus jamaicensis* [NC_002009]; flying fox *Pteropus scapulatus* [NC_002619]. Eulipotyphla: mole *Talpa europaea* [NC_002391]. Carnivora: dog *Canis familiaris* [NC_002008]; cat *Felis catus* [NC_001700]; harbor seal *Phoca vitulina* [NC_001325]. Perissodactyla: horse *Equus caballus* [NC_001640]; white rhinoceros *Ceratotherium simum* [NC_001808]. Cetartiodactyla: hippopotamus *Hippopotamus amphibius* [NC_000889]; cow *Bos taurus* [NC_001567]; fin whale *Balaenoptera physalus* [NC_001321]; pig *Sus scrofa* [NC_000845].

We selected sequences of all deep-diverging lineages within the groups of interest, Afrotheria, Xenarthra, rodents, lagomorphs and primates. Within Laurasiatheria, lineages were selected to give the deepest splits as long as intra-ordinal placement was unambiguous. Uncertainty regarding the position of llama and the New Zealand long-tailed bat saw them excluded. (For taxa included in each of the above mentioned superorders, see Waddell et al., 2001). The hedgehog was omitted because it may be misplaced (see Waddell et al., 1999b) perhaps due to a high rate of nucleotide substitution that is also non-stationary and affecting amino acid composition. (A reanalysis of the position of the hedgehog mtDNA sequence along with its near relative gymnure is in Lin et al., 2002). For the outgroup, mitochondrial genomes from four marsupials (opossum *Didelphis virginiana* [NC_001610], wallaroo *Macropus robustus* [NC_001794], bandicoot *Isoodon macrourus* [NC_002746] and brush-tailed possum *Trichosurus vulpecula* [NC_003039]), plus a platypus *Ornithorhynchus anatinus* [NC_000891], were used (see Phillips et al., 2001).

SeAl version 1.0 a1 (http://evolve.zps.ox.ac.uk/software.html) was used for aligning RNA and protein-coding datasets manually. RNA sequences were aligned using secondary structure (http://www.rna.icmb.utexas.edu/RNA/). Alignments were made independently by Y.-H.L and P.J.W, and then edited to remove regions of ambiguity. The five outgroups were similar enough that their inclusion did not require removal of further sites. The first dataset comprised RNA sequences (rRNAs + tRNAs) and the second the 12 protein genes coded on the H-strand (both as 1st and 2nd position nucleotides [1 + 2] plus translated to amino acids). The RNA and protein data sets were also combined as nucleotides. The RNA and protein data sets allow independent estimates of the gene phylogeny in that they share no nucleotides in common. Data sets are available from (http://awcmee.massey.ac.nz/software.htm).

PAUP* 4d65 was used for all analyses except for Maximum Likelihood on amino acids; this used ProtML in the MOLPHY package (Adachi and Hasegawa, 1996), and protein LogDet which used the programs of Penny et al. (1999) and Waddell et al. (1999b). Trees were compared quantitatively using the partition metric (Steel and Penny, 1993). To avoid local optima, ProtML searches were seeded with multiple near optimal trees from different methods. The inequality test of Lockhart et al. (1998) was used to test for covarion evolution as opposed to i.i.d. models with unequal substitution rates across sites. This tests whether sites are always in the same rate class (rates across sites models), or whether sites vary in their rate of evolution as the tertiary structure of the macromolecule evolves (Penny et al., 2001). A triplet Markov analysis (analyzing three sequences simultaneously, rather than pairs of sequences) was undertaken using the program 'Gambit' from Lake (1997). This gives closed form estimates of the general transition matrices for each of three lineages.

3. Results

3.1. Complete mtDNA sequences

The pika and vole sequences are reported under GenBank numbers AF348080 and AF348082, respectively. The sequences have the standard gene order for mammals, and are 16,968 and 16,312 nucleotides long, respectively. There were no notable features in their gene organisation, total length, start and stop codons, etc.

3.2. Choice of analytic methods

With well over a hundred variants of methods differing in optimality criterion, search strategy, and the assumed mechanism of evolution it is not surprising that there are differences in results between the methods of analysis. This makes it difficult to quantitatively compare trees from independent data sets because trees will vary slightly with the analysis. We had, from previous experience, decided to compare the ProtML tree for the amino acid data set, with the maximum likelihood (ML) tree on nucleotides (using PAUP*). In addition, we wanted to know if the four main groups of placentals (Xenarthra, Afrotheria, Supraprimates, and Laurasiatheria) appearing in Waddell et al. (1999b) and subsequent work, would appear on this mitochondrial data set, especially with the pairing of (Xenarthra, Afrotheria) (Supraprimates, Laurasiatheria), see Fig. 1. Supraprimates means above or beyond Primates and is the name given in Waddell et al. (2001) to a group in Waddell et al. (1999a,b) that includes Euarchonta and Glires. Another name 'Euarchontaglires' was suggested for this group by Murphy et al. (2001) in work submitted, accepted and published slightly later. Another clade from Waddell et al. (1999a,b), consisting of Supraprimates and Laurasiatheria was named Boreotheria in Waddell et al. (2001) and as a junior synonym 'Boreoeutheria' by Murphy et al. (2001).

