Research article

Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades

Shani Blanga-Kanfi[1], Hector Miranda[2,4], Osnat Penn[3], Tal Pupko[3], Ronald W DeBry[2] and Dorothée Huchon*[1]

Address: [1]Department of Zoology, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel, [2]Department of Biological Sciences, University of Cincinnati, Box 210006, Cincinnati, Ohio, 45221-0006, USA, [3]Department of Cell Research and Immunology, George S Wise Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel and [4]Department of Biology, Texas Southern University, 3100 Cleburne Street, Houston, TX 77004, USA

Email: Shani Blanga-Kanfi - shanib@post.tau.ac.il; Hector Miranda - mirandahc@tsu.edu; Osnat Penn - penn@post.tau.ac.il; Tal Pupko - talp@post.tau.ac.il; Ronald W DeBry - ron.debry@uc.edu; Dorothée Huchon* - huchond@post.tau.ac.il

* Corresponding author

Abstract

Background: Rodentia is the most diverse order of placental mammals, with extant rodent species representing about half of all placental diversity. In spite of many morphological and molecular studies, the family-level relationships among rodents and the location of the rodent root are still debated. Although various datasets have already been analyzed to solve rodent phylogeny at the family level, these are difficult to combine because they involve different taxa and genes.

Results: We present here the largest protein-coding dataset used to study rodent relationships. It comprises six nuclear genes, 41 rodent species, and eight outgroups. Our phylogenetic reconstructions strongly support the division of Rodentia into three clades: (1) a "squirrel-related clade", (2) a "mouse-related clade", and (3) Ctenohystrica. Almost all evolutionary relationships within these clades are also highly supported. The primary remaining uncertainty is the position of the root. The application of various models and techniques aimed to remove non-phylogenetic signal was unable to solve the basal rodent trifurcation.

Conclusion: Sequencing and analyzing a large sequence dataset enabled us to resolve most of the evolutionary relationships among Rodentia. Our findings suggest that the uncertainty regarding the position of the rodent root reflects the rapid rodent radiation that occurred in the Paleocene rather than the presence of conflicting phylogenetic and non-phylogenetic signals in the dataset.

Background

The order Rodentia is the most diverse among placental mammals: extant rodent species represent half of the placental diversity (2,277 species divided into 33 families) [1]. Morphological phylogenetic approaches have identified characters supporting a common origin (monophyly) of rodents, and clustered rodents and lagomorphs (rabbits, pikas) in a clade called Glires [2]. Morphological studies also generally agree on the number and content of rodent families [1,3,4]. However, the description of the relationships among rodent families has been confounded by rampant convergent evolution of morphological characters [5]. Based on morphological characters, rodents have been divided into either two or three subor-

ders. The first system, suggested by Brandt, divides rodents into three suborders, Myomorpha, Sciuromorpha, and Hystricomorpha, based on the position of masticatory muscles (the masseters) [6]. However, it has since been proven that this character is homoplasic and that this classification does not reflect evolutionary relationships [7,8]. The second system, proposed by Tullberg, divides rodents into two suborders, Sciurognathi and Hystricognathi, based on the position of the incisors and the angle of the jaw [9]. The monophyly of Hystricognathi has been accepted, based on the identification of additional morphological synapomorphies, but the Sciurognathi are usually considered to be paraphyletic [10]. Debates on the relationships within Sciurognathi and their relationships with Hystricognathi are the subject of numerous morphological papers [reviewed in [11]]. Molecular studies were expected to clarify the relationships among rodents. However, early studies based on molecular data complicated the understanding of rodent evolution by suggesting that rodents are paraphyletic [12-14]. These results initiated lively debates concerning evolutionary relationships among rodents and their place among placental mammals [15-17]. Phylogenetic conclusions supporting rodent paraphyly have been criticized, because they were based on a very limited taxonomic sampling. It has been suggested that increasing the sampling of rodent diversity [2] and/or mammalian diversity [18] would have supported rodent monophyly. Additionally, over-simplified models have been shown to erroneously support rodent paraphyly [19]. Recent analyses based on a representative sampling of rodent taxonomic diversity and using model-based methods of sequence analysis have strongly supported the monophyly of rodents [20-24].

Within Rodentia, molecular analyses suggest that rodents are divided into seven well-supported clades: 1-Anomaluromorpha (scaly-tailed flying squirrels, springhares), 2-Castoridae (beavers), 3-Ctenohystrica (gundi, porcupines, guinea-pigs), 4-Geomyoidea (pocket gophers, pocket mice), 5-Gliridae (dormice), 6-Myodonta (rats, mice, jerboas), and 7-Sciuroidea (mountain beavers, squirrels, woodchucks) [25-28]. However, several evolutionary relationships within Rodentia are still debated. Recent studies have suggested that these seven clades are clustered into three main lineages: 1 – Anomaluromorpha, Castoridae, Geomyoidea, and Myodonta together form the "mouse-related clade"; 2 – Sciuriodea and Gliridae form the "squirrel-related clade"; and 3 – Ctenohystrica forms the third lineage [29-32]. However, most studies have not been able to solve the relationships among these three clades. Recently, Montgelard et al. [32] analyzed mitochondrial genes as well as nuclear exonic and intronic sequences, and found significant support in favor of a basal position of the "mouse-related clade". This result was dependent on the removal of the fastest evolving characters from the dataset, suggesting that

mutational saturation might explain the inconclusive placement of the rodent root.

More generally, Rodriguez-Ezpeleta et al. [33] have shown that weakly supported nodes can sometimes be explained by the presence of conflicting phylogenetic and non-phylogenetic signal in a dataset. Three methods to reduce the non-phylogenetic information have been suggested: identification and removal of fast-evolving positions, character-recoding (e.g., RY coding), and the use of a site-heterogenous mixture model (e.g., CAT) [34].

Here, we aimed to resolve rodent relationships at the family level and above. We established a comprehensive dataset including six nuclear gene fragments from 41 rodent species together with eight outgroup species. We were able to solve most evolutionary relationships among rodent families. In order to minimize conflicting signals and thus solve the debated basal rodent relationships, we applied the three methods suggested by Rodriguez-Ezpeleta et al. [33]. We show that none of these methods, nor the use of more complex evolutionary models, can significantly solve basal rodent relationships. Additionally, some of our analyses, surprisingly, suggest a basal position of the squirrel-related clade and significantly reject the basal position of the "mouse-related clade" supported by Montgelard et al. [32]. We thus propose that the lack of resolution at the base of the rodent tree may reflect rapid rodent radiation, rather than conflicting phylogenetic signals.

Results and discussion
The rodent phylogeny
Maximum likelihood (ML) and Bayesian phylogenetic analyses, based on the combined nucleotide datasets, result in a well-resolved phylogeny (Figure 1), in agreement with the division of rodents into three major clades: the mouse-related clade (Bootstrap Percentage (BP) = 96, Posterior Probability (PP) = 1.0), the squirrel-related clade (BP = 86, PP = 1.0) and the Ctenohystrica (BP = 100, PP = 1.0).