3.3. Unrooted placental tree

The first analysis is the unrooted tree of ingroup taxa, which is expected to be more congruent with other data if

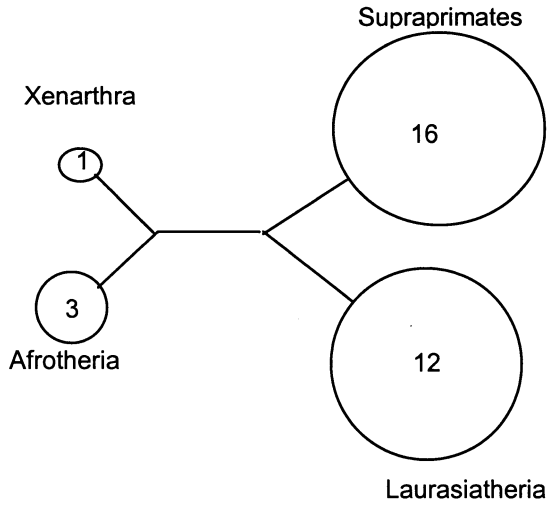

Fig. 1. Predicted relationship between four groupings of placentals based on nuclear data (Madsen et al., 2001; Murphy et al., 2001; Waddell et al., 1999a, 2001). In the present dataset there is one Xenarthran mt genome, three Afrotherians, 16 Supraprimates, and 12 Laurasiatherians.

rooting is a problem. Fig. 2 shows the protein ML tree for the 32 placentals for the 12 proteins coded on the same DNA strand. Similar trees were inferred for the three nucleotide datasets (1 + 2; RNA; combined 1 + 2 + RNA, trees not shown). In general the trees are highly similar, apart from the position of the tree shrew (*Tupaia*) which is locally unstable; results for the tree shrew are shown in Table 1. It can occur sister to Primates as a member of the Euarchonta (which is its expected position) but it was more frequently found basal to the lagomorphs (pika and rabbit), and sometimes basal to the Supraprimates (but still within that grouping). In only one analysis is it found outside the Supraprimates, joining with the armadillo (see Table 1); this maybe a long edge artefact. Using a partially-complete sloth mitochondrial genome (Trish McLenachan pers. comm.) keeps the tree shrew within the Supraprimates and is consistent with expectation (data not shown). Given that the position of the tree shrew does vary within the Supraprimates, the rest of the tree is virtually the same for both the protein and RNA data. There are two one-step rearrangements of note – mole joins with bats within the laurasiatherians, while within the afrotherians aardvark and tenrec are united. Thus there is excellent agreement between the trees from the RNA and the protein coding genes.

3.4. Probability of similar trees from different data

It is necessary to evaluate congruence objectively. For 31 taxa (excluding tree shrew), the probability of randomly selecting two trees with only two differences on the partition tree comparison metric is $\approx 0.5 \times 10^{-36}$ with all trees equally likely (Steel and Penny, 1993). Similarly, Fig. 2 is virtually the same as the tree on the combined DNA data set (RNA coding plus 1st and 2nd position of protein coding genes). The only difference between trees from the combined and RNA datasets is the one-step rearrangement within Afrotheria. However, in this case (comparing trees from the amino acid (or RNA) and combined data sets) the datasets are not independent. The important conclusion is that two data sets with no sites in common (the protein and RNA datasets) give extremely similar trees, meaning that the mammalian trees are converging as additional taxa are added.

Restricting the comparisons to orders, the tree in Fig. 2 is also highly congruent with the super-ordinal classification of mammals in Waddell et al. (1999a). There are 13 orders in the Fig. 2 (not including the subgroups of Cetartiodactyla which might be considered orders). So for a 13 taxon unrooted binary tree, and assuming all trees to be equally likely, the probability of eight out of ten partitions being identical by chance is 1.78×10^{-8} (Steel and Penny, 1993). Thus, in addition to the afore-mentioned analyses, the mtDNA data is strongly congruent with the new classification. In contrast, only two partitions are in common between the mtDNA trees and morphological trees (e.g. Shoshani and McKenna, 1998).

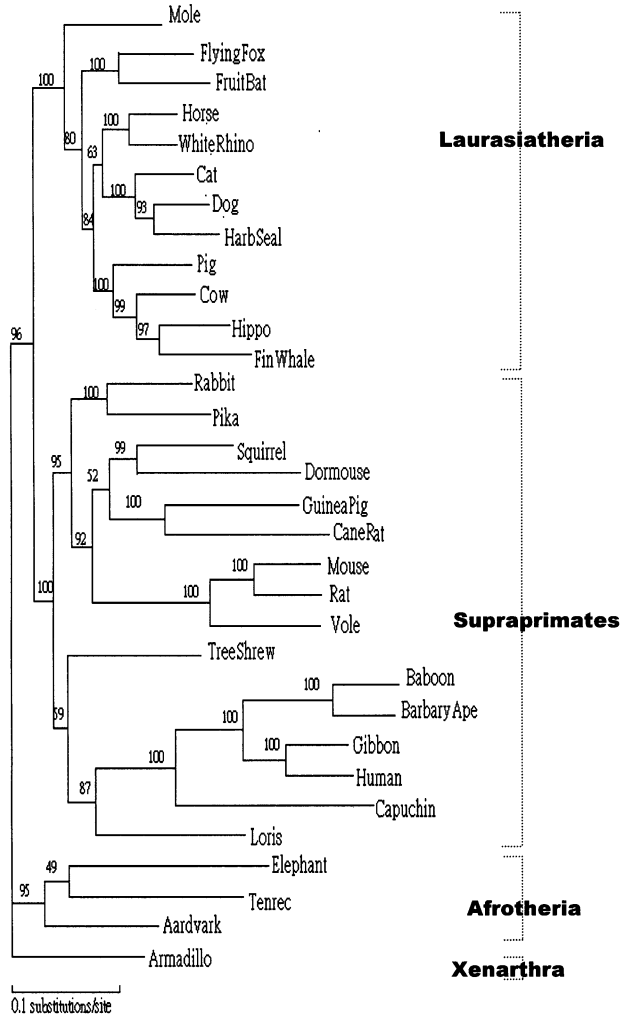

Fig. 2. The unrooted ProtML tree of 32 placentals based on the amino acid dataset with RELL bootstrap values shown. The four-way split predicted from nuclear data (Fig. 1) is found on this tree.

Another way to consider the high congruence we are seeing is to consider the partitions near the root. Fig. 1 shows the relationships and composition of the four main groups of placentals postulated in Waddell et al. (1999a), Madsen et al. (2001), and Murphy et al. (2001), together with the number of representatives of each group used in this study – one xenarthran (armadillo), three afrotherians, 12 laurasiatherians, and 16 supraprimates – giving 32 species. We find this same arrangement now based only on mitochondrial data. The probability of randomly selecting a tree with the same taxa in the same configuration is approximately 2.5×10^{-14}. The calculation is based on the following. Let $b(n) = (2n - 5)!!$ be the number of unrooted binary trees on n taxa, where the double factorial notation multiplies by every second number (in this case $1 \times 3 \times 5 \times \cdots 2n - 5$). The number of rooted trees for n taxa is $b(n + 1)$. The taxa in each of the four subsets can be arranged in $b(n + 1)$ rooted trees, and still be consistent with the tree in Fig. 1. Thus the number of 32 taxon trees which have this structure is b(2)*b(4)*b(13)*b(17)/3. Hence the probability of obtaining this basic tree on an independent data set is (b(2)*b(4)*b(13)*b(17)*3) divided by the number of unrooted trees on n taxa, b(32). Care is required in interpreting this value. It is not the probability that the tree is correct (there could be another tree almost as good on the same data sets). Rather, it is more comparable to the g-statistic of Huelsenbeck (1991) showing a strong signal in the datasets. However it is a more direct measure, and for a specific signal deep in the placental tree. To be considered more than a 'strawman' however it needs to be set up in a slightly different way. A morphologist might suggest that the pattern observed at the ordinal level was little better than chance. We have one xenarthran order, three afrotherian orders, seven laurasiatherians, and four supraprimates – giving 14 or 15 orders (a morphologist might treat whales as a separate order). The P value of such agreement by chance is 2.56×10^{-8}, while even if the morphologist argues that he expected whales and artiodactyls plus lagomorphs and rodents to fall together he is faced with a P of 2.06×10^{-7}. This is a clear counter argument to anyone claiming we are no closer to resolving the deep placental mammal tree than previously when we had major disagreement at the ordinal level between morphology and each individual molecular data set.

3.5. Mammalian tree structure and stability

Before focussing on the position of the new sequences in the tree, consider further the overall structure and stability. The laurasiatherian taxa (bats, carnivores, artiodactyls, perissodactyls, whales and Eulipotyphlans or core insectivores) are always monophyletic in our analyses. There is some local variation in positions within the Laurasiatheria. For example whether bats and Eulipotyphla form a group, or whether the latter are deeper is not certain. However, the latter resolution is being seen more frequently with greater taxon sampling (for example, Lin and Penny, 2001; Lin et al., 2002; Waddell et al., 2001). The afrotherians, represented here by elephant, tenrec and aardvark are united in this tree, although the bootstrap support is low. A hyrax or a dugong genome may help stabilise the tree in this region. Again, the single Xenarthran (armadillo) groups with the Afrotheria, agreeing with the trees in Waddell et al. (1999a, 2001), Madsen et al. (2001), and Murphy et al. (2001).

3.6. Systematics within rodents

This leaves relationships within the supraprimates (represented here by primates/lagomorphs/rodents/tree shrew) to be considered further. Focusing on the new sequences, there is no ambiguity in support for Lagomorpha since pika and rabbit always come together. The three murids (vole/rat/mouse) also always come together with 100% bootstrap support. The vole joins about one third of the way up on the rat/mouse lineage (which had been the longest internal edge in the tree). The two hystricomorph rodents (cane rat and guinea pig) are united, in agreement with Mouchaty et al. (2001). Similarly the squirrel and the dormouse are united, though this result is not predicted on current classifications – squirrel is in the Sciuromorpha and dormouse is usually assigned to a basal position among myomorph rodents. However, Kramerov et al. (1999) report a squirrel and dormouse grouping based on their having a similar copy number of a retrotransposon. Huchon et al. (2000) report a relatively close association between dormouse and squirrel and with DNA profiles on caesium chloride gradients, the dormouse (Gliridae) does not fit within the myomorph rodents (Douady et al., 2002). With sequence data this grouping is found in Murphy et al. (2001), and the largest concatenation of Waddell et al. (2001). Montgelard et al. (2002) report a glirid/sciuroid grouping based on short mitochondrial sequences. Given the present results, together with the five previous ones, it appears that the squirrel/dormouse association is likely. The sciuromorphs generally were not closer to the murids than the hystricognaths, though this will depend on where the rodent subtree is rooted. Given that aspects of the mutational mechanism appear changed in murid rodents (see later) then any final conclusions on this point may have to await improved taxon sampling.

3.7. Rooted placental tree

The next step is to root the placental tree using four marsupials and platypus (monotreme) as the outgroup (Fig. 3). It is at this point that it could be said, 'all hell breaks loose', the position of the root differs markedly to those obtained with either morphological or nuclear data. With the outgroup added, most of the mitochondrial datasets and most methods (Table 1) move murids deepest in the placentals – to the base of the tree (the main exception is the first two codon positions of the protein coding genes). This makes rodents paraphyletic. However, the same four-way division of placentals (Afrotheria, Laurasiatheria, etc.)

is still maintained as in Waddell et al. (1999a), and there are no major rearrangements on the ingroup tree. The same rooting is found with the LogDet correction on amino acids (Penny et al., 1999; Waddell et al. 1999a), even with all constant sites are removed.