The mouse-related clade
The mouse-related clade comprises three main lineages: Myodonta, Anomaluromorpha (Anomaluridae and Pedetidae), and Castorimorpha (Geomyoidea and Castoridae). The monophyly of this clade was first found in molecular studies [25,30-32] and later corroborated by a morphological analysis of extant and fossil taxa [8]. However, two recent molecular analyses cast doubt on the validity of the mouse-related clade. First, analysis of complete mitochondrial protein-coding genomes placed *Anomalurus* as a sister taxon of the Hystricognathi [24]. Second, structural analysis of B1 retroposon elements suggested that Castoridae could be an early diverging family within rodents [35]. Our analysis strongly rejects both of these possibilities. The best alternative to monophyly of
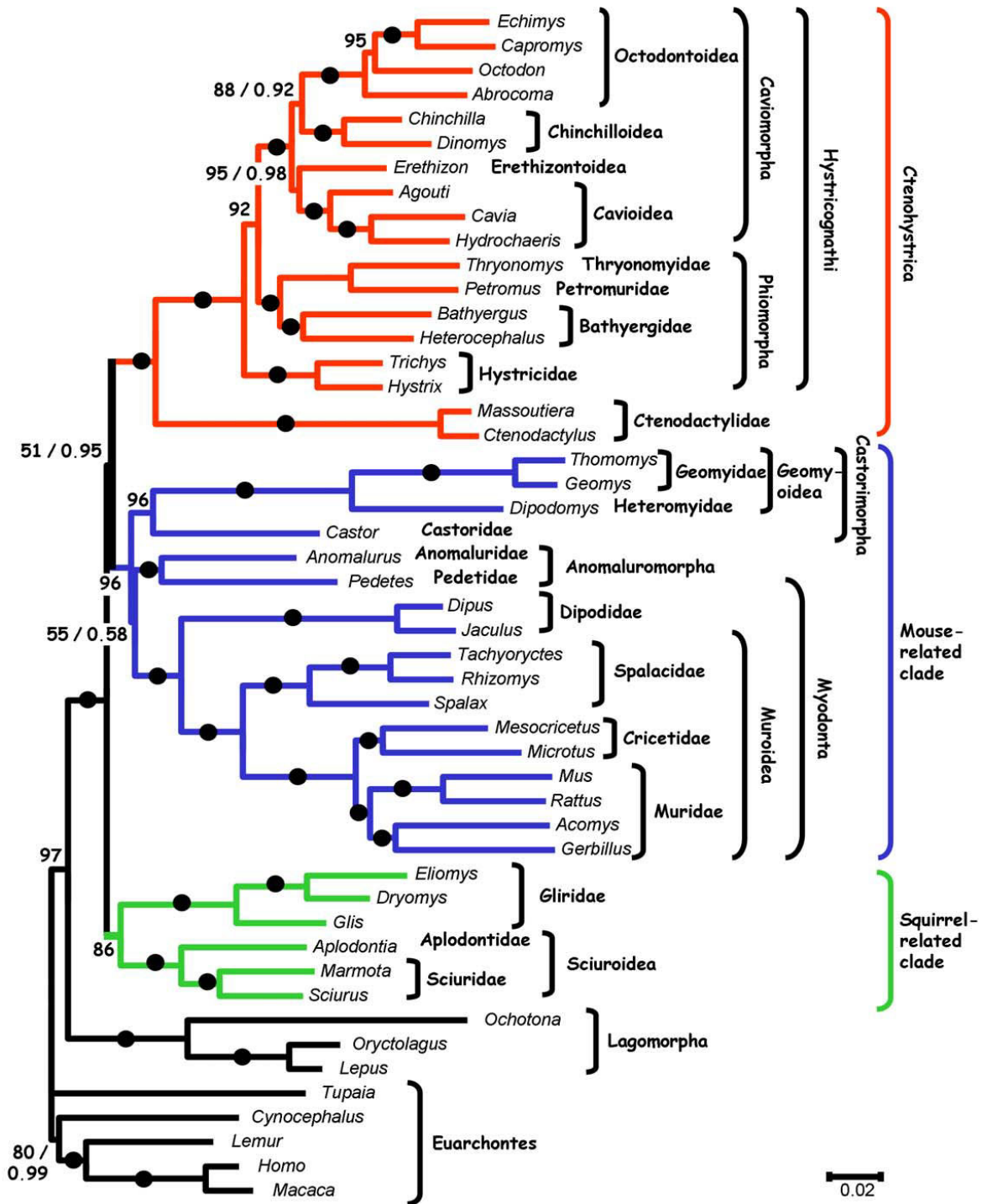
Figure 1
ML tree (-ln likelihood = 85,018.88) obtained for the concatenated dataset under the GTR+Γ+I model of sequence evolution. For each node the ML bootstrap percentage (BP) and the Bayesian posterior probabilities (PP) are given at the right and left of the slash, respectively. Branches with maximal support value (i.e., BP = 100, PP = 1.0) are indicated by black dots.

the mouse-related clade is significantly less likely than the ML tree, based on the approximately unbiased (AU) test (Table 1, *p*-value = 0.02). Analyses using RY coding or removal of third-codon positions, as well as partitioned analyses strongly support the monophyly of the mouse-related clade (Table 2, BP = 94–100). When protein sequences are analyzed, the monophyly of the mouse-related clade is still supported, albeit with lower bootstrap support (BP = 77). It is likely that the disagreement between our analysis and that of Horner et al. [24] stems from the fact that the latter was based on only six non-muroid species. With regard to the B1 retroposon study, the position of Castoridae presented by Veniaminova et al. [35] may be an artifact, because analysis of SINE insertion loci in rodents supports the monophyly of the mouse-related clade [36].

Previous phylogenetic reconstructions were unable to solve the relationships among the three main lineages of the mouse-related clade (Myodonta, Anomaluromorpha, and Castorimorpha), and all three possible evolutionary relationships have been suggested [22,25,26,28,31,32,36]. Our phylogenetic inference based on the full nucleotide dataset suggests the grouping of Anomaluromorpha with Myodonta (Figure 1). However, bootstrap and Bayesian support is at best moderate across the analyses considered (Table 2, BP = 37–72, PP = 0.58). In agreement with the bootstrap analysis, an AU test does not reject either alternative hypotheses (Table 1, *p*-value = 0.159–0.604). Additional data are thus needed to resolve the relationships at the base of the mouse-related clade. All other nodes within the mouse-related clade are well supported and alternatives are rejected based on an AU test (data not shown).

The squirrel-related clade
The grouping of Gliridae and Sciuridae has been recognized in morphological studies based on middle ear fea-

tures [37], arterial pattern [38], and by most molecular analyses. Nevertheless, high support values have seldom been obtained to support this relationship [22,25-27,29,31,32]. This node is well supported in our study (BP = 86, PP = 1.0). It is also supported in our analyses using different coding and partitions approaches (Table 2, BP = 86–98). However, alternatives to the monophyly of this clade are not rejected according to the AU test (Table 1, *p*-value = 0.123).

The Ctenohystrica
The clustering of Ctenodactylidae and Hystricognathi is highly supported (BP = 100, PP = 1.0). Previous knowledge of relationships within hystricognaths has been based either on a single gene (vWF or 12sRNA) for many hystricognath species (22–23 species) [39,40] or on multiple genes (3–6 genes) for fewer species (8–13 species) [22,29,32,41]. The present dataset expands that of Huchon et al. [29] by the addition of two nuclear gene fragments and four hystricognath taxa (in particular, a second representative of the Hystricidae). This expanded dataset allows us to solve the debated relationships within Hystricognathi. We find strong support for a basal position of Hystricidae within Hystricognathi (Figure 1, Table 2, BP = 89–95, PP = 1.0), while this position was previously only weakly supported [25,26,29,40]. However, AU tests do not reject alternative positions of Hystricidae (Table 1, *p*-value = 0.128).