We have two hypotheses about this rooting, and they lead

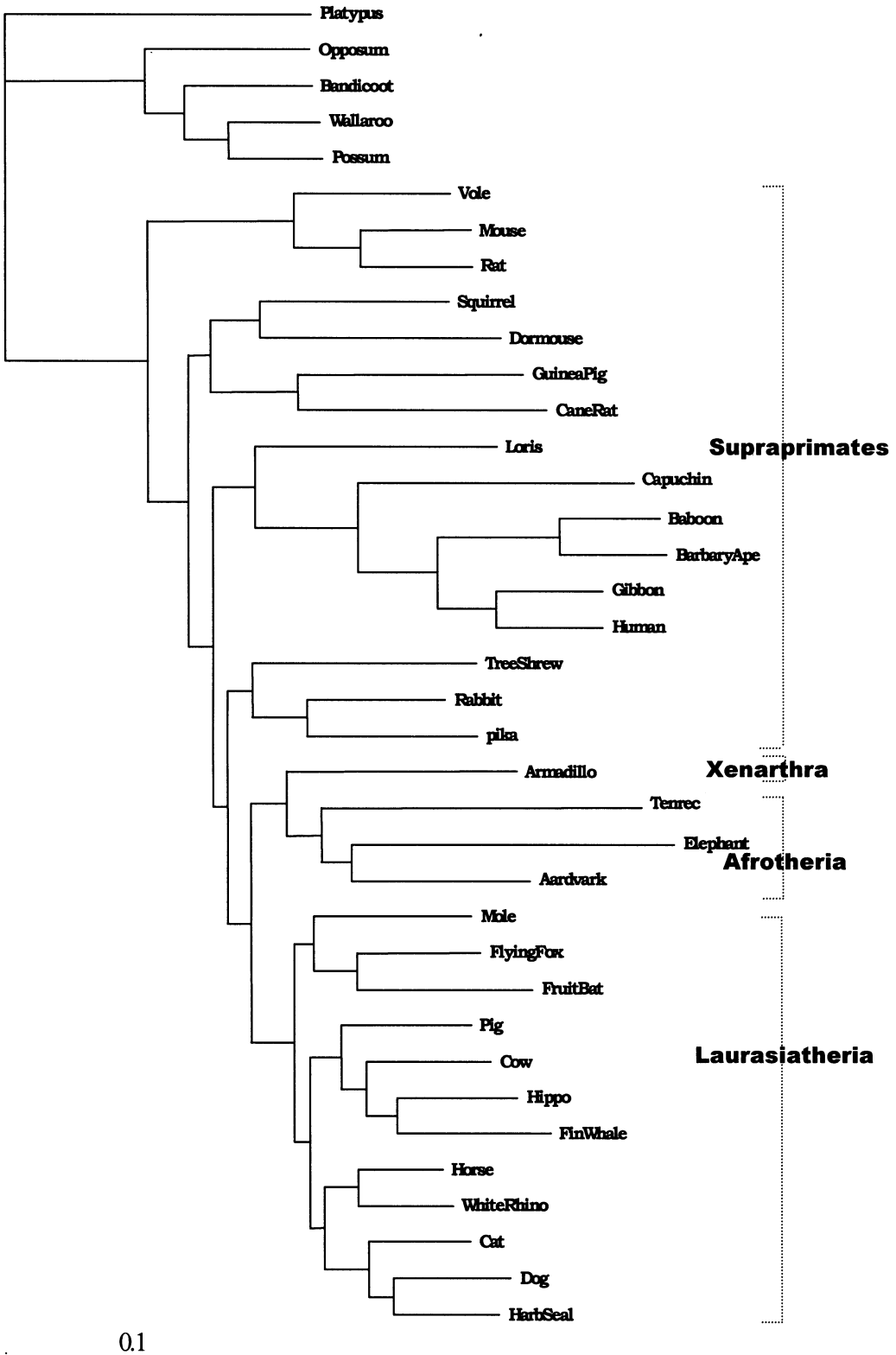

Fig. 3. The tree of Fig. 2 rooted with four marsupial and a monotreme sequences. There are no constraints on this tree and the root comes onto the murid rodent lineage (c.f. Phillips et al., 2001).

to different predictions if murid rodents are omitted from the dataset. The alternatives are:

A. The rooting at the base of murid rodents is an artefact of a different mutational process. If the murids are omitted the root reverts to the Afrotherian/Xenarthran part of the tree.
B. The rooting at the base of murids rodents is genuine. The root will stay with the other rodents if murids are omitted.

In fact, when the murid rodents are omitted, the root comes to the base of the afrotherian group – not to the remaining rodents. This shift of the root away from the rodents contradicts their being basal – it is some feature of the murid rodents that is interacting with the outgroup, not a general similarity of their sequences. The same result was also found using the ML for DNA sequences with a gamma correction. Indeed, forcing the gamma shape to be more extreme with the full data set (with murids) led to the outgroups joining on the internal edge separating the Afrotherians plus Xenarthra (though some changes to the ingroup then started to appear).

3.8. Detecting a change in mutational processes in murid rodents

The most likely explanation why the murids are attracted to the root, is a change in the evolutionary process in murid rodents, a change that is uncorrected for by the tree-building programs. Earlier we noted that it is known from cancer research that some DNA-repair mechanisms are less efficient in murids. A way of testing for this is by using a triplet Markov method (Lake, 1997) to analyse three sequences simultaneously using tensors (three-dimensional matrices). The $4 \times 4 \times 4$ tensor has 63 independent entries (64 – 1). The 4×4 Markov transition matrices from the root to each of the three species require estimating 36 (3×12) parameters, and three independent values (4 – 1) are required for the composition of nucleotides at the root. If the sequences are sufficiently long, the tensor has sufficient information to recover the full model for the three species. Results with mitochondrial genomes (excluding D-loops) for squirrel, guinea pig and vole are given in Table 2. The three matrices are the estimated Markov transition matrices from the root to the observed nucleotide frequencies for squirrel, guinea pig, and vole respectively. The results are consistent with a higher C/T ratio in the vole. This is preliminary evidence that the murid has the most divergent substitution process, and the triplet Markov analysis is a productive area for further research.

The results estimating the Markov transition matrices indicate that a difference in the mutational process on murids is a viable explanation for the unexpected position of the root. It is consistent also with evidence of C/T composition shift within the outgroup, (Phillips et al., 2001) and a shift in amino acid composition within murids (Waddell et al., 1999a). A change on the murid lineage of the amino acid sites that are free to vary is another possibility (this is a change in the covarion structure of the proteins – Penny et al., 2001). However, the test of Lockhart et al. (1998) gives no evidence for a change in covarion structure (results not shown).

Another way of testing the murid rooting is to constrain the rodents to be strictly monophyletic, and see whether the root now moves just outside the rodents. This is its expected position if the root really did belong there. The result of such an experiment (constraining rodents to be monophyletic) using ML for DNA sequences (in PAUP*) is shown in Fig. 4. The root now moves to quite a different place on the tree, onto the combined Xenarthran (armadillo) plus Afrotheria. This is four steps on the tree away from the base of the rodents, and is one of the expected positions for the root based on previous analyses. This major shift in the root is expected if there was a difference in substitution process in the three murid rodents and not in the four other rodents in the sample. Given the prior information of a change in process for nuclear genes (Cortopassi and Wang, 1996; Holmquist and Filinski, 1994; Op het Veld et al., 1997) and in mitochondria (Karlin and Mrázek, 1997) it would seem that misrooting on murids should be considered in all types of sequence data.

4. Discussion

4.1. Addition of pika and vole mtDNA to placental tree

As reported here, with the addition of vole and pika mitochondrial genomes, the mtDNA tree seems to be making more sense. There is good agreement between the RNA

Table 2
Estimated transition matrices from the root to each of three rodents[a]

	Root to squirrel				Root to guinea pig				Root to vole			
	T	C	A	G	T	C	A	G	T	C	A	G
T	0.972	0.082	0.024	0.007	0.965	0.090	0.022	0.010	0.940	0.053	0.028	0.019
C	0.017	0.908	0.012	0.006	0.027	0.892	0.014	0.003	0.039	0.912	0.023	0.008
A	0.010	0.010	0.945	0.051	0.007	0.016	0.939	0.039	0.018	0.031	0.929	0.047
G	0.001	0.001	0.018	0.937	0.001	0.002	0.025	0.948	0.002	0.004	0.021	0.926

[a] The average standard deviation for off-diagonal entries is approximately ± 0.003. The estimated composition at the root is given by the vector [0.285, 0.245, 0.280, 0.189]. Multiplying this vector by the appropriate transition matrix gives the observed mitochondrial nucleotide composition for each species.

and protein coding datasets. The unrooted trees are showing strong congruence with both prior hypotheses (Waddell et al., 1999b) and recently expanded nuclear data sets (Madsen et al., 2001; Murphy et al., 2001; Waddell et al., 2001).

The unrooted placental tree from mtDNA appears to be close to showing only local rearrangements due to errors (that is, the tree is expected to differ from the historical tree by one or two non-adjacent local interchanges). Thus,