Phylogenetic relationships among South-American hystricognaths (i.e., Caviomorpha) have long been debated. Caviomorphs have been found to comprise four distinct lineages (Cavioidea, Chinchilloidea, Erethizontoidea, and Octodontoidea) [40]. Our results confirm that chinchilla rats (Abrocoma) are not related to *Chinchilla* but rather belong to the Octodontoidea (BP = 100, PP = 1.0) [40,42,43]. Previous molecular trees did not resolve the relationships among the four caviomorph lineages with

Table 1: Results of likelihood-based tests of alternative topologies

Constraint	Diff -ln L	AU	KH	wSH
Unconstrained (ML tree)	<79199.95>	0.855	0.537	0.999
Ctenohystrica at the base of the rodent tree	0.5	0.650	0.463	0.917
Myodonta at the base of the mouse-related clade	1.1	0.604	0.386	0.928
Mouse-related clade at the base of the rodent tree	2.5	0.391	0.296	0.853
Anomaluromorpha at the base of the mouse-related clade	3.0	0.159	0.156	0.733
[Cavioidea+Erethizontoidea] not monophyletic	7.4	0.148	0.087	0.360
[Chinchilloidea+Octodontoidea] not monophyletic	8.0	0.206	0.149	0.471
Paraphyly of the squirrel-related clade	11.9	0.123	0.085	0.293
Caviomorpha at the base of Hystricognathi	12.2	0.128	0.079	0.278
Phiomorpha at the base of Hystricognathi	12.2	0.128	0.079	0.278
Paraphyly of the mouse-related clade	31.4	**0.024**	**0.017**	**0.050**
Paraphyly of Ctenohystrica	51.8	**3e-4**	**< 1e-10**	**< 1e-10**

Diff -ln L: observed log-likelihood difference between the ML topology and the alternative. AU: *p*-values of the approximately unbiased test. KH: *p*-values of the Kishino-Hasegawa test. wSH: *p*-values of the weighted Shimodaira – Hasegawa test. Significant *p*-values are indicated in bold.

Table 2: Maximum likelihood bootstrap support of main rodent relationships under different coding models

Coding model /Node	Unpartitioned models				Partitioned DNA models		
	All three codon positions	First two codon positions and third codon with RY coding	First two codon positions only	Protein sequences	A partition per gene (6 partitions)	A partition per codon position (3 partitions)	A partition per codon position per gene (18 partitions)
Squirrel-related clade at the base of the rodent tree	51	35	40	41	46	63	75
Mouse-related clade at the base of the rodent tree	15	28	less than 1	less than 1	17	9	15
Ctenohystrica at the base of the rodent tree	25	30	48	47	31	22	8
Monophyly of squirrel-related clade	86	90	89	98	92	92	95
Monophyly of mouse related clade	96	99	94	77	100	97	99
Monophyly of Ctenohystrica	100	100	99	100	100	100	100
[Anomaluromorpha + Myodonta] monophyly	55	54	65	37	50	72	48
Hystricidae at the base of the Hystricognathi	92	91	89	89	95	91	91

high bootstrap support, and various alternative topologies have been suggested [22,25,26,29,40]. Our data support a sister clade relationship between Cavioidea and Erethizontoidea (BP = 95, PP = 0.98), and a sister clade relationship between Chinchilloidea and Octodontoidea (Figure 1, BP = 88, PP = 0.92). In spite of these high support values, AU tests indicate that the best alternatives to these arrangements within Caviomorpha cannot be rejected (Table 1, *p*-value = 0.148–0.206). Similarly, analyses using RY coding or removal of third codon positions, as well as protein sequence analysis, support other relationships within Caviomorpha (data not shown). This suggests that additional species sampling is needed in order to robustly solve caviomorph relationships at the superfamily level.

Solving the base of the rodent tree
The most important unresolved relationship in rodent systematics is the one at the base of the rodent tree. To date, no phylogenetic analysis has been able to resolve this question with strong support, whether based on nucleotide sequence data [24,25,29,31], SINE data [36,44], or morphological data [8,10]. The only exception is the analysis of Montgelard et al. [32], which supports a basal position of the mouse-related clade after removal of fast-evolving nucleotide positions. Our nucleotide-based ML and Bayesian analyses (all three codon positions; Figure 1) place the squirrel-related clade at the base of the rodent tree. Our Bayesian analysis with the data partitioned by gene and partially partitioned by codon position (1st- and 2nd-position sites combined within genes, 3rd-position sites for each gene separate) appears to provide strong support for this relationship (PP > 0.90), but

the partitioned ML bootstrap support values are much lower (Table 2, BP = 51). It is possible for Bayesian PP values to be artificially inflated under circumstances of a near-trichotomy [45]. With this single Bayesian analysis being the only suggestion of strong support, and with the corresponding ML bootstrap support being so low, we hesitate to give much weight to the partitioned Bayesian result at the present time.
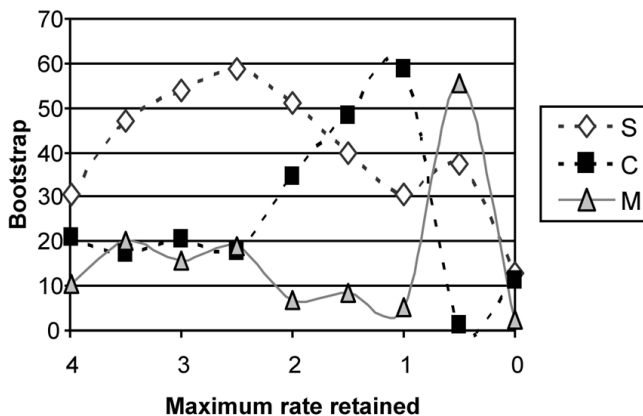
Recently, Rodriguez-Ezpeleta et al. [33] have shown that the presence of conflicting phylogenetic and non-phylogenetic signal in a dataset may result in weakly supported nodes. They suggested various approaches to remove the non-phylogenetic signal and thus increase the ability to resolve difficult phylogenetic relationships. To evaluate whether the resolution of the basal relationships among rodents could be improved by reducing non-phylogenetic signal in our dataset, we tested all the approaches suggested by Rodriguez-Ezpeleta et al. [33].

Character recoding
Third codon-position sites evolve the fastest, and are thus the most likely source of non-phylogenetic signal. In an attempt to reduce the non-phylogenetic signal, we performed analyses using RY coding for these positions. We also explored the extreme solution of removing all third codon-position sites. However, none of the three possible basal branching topologies was highly supported under these alternatives (Table 2). The only signal that can be seen is that a basal position of the mouse-related clade is not supported by the analysis of either the nucleotide dataset with only the first two codon positions, or the protein sequence dataset (Table 2, BP < 1).