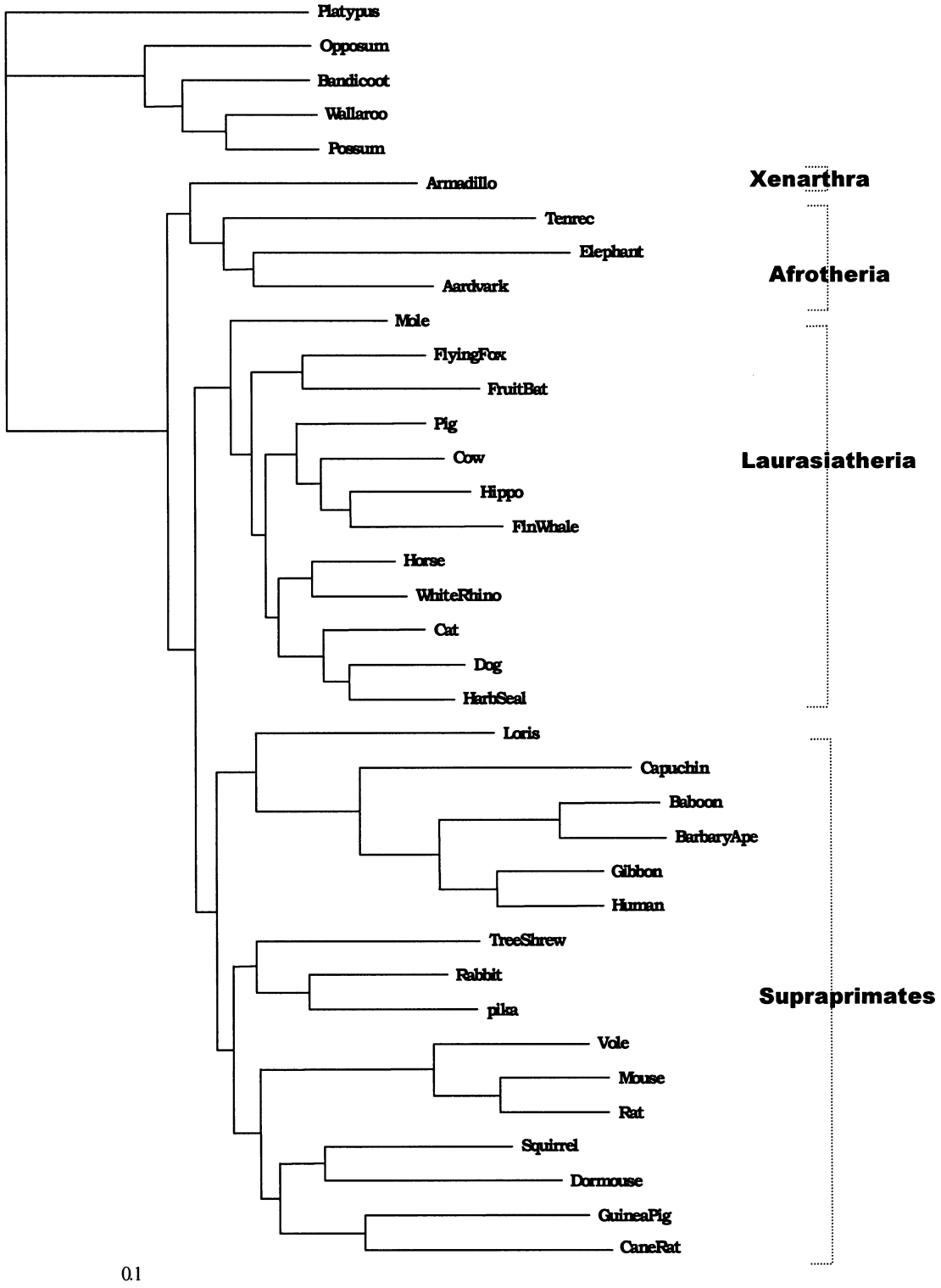

Fig. 4. The alternative tree to Fig. 3 where the rodents are constrained to be strictly monophyletic. If the correct rooting was on the murid rodents (as in Fig. 3) then the root should now come directly outside the rodents (and within Glires). Instead the root moves to a different part of the tree (Xenarthrans), similar to the rooting with nuclear or morphological data. This major shift in the position of the root is further evidence against the root belonging on the murid rodent lineage.

there is much more than just convergence of nuclear and mitochondrial data towards four basic groups of placentals – Afrotheria, Xenarthra, Laurasiatheria and Supraprimates (important as this is). Considering orders yet to be sampled adequately and given about six spots for local rearrangement (and some of these involve three way splits) we could expect perhaps one to three differences from our current best estimates (e.g. Waddell et al., 2001).

4.2. Rooting in murid rodents an artefact

Given the preliminary result using tensors (Table 2) and the major change in the position of the root when rodent monophyly is constrained, our working hypothesis is that the apparent rooting in murid rodents is an artefact arising from the change in their DNA repair processes. It is apparent that, even within mammals, the base compositional shifts are greater than expected and therefore a sign of non-stationary evolution which may well distort the trees. It is useful to note that, because constant sites are excluded, the tests of Penny et al. (1999) and Waddell et al. (1999a) are more powerful at detecting non-stationarity than earlier tests such as those in PAUP*. Detection of significant composition shifts is one of the few warning signs with that our model's assumptions are broken.

4.3. Rodent monophyly and Glires

Given the above, there is now generally good evidence from mtDNA of rodent monophyly, and also (apart from the problem of the tree shrew) of rodents joining with lagomorphs to give Glires. Adding the pika and vole data has helped stabilise the tree, but a large part of this result is from concentrating on the unrooted tree (and with exclusion of the clearly non-stationary hedgehog sequence). Similarly, the addition of three primate mitochondrial genomes (Arnason et al., 2001) has broken up what had been the longest internal branch of the placental tree, and this too appears to have increased the stability of the tree. (Before those sequences were available there was a tendency for lagomorphs and/or elephant to move across to the long internal edge before the apes split, Waddell et al., 1999a) Although the vole has reduced the length of the edge leading to murid rodents, it is still the longest internal edge in the placental tree, and a prime target for further taxon sampling (e.g. *Spalax* or mole rats). The problem with the tree shrew shifting about might be due to unusual base composition or poor taxon sampling within tree shrews (a lack of flying lemur, and a fairly long edge to tree shrew).

4.4. Adding extra taxa to break the long branch attraction

The tendency for the root to join incorrectly with murid rodents, resulting in rodent paraphyly, shows that long branch attraction is a real problem, even within placentals and with amino acid sequences (Waddell et al., 1999a). Although improved models will help, the immediate solution is probably additional mitochondrial genomes. Action on proposals to accelerate the rate of sequencing of mitochondrial genomes (Pollock et al., 2000) may see this situation improve rapidly. However, we should remember that deep within taxonomic groups we may not have the luxury of increased taxon sampling. Accordingly, results such as those in the present paper are both encouraging and sobering in regard to some of the outstanding problems in uncovering deep phylogenetic splits. If ~90 million years of placental evolution can see errors, how well are we doing with bacterial genomes that diverged billions of years ago and show tremendously long duration unbranched lineages?

Is mtDNA protein data is as reliable as nuclear data? Certainly the apparent attraction between the hedgehog sequence, murids and outgroups is a major issue, though it is not unique to mtDNA data. Both the nucleotide (Madsen et al., 2001) and the protein (Waddell et al., 2001) sequence trees of BRCA1 show misrooting and rodents becoming non-monophyletic. In addition both the latest nuclear and mtDNA data sets show problems with the location of the tree shrew, although single genes (particularly TP53) that show very clean data are very strongly in favour of Euarchonta (Waddell et al., 2001). Finally, the retention index of the best nuclear and mtDNA protein datasets seem to be about 0.4–0.5 (Waddell et al., 2001). In retrospect, much of the problem with the mtDNA protein data was the order in which it was collected, with some of the problematic taxa (murids, hedgehog) being collected early when taxon sampling was worst. However in either case we urge caution in not over-interpreting the bootstrap results of either amino acid or nucleotide analyses. Congruence of multiple independent data sets now suggests that the molecular tree is largely correct, but reliable statistical testing of clades may need to await SINE data (Waddell et al., 2001).

4.5. Molecular datasets agreeing, less with morphological data

We have further quantified congruence between independent estimates from the mtDNA data sets. While there is very good agreement amongst molecular data sets (with $P \ll 0.0001$), there is less agreement with morphological based trees. Quantifying congruence between datasets is under-utilised in phylogenetics. It was primarily this approach that led to the improved estimate of placental phylogeny by Waddell et al. (1999a,b). Congruence may be the answer to the 'total evidence approach' whereby misleading data can swamp signal in good data. With data sets of tens of thousands of sites, bootstrap proportions (or even more liberal Bayesian posteriors, Waddell et al., 2001) tend towards one, yet this can only indicates that sampling error is minor but gives the reader no gauge of potential systematic error. This is unsatisfactory, indicating phylogenetics is not yet a mature science.

4.6. Current analytic methods are not perfect

There is a natural tendency to say that the rat/mouse/vole sequences are 'bad data'. But of course the data is correct (excluding a few sequencing errors); rather our analytical methods are primarily 'wrong' – they make erroneous assumptions about the mechanisms of evolution. In truth, our current models and methods are incomplete. Current models are based on a stochastic mechanism and with expected numbers of changes between nucleotides. There are many signals in sequences in addition to a historical (phylogenetic) signal and we need to consider them all. Other signals (perhaps from a mutational bias) might be a nuisance with respect to phylogeny but could, for example, be very interesting for understanding changing mutational processes – or changing protein 3D structure and function through time. There is no reason to expect all the evolutionary changes to be free of convergences and parallelisms and thus allow us to reconstruct history easily.

If a tree like Fig. 4 is correct, then a notable feature is the presence of small-generalised insectivores branching near the root of all the major lineages (except Xenarthra). Given this, the principle of parsimony, and the difficulty of going from specialised forms back to very general forms, support the conclusion that small generalised insectivores were ancestral not just to the whole placental tree, but also at the root of all the major groups namely Laurasiatheria, Afrotheria, Supraprimates and Xenarthra. It thus appears that the more derived body forms did not occur until after each of these major groups diverged. This agrees with fossil indications in contrast to earlier molecular trees that seemed to suggest early transitions to more derived forms (e.g. Glires-like) early in the tree.

4.7. Change in rate, or change in mutational process?

Work in the past has concentrated on changes in the 'rate' of evolution (for example, Hendy and Penny, 1989). However, a simple rate change would imply that all values in a Markov rate matrix increased (or decreased) in proportion. In retrospect, it is difficult to find a mechanism that would change all such values equally. There are up to 70 enzymes involved in DNA replication and repair, and they fit into a range of different categories. These include photolyases (repair of pyrimidine dimers); DNA repair methyl transferases (repair methylation and similar damage), base excision repair (removal of abnormal or damaged nucleotides), and mismatch repair – see reviews by Memisoglu and Samson (1996) and Yu et al. (1999). Each of these major systems consists of a large group of enzymes, and it should be expected that with alterations to these enzymes the error rates on all nucleotide transitions are not affected equally. We refer to a change of 'process' (not just 'rate') when there is a marked change in some nucleotide interconversions compared with others.

Now, approximately 100 years after most orders of mammals were correctly recognised, the superordinal tree of mammals is rapidly resolving. Once a stable tree is found then many additional questions can be studied – times of divergence, biogeography, rates of speciation, likely transitions between niches (such as terrestrial insectivore to omnivore), and detection of selection pressures. The major result here is to show that the mtDNA data, at least in unrooted form, is congruent with the nuclear data (Madsen et al., 2001; Murphy et al., 2001; Waddell et al., 2001).

Acknowledgements

We thank Michael Sorensen for the pika sample; Cheng Hsi-Chi of Taiwan Endemic Species Research Institute (TESRI) for the vole; Barbara Holland and Rissa Ota for the triple Markov analysis; Jim Lake for his triple Markov program; Matt Phillips, David Archibald, and Mike Sorenson for discussions on mammalian evolution; and the NZ Marsden Fund for financial support.

References

Adachi, J., Hasegawa, M., 1996. Comput. Sci. Monogr. 28. MOLPHY: version 2.3: Programs for molecular phylogenetics based on maximum likelihood. Inst. Stat. Math. Tokyo. http://bioweb.pasteur.fr/seqanal/interfaces/MolPhy.html

Arnason, U., Gullberg, A., Burguete, S., Janke, A., 2001. Molecular estimates of primate divergences and new hypotheses for primate dispersal and human origins. Hereditas 133, 217–228.

Cortopassi, G.A., Wang, E., 1996. There is substantial agreement among interspecies estimates of DNA repair activity. Mech. Ageing Dev. 91, 211–218.

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

Douady, D., Carels, N., Clay, O., Catzeflis, F., Bernardi, G., 2002. Diversity and phylogenetic implications of CsCl profiles from rodent DNAs. Mol. Phylogenet. Evol. 17, 219–230.

Hendy, M.D., Penny, D., 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38, 297–309.