Removal of fast-evolving positions

Nine datasets were delimited by retaining sites based on their inferred site-specific rates: (1) all sites (6,255 base pairs (bps); rates range from -0.698 to 3.989); (2) sites with rate ≤ 3.5 (6,114 bps); (3) sites with rate ≤ 3.0 (6,058 bps); (4) sites with rate ≤ 2.5 (5,997 bps); (5) sites with rate ≤ 2.0 (5,896 bps); (6) sites with rate ≤ 1.5 (5,759 bps); (7) sites with rate ≤ 1 (5,444 bps); (8) sites with rate ≤ 0.5 (4,997 bps); and (9) sites with rate ≤ 0.0 (4,179 bps). The bootstrap support as a function of the maximal evolutionary rate of site retained is presented in Figure 2. Removal of the fastest evolving sites (rate removed ≥ 2.5) improves the support in favor of a basal position of the squirrel-related from 30% to 59% while support for alternative topology remains below 25%. However, no clear trend can be found as bootstrap support remains below 60% in all analyses. It is worth noting that the topology supporting a basal position of the mouse-related clade is again the least supported, except for the dataset with maximum rate ≤ 0.5. We do not believe that this result effectively supports an early divergence of the mouse-related clade, because slight modification of the rate cutoff substantially changes the topology. For example, while the dataset with maximum rate ≤ 0.5 supports a basal position of the mouse-related clade, the dataset with maximum rate ≤ 0.6 supports a basal position of the Ctenohystrica. Note that

Montgelard et al. did not study the effect of varying their cut-off value. Finally, the support for all three topologies drops when sites with rate higher than zero are removed, possibly reflecting the fact that only 801 out of 2,858 informative characters remained in this dataset.

Similarly, seven datasets were considered by retaining sites based on their consistency index (CI): (1) all sites (6,255 bps; CI range 0.0625–1); (2) sites with CI > 0.1 (6,210 bps); (3) sites with CI > 0.2 (5,864 bps); (4) sites with CI > 0.3 (5,432 bps); (5) sites with CI > 0.4 (4,833 bps); (6) sites with CI > 0.5 (4,119 bps); and (7) sites with CI > 0.6 (4,023 bps). When retaining sites according to their maximal CI value (Figure 3), we observe an increase in the bootstrap support in favor of a basal position of the squirrel-related clade from 30.6% to 68.2%, which might suggest that this represents the phylogenetic signal. This support drops when sites with CI ≤ 0.6 are removed, which might come from the fact that only 626 informative characters remain in this dataset.

Use of site-heterogenous mixture model

The phylogenetic trees obtained under the CAT model did not help resolving the basal rodent relationships. The CAT analysis suggests that the squirrel-related clade is the first rodent lineage to diverge. However, no strong support in favor of this relationship is found, whether reconstruc-
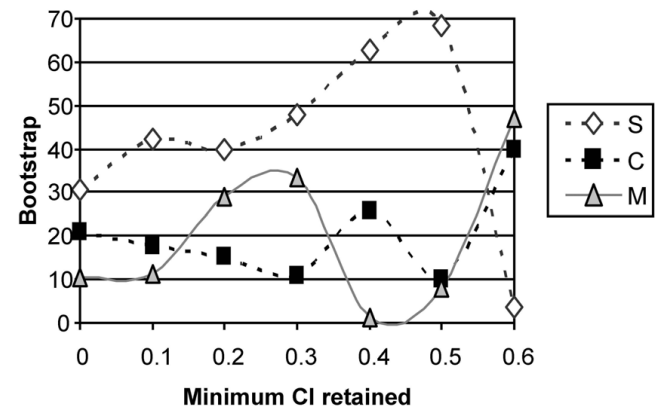

Figure 2
Bootstrap support as a function of the maximum evolutionary rate of site retained in the data. S: the best topology with the squirrel-related clade at the base of the rodent tree; M: the best topology with the mouse-related clade at the base of the rodent tree; C: the best topology with Ctenohystrica at the base of the rodent tree. Nine datasets were considered: (1) all sites (6,255 base pairs (bps); rates range from -0.698 to 3.989); (2) sites with rate ≤ 3.5 (6,114 bps); (3) sites with rate ≤ 3.0 (6,058 bps); (4) sites with rate ≤ 2.5 (5,997 bps); (5) sites with rate ≤ 2.0 (5,896 bps); (6) sites with rate ≤ 1.5 (5,759 bps); (7) sites with rate ≤ 1 (5,444 bps); (8) sites with rate ≤ 0.5 (4,997 bps); and (9) sites with rate ≤ 0.0 (4,179 bps).

Figure 3
Bootstrap support as a function of the minimum CI of sites retained in the data. S: the best topology with the squirrel-related clade at the base of the rodent tree; M: the best topology with the mouse-related clade at the base of the rodent tree; C: the best topology with Ctenohystrica at the base of the rodent tree. Seven datasets were considered: (1) all sites (6,255 bps; CI range 0.0625–1); (2) sites with CI > 0.1 (6,210 bps); (3) sites with CI > 0.2 (5,864 bps); (4) sites with CI > 0.3 (5,432 bps); (5) sites with CI > 0.4 (4,833 bps); (6) sites with CI > 0.5 (4,119 bps); and (7) sites with CI > 0.6 (4,023 bps).

Table 3: Maximum log-likelihood scores and AU test *p*-values under different models of sequence evolution for three possible basal rodent relationships.

Model of sequence evolution	JTT + Γ_4		Rate shift model + Γ_4		Codon model + Γ_4 (without positive selection)	
Topology	Diff -ln L	*p*-value	Diff -ln L	*p*-value	Diff -ln L	*p*-value
Squirrel-related clade at the base	<32,963.68>	0.580	<32,936.35>	0.586	0.9	0.517
Ctenohystrica at the base	0.2	0.563	0.3	0.558	<77,788.57>	0.623
Mouse related clade at the base	3.2	**0.031**	3.1	**0.035**	4.0	0.158

Significant *p*-values are indicated in bold. The log-likelihood value of the best tree is indicated between brackets.

tions are based on nucleotide or protein sequences ($PP_{DNA} = 0.57$; $PP_{PROTEIN} = 0.75$).

Use of complex evolutionary models for protein sequences
The use of more complex evolutionary models did not completely solve basal rodent relationships. Again, a basal position of the mouse-related clade is generally the least likely, and this hypothesis is even rejected using AU tests under either the JTT+Γ model or the rate-shift model (Table 3). However, a basal position of the Ctenohystrica cannot be excluded. This finding is in agreement with both the nucleotide analysis based on the first two codon positions and the nucleotide analysis with fast-evolving sites removed.

The nucleotide sequences were also analyzed using codon models. No support for positive selection was found, and hence, we only report the results obtained using the M8a model, which does not allow sites to evolve under positive selection. Under this model, a basal position of the Ctenohystrica is the most likely. However, the fit of the data to this topology is not significantly better than alternative topologies (0.9 and 4.0 log-likelihood point differences, for the topology with a basal position of the squirrel-related clade and the topology with a basal position of the mouse-related clade, respectively). The three possible rootings of the rodent tree are thus not statistically different based on AU tests (Table 3).

Conclusion
Our phylogenetic reconstructions provide a well-resolved rodent tree, except for a few nodes and the basal relationships among the main rodent clades. Unlike Montgelard et al. [32], removing fast evolving characters did not improve the resolution at the base of the rodent tree. This lack of resolution remained when all the other methods suggested by Rodriguez-Ezpeleta et al. [33] to increase tree resolution were applied. Surprisingly, using the JTT and the rate-shift models, we were able to reject a basal position of the mouse-related clade supported by Montgelard et al. [32] and support instead a basal position of the squirrel-related clade (a topology rejected by Montgelard et al. [32]). This suggests that removing fast evolving positions is not a panacea to solve phylogenetic conflicts,

since different datasets can lead to significantly different results when using this approach.

More generally, our results suggest that the low support at the base of the rodent tree cannot be attributed only to the presence of conflicting non-phylogenetic signal, since removing such non-phylogenetic signal failed to significantly increase the tree resolution. We thus hypothesize that this lack of resolution reflects rapid radiation at the base of the rodent tree and possibly incomplete lineage sorting. Indeed, rodents were already highly diversified in the Paleocene and Early Eocene. Many extinct families are identified in these geological periods (i.e., Decipomyidae, Alagomyidae, Ivanantoniidae, Sciuravidae, Ischyromyidae, Theridomorpha, and Yuomyidae). According to recent phylogenetic work based on fossils and extant taxa [8], some of these ancient families are sister clades of extant clades. In particular, Theridomorpha might be related to Sciuroidea, Sciuravidae to the mouse-related clade, and Yuomyidae to the Ctenohystrica [8]. This supports the idea that the divergences among Ctenohystrica, the mouse-related clade, and the squirrel-related clade occurred during the explosive radiation of rodents in the Paleocene.

Our results further suggest that a basal position of the mouse-related clade is the least likely, while a basal position of the squirrel-related clade may be the most likely. Interestingly, structural analysis of B1 retroposon elements also provides additional support in favor of an early divergence of the squirrel-related clade [35,44]. The basal position of the squirrel-related clade may further be supported by the fact that the earliest fossils representative of the Gliridae, Aplodontidae, and Sciuridae families are protrogomorphous, while most early Ctenohystrica and most early representatives of the mouse-related clade are hystricomorphous [see review of character states in [8]]. Consequently, an early divergence of the squirrel-related clade appears to be the most parsimonious evolutionary scenario, given our current knowledge.

Methods
Taxon sampling
All suborders and super families of the order Rodentia listed by Carleton and Musser [46] are included in the

analysis (Additional file 1). The tree was rooted with the closest rodent outgroups: representative lagomorphs, representative primates, Cynocephalus (the flying lemur, order Dermoptera), and Tupaia (tree shrew, order Scandentia). Rodents together with lagomorphs, primates, flying lemurs, and tree shrews form a clade called Euarchontoglires or Supraprimates [21,47-49].

DNA amplification and sequencing

Ethanol-preserved samples, frozen tissue samples, or previously purified genomic DNAs were obtained from the donor institutions listed in Additional file 2. Total DNA was extracted according to Sambrook, Fritsch, and Maniatis [50] with slight modifications. Fragments from the following six nuclear genes were sequenced: the alpha 2B adrenergic receptor (ADRA2B); the cannabinoid receptor 1 (CB1); the growth hormone receptor (GHR); the interphotoreceptor retinoid binding protein (IRBP); the recombination activating gene 2 (RAG2); and the von Willebrand factor (vWF). These nuclear genes were chosen for the following reasons: (i) a large number of sequences are already available for those genes, especially within rodents; (ii) these genes have been shown to contain phylogenetic information within rodents and between mammalian orders [21,26,30]; (iii) these genes are not genetically linked to one another (their location is variable, on chromosomes 2, 4, 15, 14, 2, and 6 in Mus, chromosomes 3, 5, 2, 16, 3, and 4 in Rattus, and chromosomes 2, 6, 5, 10, 11, and 12 in Homo); and (iv) no interactions among these proteins were previously reported.

Amplification of ADRA2B, IRBP, and vWF was performed as described in Huchon et al [29]. Amplification of CB1 was performed in two steps. A first amplification was performed with primers CB1-D1: 5'-GGCTCAAATGACAT-TCAGTACGAA-3' and CB1-R1: 5'-GAGTCCCCCATGCTGTTATCTAGAGGCTG-3', followed by a re-amplification of the initial PCR product using primers CB1-D2: 5'-CAGTACGAAGATATCAAAGGAGA-

CATGGC-3' and CB1-R2: 5'-GAGTCCCCCATGCTGT-TATCTAGAGGCTG-3'. Amplification of RAG2 and GHR was performed similarly. For RAG2, the first amplification was performed with primers RAG2-D1 5'-CGCTGCACA-GAGAAAGACTT-3' and RAG2-R1: 5'-AAGGATTTCTT-GGCAGGAGT-3', followed by a re-amplification of the initial PCR product using primers RAG2-D2: 5'-TAYAGY-CGAGGGAAAAGYATGGG-3' and RAG2-R2: 5'-GACAAGTGGATGAGTGTGCGTTC-3'. For GHR the first amplification was performed with primers GHR-D1 5'-TAGGAAGGAAAATTRGARGARGTNAA-3'and GHR-R1: 5'-AAGGCTANGGCATGATRTTRTT-3', followed by a re-amplification of the initial PCR product using primers GHR-D2: 5'-GGAAAATTRGAGGAGGTGAAYACNATHTT-3' and GHR-R2: 5'-GATTTTGTTCAGTTGGTCRGTRCT-NAC-3' or GHR-R1. Purification of the PCR products and sequencing were performed according to Huchon et al. [29]. Sequence accession numbers are available in Additional file 1.

Sequence alignment

DNA sequences were translated and the corresponding protein sequences were aligned using both PROBCONS [51] and MAFFT [52]. PROBCONS alignments were conducted with three consistency steps and 500 iterative refinement repetitions. MAFFT alignments were conducted with the L-INS-i option. The positions that differed between both alignments were removed using SOAP [53]. The DNA alignments were then computed based on the protein alignments using the program PAL2NAL [54]. The number of DNA positions included in each gene partition after using SOAP is indicated in Table 4. The DNA and protein sequence alignments are provided in Additional file 3 and Additional file 4, respectively.

ML analyses of the concatenated dataset

Phylogenetic tree reconstructions were performed on the concatenated nucleotide dataset using the ML criterion. The program MODELTEST 3.07 [55] was used to deter-

Table 4: Number of positions in each gene partition.

	Constant	Variable uninformative	Informative	Total
A2AB	500	98	464	1,062
CB1	694	70	319	1,083
GHR	219	103	470	792
IRBP	440	172	627	1,239
RAG2	430	111	365	906
vWF	388	172	613	1,173
Total	2,671	726	2,858	6,255

mine the best probabilistic model of DNA sequence evolution using the Akaike Information Criterion. The best model was found to be GTR+Γ+I. ML searches for the best trees were performed using the program PAUP* [56]. The parameters of the model and the ML tree were then determined by successive approximation [57]. The initial parameter values were those estimated by MODELTEST 3.07, and those values were used for a first round of heuristic search starting with a Neighbor-Joining (NJ) tree and using TBR branch-swapping. Parameters were then estimated on the resulting tree and used for another round of heuristic search. The process was repeated until all parameter values were stable. Bootstrap percentages were estimated from 100 pseudo-replicates using the best estimated parameters, a NJ starting tree, and TBR branch-swapping.