Holmquist, G.P., Filinski, J., 1994. Organisation of mutants along the genome: a prime determinant of genome evolution. Trends Ecol. Evol. 9, 65–69.

Huchon, D., Catzeflis, F.M., Douzery, E.J.P., 2000. Variance of molecular datings, evolution of rodents and the phylogenetic affinities between Ctenodactylidae and Hystricognathi. Proc. R. Soc. Lond. Ser. B 267, 393–402.

Huelsenbeck, J.P., 1991. Tree-length distribution skewness: an indicator of phylogenetic information. Syst. Zool. 40, 257–270.

Karlin, S., Mrázek, J., 1997. Compositional differences within and between eukaryotic genomes. Proc. Natl. Acad. Sci. USA 94, 10227–10232.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution, Cambridge Univ. Press, Cambridge.

Kramerov, D., Vassetzky, N., Serdobova, I., 1999. The evolutionary position of dormice (Gliridae) in Rodentia determined by a novel short retroposon. Mol. Biol. Evol. 16, 715–717.

Lake, J., 1997. Phylogenetic inference: how much evolutionary history is knowable? Mol. Biol. Evol. 14, 213–219.

Lin, Y.-H., Penny, D., 2001. Implications for bat evolution from two new complete mitochondrial genomes. Mol. Biol. Evol. 18, 684–688.

Lin, Y.-H., McLenachan, P.A., Phillips, M.J., Gore, A.R., Ota, R., Hendy, M.D., Penny, D., 2002. Four new mitochondrial genomes, and the

increased stability of evolutionary trees of mammals from improved taxon sampling. Mol. Biol. Evol. (in press).

Liu, F.-G.R., Miyamoto, M., 1999. Phylogenetic assessment of molecular and morphological data for eutherian mammals. Syst. Biol. 48, 54–64.

Lockhart, P.J., Steel, M.A., Barbrook, A.C., Huson, D.H., Charleston, M.A., Howe, C.J., 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. 15, 1183–1188.

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., Springer, M.S., 2001. Molecules reveal parallel adaptive radiations in two major clades of placental mammals. Nature 409, 610–614.

Memisoglu, A., Samson, L., 1996. DNA repair functions in heterologous cells. Crit. Rev. Biochem. Mol. Biol. 31, 405–447.

Montgelard, C., Bentz, S., Tirard, C., Verneau, O., Catzeflis, F.M., 2002. Molecular systematics of Sciurognathi (Rodentia). Mol. Phylogenet. Evol. 22, 220–233.

Mouchaty, S.K., Catzeflis, F., Janke, A., Arnason, U., 2001. Molecular evidence of an African phiomorpha-South American caviomorpha clade. Mol. Phylogenet. Evol. 18, 127–135.

Murphy, W.J., Eizirik, E., O'Brien, S.J., Madsen, O., Scally, M., Douady, C.J., Teeling, E., Ryder, O.A., Stahope, M.J., de Jong, W.W., Springer, M.S., 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294, 2348–2351.

Op het Veld, C.W., van Hees-Stuivenberg, S., van Zeeland, A.A., Jansen, J.G., 1997. Effect of nucleotides excision repair on hprt gene mutations in rodent cells exposed to DNA ethylating agents. Mutagenesis 12, 417–424.

Penny, D., Hasegawa, M., Waddell, P.J., Hendy, M.D., 1999. Mammalian evolution: timing and implications from using the Log determinant transform for proteins of differing amino acid composition. Syst. Biol. 48, 76–93.

Penny, D., McComish, B.J., Charleston, M.A., Hendy, M.D., 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J. Mol. Evol. 53, 711–723.

Phillips, M.J., Lin, Y.-H., Harrison, G.L., Penny, D., 2001. Complete mitochondrial sequences for two marsupials, a bandicoot and a brushtail possum. Proc. R. Soc. Lond. Ser. B 268, 1533–1538.

Pollock, D.D., Eisen, J.A., Doggett, N.A., Cummings, M.P., 2000. A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. Mol. Biol. Evol. 17, 1776–1788.

Reyes, A., Gissi, C., Pesole, G., Catzeflis, F.M., Saccone, C., 2000. Where do rodents fit? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*. Mol. Biol. Evol. 17, 979–983.

Schmitz, J., Ohme, M., Zischler, H., 2000. The complete mitochondrial genome of *Tupaia belangeri* and the phylogenetic affiliation of Scandentia to other eutherian orders. Mol. Biol. Evol. 17, 1334–1343.

Shoshani, J., McKenna, M.C., 1998. Higher taxonomic relationships among extant mammals based on morphology, with selected comparisons of results from molecular data. Mol. Phylogenet. Evol. 9, 572–584.

Steel, M.A., Penny, D., 1993. Distributions of tree comparison metrics – some new results. Syst. Biol. 42, 126–141.

van Valen, L.M., 1964. A possible origin for rabbits. Evolution 18, 484–491.

Waddell, P.J., Cao, Y., Hauf, J., Hasegawa, M., 1999a. Using novel phylogenetic methods to evaluate mammalian mtDNA. Syst. Biol. 48, 31–53.

Waddell, P.J., Okada, N., Hasegawa, M., 1999b. Toward resolving the interordinal relationships of placental mammals. Syst. Biol. 48, 1–5.

Waddell, P.J., Kishino, H., Ota, R., 2001. A phylogenetic foundation for comparative mammalian genomics. Genome Inf. 12, 141–154.

Wood, A.E., 1957. What, if anything, is a rabbit? Evolution 11, 417–425.

Yu, Z., Chen, J., Ford, B.N., Brackley, M.E., Glickman, B.W., 1999. Human DNA repair systems: an overview. Environ. Mol. Mutagen. 33, 3–20.