Phylogenetic trees were also reconstructed based on protein sequences. The protein sequence alignment is provided in Additional file 4. The program PROTTEST 1.3 [58] was used to estimate the best model of protein sequence evolution. The best model was found to be JTT+Γ+I. Phylogenetic trees were then reconstructed with the program PHYML [59] using the ML model identified by PROTTEST.

ML analyses of the partitioned dataset
Three different partitioned ML-analyses were conducted on the nucleotide dataset with RAxML [60]. The first analysis considered each gene as an independent partition (six partitions). The second analysis considered each codon position as an independent partition (three partitions). The third analysis considered each codon position of each gene as an independent partition (18 partitions). The GTR+Γ+I model was applied to all partitions, individual α-shape parameters, substitution rates, and base frequencies were estimated and optimized separately for each partition. Bootstrap support was estimated using 100 pseudo-replicates.

Bayesian analyses
Bayesian analyses were performed on the nucleotide dataset using the program MrBayes v3.1.2 [61]. Prior distributions for parameters in the Bayesian analyses were: topology, uniform; branch lengths, exponential (λ = 10); alpha parameter of the Γ distribution, uniform (0.05,50.0); p_{inv}, uniform (0,1); κ, beta (1.0,1.0); R-matrix, Dirichlet (1,1,1,1,1,1); base frequencies, Dirichlet (1,1,1,1). Each run included two independently started chains, each beginning with a different, randomly chosen tree. From each starting tree, four related MCMC chains (one cold and three incrementally heated) were run. The temperature parameter λ was set to produce chain-swap frequencies in the range of 10–30%. Posterior distribution estimates were based on sampling the cold chain

every 250 generations. Initial runs were allowed to stop at 10^7 cycles if the percent standard deviation among bifurcation split probabilities for the two separate chains was less than 0.01. In those cases, the first 5×10^6 cycles were discarded as burn-in. Additional longer runs were performed if needed, with the first 50% of samples discarded as burn-in. The dataset was divided into 12 partitions, two for each gene. For each gene, the first- and second-position sites were combined into a single partition, and the third-position sites in a separate partition. Analyses were conducted with each partition assigned either the HKY85+Γ or the GTR+Γ model, with the exception of the 1st- plus 2nd-position site partition for CB1, which used either HKY85+I or GTR+I, because of the extremely low number of variable sites in that partition. The GTR models were preferred by Bayes' Factors, while the HKY85 models were favored by the Bayesian information criterion (BIC). However, both models gave very similar results, as did a separate set of analyses with the CB1 1st- plus 2nd-position sites excluded.

Bayesian analyses under the CAT +Γ_4 model were performed using the program Phylobayes 2.1c [34,62]. Both for the DNA and protein datasets, two chains were run for 100,000 cycles and trees were sampled every 100 cycles after the first 25,000 cycles. As recommended in Phylobayes, the maximum difference in bipartition frequencies between the two chains was below 0.1, indicating a "good run" (for DNA, maxdiff = 0.051; for protein, maxdiff = 0.044). The phylogenetic trees obtained under the CAT +Γ_4 model are available in Additional file 5.

Testing alternative hypotheses
The ML tree was compared to several constrained topologies using various likelihood-based tests as implemented in the program CONSEL v0.1i [63]. Eleven alternative topologies were considered. 1 – The best tree placing the Ctenohystrica at the base of Rodentia. 2 – The best tree placing the mouse-related clade at the base of Rodentia. 3 – The best tree placing Myodonta at the base of the mouse-related clade. 4 – The best tree placing Anomaluromorpha at the base of the mouse-related clade. 5 – The best alternative that does not support monophyly of Chinchilloidea+Octodontoidea. 6 – The best alternative that does not support monophyly of Cavioioidea+Erethizontoidea. 7 – The best tree placing Phiomorpha at the base of the Hystricognathi. 8 – The best tree placing Caviomorpha at the base of the Hystricognathi. 9 – The best alternative that does not support monophyly of the mouse-related clade. 10 – The best alternative that does not support monophyly of the squirrel-related clade. 11 – The best alternative that does not support monophyly of the Ctenohystrica. The best alternatives were built using constrained ML heuristic searches. Each search was conducted starting with an NJ tree, using the TBR branch-

swapping option, and the parameters of the unconstrained ML tree. Site-wise log-likelihoods were computed with PAUP* using the parameters of the best ML tree.

Removal of fast-evolving positions

Following the approach of Rodriguez-Ezpeleta et al. [33], fast evolving sites were determined according to their site-wise rates calculated with the program Rate4Site [64] using the Tamura-Nei substitution model [65] with 16 discrete rate categories used to approximate the Gamma distribution. Rates were computed for three topologies: the ML tree topology obtained as describe above (i.e., the topology with the squirrel-related clade at the base), and the two alternative topologies (the mouse-related clade at the base and the Ctenohystrica at the base). All other nodes were identical between the topologies. Nucleotide sites were classified according to their average rate over the three topologies. Rates were normalized so that the average rate across all sites was 0. Rates ranged between – 0.698 to 3.989. The fastest-evolving sites were then progressively removed to create nine datasets: all sites; sites with rates ≤ 3.5; sites with rates ≤ 3.0; sites with rates ≤ 2.5; sites with rates ≤ 2.0; sites with rates ≤ 1.5; sites with rates ≤ 1.0; sites with rates ≤ 0.5; and sites with rates ≤ 0. Bootstrap values were computed with the program Treefinder [66]. For each dataset, the ML tree and parameters were estimated by Treefinder under the GTR+Γ+I model. These ML parameters were then used to perform a bootstrap analysis using 500 replicates.

Removing sites based on their evolutionary rate does not allow differentiation between sites with few character states (e.g. sites with only purines or only pyrimidines) from sites with all possible character states (e.g. sites with all four bases). However, if two positions evolve under the same substitution rate, we can expect the one with more character states to be less homoplasious. Consequently, we used CI values as a measure of the level of homoplasy. It is worth noting that CI and site specific rates are weekly correlated (Additional file 6). For each of the three topologies described above, the CI of each site was computed using PAUP* and sites were classified according to their average CI values across the three topologies. Seven datasets were constructed by eliminating some sites based on their CI values (retaining all sites; retaining sites with CI > 0.1; retaining sites with CI > 0.2; retaining sites with CI > 0.3; retaining sites with CI > 0.4; retaining sites with CI > 0.5; and retaining sites with CI > 0.6). Bootstrap values were then computed as described above.

Character recoding

Since the third codon position generally evolves at the highest rate, we performed two types of ML analyses that were intended to reduce the impact of this high rate. Phylogenetic trees were reconstructed either without third codon position or using RY coding at the third codon

position. The program MODELTEST 3.07 was used to infer the model of sequence evolution and phylogenetic trees were reconstructed using PAUP* as described above.

Use of complex evolutionary models

It has been shown that using codon-based models can improve phylogenetic inference from protein-coding genes [67]. This method is computationally intensive, thus precluding the computation of bootstrap values for a dataset of 49 species. To this end, maximal log-likelihoods of the three topologies involving basal rodent relationships were compared using the program CONSEL [63]. For this comparison, site-wise log-likelihoods of each topology under two codon models were computed using the program SELECTON version 2.3 [68]. The first model, M8, allows for positive selection operating on the protein [69]. The second model, M8a, does not allow for positive selection [70]. In both codon analyses, four Gamma-rate categories were used to account for among-site rate variation.

Heterotachy (covarion) is define as the variation of the evolutionary rate within a given site, and it was shown to have an impact on phylogenetic inferences [e.g., [71]]. Consequently, the covarion model developed by Galtier [72], extended to amino-acids [73] was used to compute the site-wise log-likelihoods of the three topologies involving basal rodent relationships. Four Gamma-rate categories were used to account for among-site rate variation. The C++ code for computing site-wise log-likelihoods under the covarion model is available from the authors upon request.

The support for each topology under the codon and covarion models was compared to the support under the JTT model. The program Rate4Site was used to obtain the site-wise log-likelihood of the three topologies tested, and these site-wise log-likelihoods were used as input to CONSEL.

Authors' contributions

SBK and HM carried out the sequencing and participated in the sequence alignment and analyses. OP conducted the phylogenetic analysis based on complex evolutionary models. TP supervised the analysis based on complex evolutionary models. RWD and DH conceived of the study, participated in its design and coordination, and wrote the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1
Accession number table. List of all specimens and accession numbers for each gene sequenced.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-71-S1.doc]

References

1. Wilson DE, Reeder DM: *Mammal species of the world: a taxonomic and geographic reference* 3rd edition. Baltimore, MD, Johns Hopkins University Press; 2005.

2. Luckett WP, Hartenberger J-L: **Monophyly or polyphyly of the order Rodentia: possible conflict between morphological and molecular interpretations.** *J Mammal Evol* 1993, **1(2):**127-147.

3. Hartenberger J-L: **The order Rodentia: major questions on their evolutionary origin, relationships and suprafamilial systematics.** In *Evolutionary relationships among rodents: a multidisciplinary analysis* Edited by: Luckett WP, Hartenberger J-L. New York and London, Plenum Press; 1985:1-33.

4. McKenna MC, Bell SK: *Classification of mammals above the species level* New York, Columbia University Press; 1997.

5. Jaeger J-J: **Rodent phylogeny: new data and old problems.** In *The phylogeny and classification of the Tetrapods Volume 2.* Edited by: Benton MJ. Oxford, Clarendon Press; 1988:177-199.

6. Brandt JF: **Beiträge zur nahern Kenntniss der Säugethiere Russlands.** *Mem Acad Imp St Petersbourg Ser* 1855, **69:**1-375.

7. Vianey-Liaud M: **Possible evolutionary relationships among Eocene and Lower Oligocene rodents of Asia, Europe and North America.** In *Evolutionary relationships among rodents: a multidisciplinary analysis* Edited by: Luckett WP, Hartenberger J-L. New York and London, Plenum Press; 1985:277-309.

8. Marivaux L, Vianey-Liaud M, Jaeger JJ: **High-level phylogeny of early Tertiary rodents: dental evidence.** *Zool J Linn Soc* 2004, **142(1):**105-134.

9. Tullberg T: **Ueber das System der Nagetiere: Eine phylogenetische Studie.** *Nova Acta Reg Soc Sci Upsala Ser3* 1899, **18:**1-514.

10. Luckett WP, Hartenberger J-L: **Evolutionary relationships among rodents: comment and conclusions.** In *Evolutionary relationships among rodents: a multidisciplinary analysis* Edited by: Luckett WP, Hartenberger J-L. New York and London, Plenum Press; 1985:227-276.

11. Luckett WP, Hartenberger J-L: *Evolutionary relationships among rodents a multidisciplinary analysis* New York and London, Plenum Press; 1985.

12. D'Erchia AM, Gissi C, Pesole G, Saccone C, Arnason U: **The guinea-pig is not a rodent.** *Nature* 1996, **381:**597-600.

13. Graur D, Hide WA, Li W-H: **Is the guinea-pig a rodent?** *Nature* 1991, **351:**649-652.

14. Reyes A, Gissi C, Pesole G, Catzeflis F, Saccone C: **Where do rodents fit? Evidence from the complete mitochondrial genome of** *Sciurus vulgaris*. *Mol Biol Evol* 2000, **17(6):**979-983.

15. Graur D: **Reply from D. Graur.** *Trends Ecol Evol* 1993, **8:**341-342.

16. Novacek MJ: **Mammalian phylogeny: morphology and molecules.** *Trends Ecol Evol* 1993, **8:**339-340.

17. Catzeflis FM: **Mammalian phylogeny: morphology and molecules.** *Trends Ecol Evol* 1993, **8:**340-341.

18. Philippe H: **Rodent monophyly: pitfalls of molecular phylogenies.** *J Mol Evol* 1997, **45:**712-715.

19. Sullivan J, Swofford DL: **Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics.** *J Mammal Evol* 1997, **4(2):**77-86.

20. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ: **Molecular phylogenetics and the origins of placental mammals.** *Nature* 2001, **409:**614-618.

21. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, *et al.*: **Resolution of the early placental mammal radiation using Bayesian phylogenetics.** *Science* 2001, **294(5550):**2348-2351.

22. Huchon D, Catzeflis FM, Douzery EJP: **Variance of molecular datings, evolution of rodents, and the phylogenetic affinities between Ctenodactylidae and Hystricognathi.** *Proc Biol Sci* 2000, **267(1441):**393-402.

23. Reyes A, Gissi C, Catzeflis F, Nevo E, Pesole G, Saccone C: **Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods.** *Mol Biol Evol* 2004, **21(2):**397-403.

24. Horner DS, Lefkimmiatis K, Reyes A, Gissi C, Saccone C, Pesole G: **Phylogenetic analyses of complete mitochondrial genome sequences suggest a basal divergence of the enigmatic rodent** *Anomalurus*. *BMC Evol Biol* 2007, **7:**16.

25. Adkins RM, Walton AH, Honeycutt RL: **Higher-level systematics of rodents and divergence time estimates based on two congruent nuclear genes.** *Mol Phylogenet Evol* 2003, **26(3):**409-420.

26. Adkins RM, Gelke EL, Rowe D, Honeycutt RL: **Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes.** *Mol Biol Evol* 2001, **18(5):**777-791.

27. Montgelard C, Bentz S, Tirard C, Verneau O, Catzeflis FM: **Molecular systematics of Sciurognathi (Rodentia): The mitochon-**

drial cytochrome *b* and 12S rRNA genes support the Anomaluroidea (Pedetidae and Anomaluridae). *Mol Phylogenet Evol* 2002, **22(2)**:220-233.

28. DeBry RW, Sagel RM: **Phylogeny of Rodentia (Mammalia) inferred from the nuclear-encoded gene IRBP.** *Mol Phylogenet Evol* 2001, **19(2)**:290-301.

29. Huchon D, Chevret P, Jordan U, Kilpatrick CW, Ranwez V, Jenkins PD, Brosius J, Schmitz J: **Multiple molecular evidences for a living mammalian fossil.** *Proc Natl Acad Sci USA* 2007, **104(18)**:7495-7499.

30. Huchon D, Madsen O, Sibbald MJJB, Ament K, Stanhope M, Catzeflis F, De Jong WW, Douzery EJP: **Rodent phylogeny and a timescale for the evolution of glires: evidence from an extensive taxon sampling using three nuclear genes.** *Mol Biol Evol* 2002, **19(7)**:1053-1065.

31. DeBry RW: **Identifying conflicting signal in a multigene analysis reveals a highly resolved tree: The phylogeny of Rodentia (Mammalia).** *Syst Biol* 2003, **52(5)**:604-617.

32. Montgelard C, Forty E, Arnal V, Matthee CA: **Suprafamilial relationships among Rodentia and the phylogenetic effect of removing fast-evolving nucleotides in mitochondrial, exon and intron fragments.** *BMC Evol Biol* 2008, **8**:321.

33. Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H: **Detecting and overcoming systematic errors in genome-scale phylogenies.** *Syst Biol* 2007, **56(3)**:389-399.

34. Lartillot N, Philippe H: **A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process.** *Mol Biol Evol* 2004, **21(6)**:1095-1109.

35. Veniaminova NA, Vassetzky NS, Lavrenchenko LA, Popov SV, Kramerov DA: **Phylogeny of the order rodentia inferred from structural analysis of short retroposon B1.** *Russ J Genet* 2007, **43(7)**:757-768.

36. Farwick A, Jordan U, Fuellen G, Huchon D, Catzeflis F, Brosius J, Schmitz J: **Automated scanning for phylogenetically informative transposed elements in rodents.** *Syst Biol* 2006, **55(6)**:936-948.

37. Lavocat R, Parent J-P: **Phylogenetic analyses of middle ear features in fossil and living rodents.** In *Evolutionary relationships among rodents: a multidisciplinary analysis* Edited by: Luckett WP, Hartenberger J-L. New York and London, Plenum Press; 1985:333-354.

38. Bugge J: **Systematic value of the carotid arterial pattern in rodents.** In *Evolutionary relationships among rodents: a multidisciplinary analysis* Edited by: Luckett WP, Hartenberger J-L. New York and London, Plenum Press; 1985:381-402.

39. Nedbal MA, Allard MW, Honeycutt RL: **Molecular systematics of the hystricognath rodents: evidence from the mitochondrial 12S rRNA gene.** *Mol Phylogenet Evol* 1994, **3(3)**:206-220.

40. Huchon D, Douzery EJP: **From the old-world to the new-world: a molecular chronicle of the phylogeny and biogeography of hystricognath rodents.** *Mol Phylogenet Evol* 2001, **20(2)**:238-251.

41. Poux C, Chevret P, Huchon D, de Jong WW, Douzery EJP: **Arrival and diversification of caviomorph rodents and platyrrhine primates in South America.** *Syst Biol* 2006, **55(2)**:228-244.

42. Opazo JC, Palma RE, Melo F, Lessa EP: **Adaptive evolution of the insulin gene in caviomorph rodents.** *Mol Biol Evol* 2005, **22(5)**:1290-1298.

43. Köhler N, Gallardo MH, Contreras LC, Torres-Mura JC: **Allozymic variation and systematic relationships of the Octodontidae and allied taxa (Mammalia, Rodentia).** *J Zool* 2000, **252**:243-250.

44. Veniaminova NA, Vassetzky NS, Kramerov DA: **B1SINEs in different rodent families.** *Genomics* 2007, **89(6)**:678-686.

45. Lewis PO, Holder MT, Holsinger KE: **Polytomies and Bayesian phylogenetic inference.** *Syst Biol* 2005, **54(2)**:241-253.

46. Carleton MD, Musser GG: **Order Rodentia.** In *Mammal species of the world: a taxonomic and geographic reference Volume 2*. 3rd edition. Edited by: Wilson DE, Reeder DM. Baltimore, MD, Johns Hopkins University Press; 2005:745-752.

47. de Jong WW, van Dijk MA, Poux C, Kappe G, van Rheede T, Madsen O: **Indels in protein-coding sequences of Euarchontoglires constrain the rooting of the eutherian tree.** *Mol Phylogenet Evol* 2003, **28(2)**:328-340.

48. Nishihara H, Hasegawa M, Okada N: **Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions.** *Proc Natl Acad Sci USA* 2006, **103(26)**:9929-9934.

49. Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J: **Retroposed elements as archives for the evolutionary history of placental mammals.** *PLoS Biol* 2006, **4(4)**:e91.

50. Sambrook J, Fritsch EF, Maniatis T: *Molecular cloning: a laboratory manual* 2nd edition. New York, Cold Spring Harbor Laboratory Press; 1989.

51. Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S: **ProbCons: probabilistic consistency-based multiple sequence alignment.** *Genome Res* 2005, **15(2)**:330-340.

52. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res* 2005, **33(2)**:511-518.

53. Löytynoja A, Milinkovitch MC: **SOAP, cleaning multiple alignments from unstable blocks.** *Bioinformatics* 2001, **17(6)**:573-574.

54. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.** *Nucleic Acids Res* 2006, **34**:W609-W612.

55. Posada D, Crandall KA: **MODELTEST: testing the model of DNA substitution.** *Bioinformatics* 1998, **14**:817-818.

56. Swofford DL: **PAUP*: Phylogenetic analysis using parsimony (* and other methods).** In *version 4.0b10 edn* Sunderland, Massachusetts: Sinauer Associates; 2003.

57. Sullivan J, Abdo Z, Joyce P, Swofford DL: **Evaluating the performance of a successive-approximations approach to parameter optimization in maximum-likelihood phylogeny estimation.** *Mol Biol Evol* 2005, **22(6)**:1386-1392.

58. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21(9)**:2104-2105.

59. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52(5)**:696-704.

60. Stamatakis A: **RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22(21)**:2688-2690.

61. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19(12)**:1572-1574.

62. Lartillot N, Philippe H: **Computing Bayes factors using thermodynamic integration.** *Syst Biol* 2006, **55(2)**:195-207.

63. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection.** *Bioinformatics* 2001, **17(12)**:1246-1247.

64. Mayrose I, Graur D, Ben-Tal N, Pupko T: **Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior.** *Mol Biol Evol* 2004, **21(9)**:1781-1791.

65. Tamura K, Nei M: **Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.** *Mol Biol Evol* 1993, **10**:512-529.

66. Jobb G, von Haeseler A, Strimmer K: **TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics.** *BMC Evol Biol* 2004, **4(1)**:18.

67. Ren FR, Tanaka H, Yang ZH: **An empirical examination of the utility of codon-substitution models in phylogeny reconstruction.** *Syst Biol* 2005, **54(5)**:808-818.

68. Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T: **Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach.** *Nucleic Acids Res* 2007, **35**:W506-W511.

69. Yang ZH, Nielsen R, Goldman N, Pedersen AMK: **Codon-substitution models for heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155(1)**:431-449.

70. Swanson WJ, Nielsen R, Yang QF: **Pervasive adaptive evolution in mammalian fertilization proteins.** *Mol Biol Evol* 2003, **20(1)**:18-20.

71. Inagaki Y, Susko E, Fast NM, Roger AJ: **Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1 alpha phylogenies.** *Mol Biol Evol* 2004, **21(7)**:1340-1349.

72. Galtier N: **Maximum-likelihood phylogenetic analysis under a covarion-like model.** *Mol Biol Evol* 2001, **18(5)**:866-873.

73. Penn O, Stern A, Rubinstein ND, Dutheil J, Bacharach E, Galtier N, Pupko T: **Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes.** *PLoS Comput Biol* 2008, **4(11)**:e1000214.