Interrogating Genomic Data in the Phylogenetic Placement of Treeshrews Reveals Potential Sources of Conflict.

Authors

Alexander Knyshov[1,*], Yana Hrytsenko[1], Robert Literman[2], and Rachel S. Schwartz[1]

1. Department of Biological Sciences, University of Rhode Island, South Kingstown, RI, USA
2. Center for Food Safety and Applied Nutrition, Office of Regulatory Science, U.S. Food and Drug Administration, College Park, MD, USA

* Corresponding author, e-mail: aknyshov@uri.edu

Abstract

The position of some taxa on the Tree of Life remains controversial despite the increase in genomic data used to infer phylogenies. While analyzing large datasets alleviates stochastic errors, it does not prevent systematic errors in inference, caused by both biological (e.g., incomplete lineage sorting, hybridization) and methodological (e.g., incorrect modeling, erroneous orthology assessments) factors. In our study, we systematically investigated factors that could result in these controversies, using the treeshrew (Scandentia, Mammalia) as a study case. Recent studies have narrowed the phylogenetic position of treeshrews to three competing hypotheses: sister to primates and flying lemurs (Primatomorpha), sister to rodents and lagomorphs (Glires), or sister to a clade comprising all of these. We sampled 50 mammal species including three treeshrews, a selection of taxa from the potential sister groups, and outgroups. Using a large diverse set of loci, we assessed support for the alternative phylogenetic position of

treeshrews. A plurality of loci support treeshrews as sister to rodents and lagomorphs; however, only a few loci exhibit strong support for any hypothesis. Surprisingly, we found that a subset of loci that strongly support the monophyly of Primates, support treeshrews as sister to primates and flying lemurs. The overall small magnitude of differences in phylogenetic signal among the alternative hypotheses suggests that these three groups diversified nearly simultaneously. However, with our large dataset and approach to examining support, we provide evidence for the hypothesis of treeshrews as sister to rodents and lagomorphs, while demonstrating why support for alternate hypotheses has been seen in prior work. We also suggest that locus selection can unwittingly bias results.

Key words: phylogenomics, contentious relationship, incomplete lineage sorting, likelihood, site concordance factor, sisrs

In the past two decades, DNA sequence data has become the main source of information for reconstructing phylogenetic relationships. When only a few short gene regions could be used for tree inference, stochastic error was often large, resulting in some areas of phylogenies with poor resolution and low branch support. However, with the advent of High Throughput Sequencing, the size of phylogenetic datasets has increased dramatically, making it possible to overcome stochastic errors in small datasets and establish some evolutionary relationships with greater certainty (e.g., Dunn et al. 2008). At the same time, large datasets on their own might not be able to resolve all phylogenetic relationships due to systematic error (Rodríguez-Ezpeleta et al. 2007). Reanalyses of the same dataset with different tools, as well as examination of different datasets, have resulted in support for alternate hypotheses (Chakrabarty et al. 2017; Esselstyn et al. 2017; Reddy et al. 2017). Some relationships remain controversial despite extensive work (e.g. early-diverging animals [Jékely and Budd 2021], birds [Reddy et al. 2017; Braun et al. 2019], and placental mammals [Smith et al. 2015, Murphy et al. 2021]). Moreover, phylogenetic conflict may be obscured by the inability of traditional support metrics to capture variation in very large datasets, as both non-parametric bootstrap and posterior probability metrics tend to provide strong support for conflicting topologies (Kumar et al. 2012; Salichos and Rokas 2013).

Apparent systematic discordance may be caused by an array of factors. Biological processes such as incomplete lineage sorting (ILS) or hybridization are predicted to be most impactful for areas of the tree of life with rapid diversification (i.e., short internodes) (Rannala et al. 2020). Systematic issues in data processing (such as extensive contamination or homology inference errors) can bias the phylogenetic results of any dataset and are predicted to have a stronger impact on less informative datasets applied to controversial relationships (Simion et al. 2020). Addressing these issues can explain conflicts and resolve phylogenies, as seen in taxa

ranging from tunicates to turtles (Brown and Thomson 2017; Simion et al. 2020). Finally, a lack of information in the data may cause essentially arbitrary topological resolution depending on minor factors influencing particular analyses (e.g., Simion et al. 2020).

In order to disentangle these factors, the array of phylogenetic support methods has evolved to include additional approaches to elucidate the presence, strength, and possible cause of phylogenetic conflicts (Simon 2020). For example, the SH-aLRT test (Guindon et al. 2010) assesses likelihood differences between the maximum likelihood (ML) tree and the nearest neighbor interchange (NNI) alternatives for each node, making it a good generic test for informativeness of loci. To interrogate locus-specific phylogenetic support, likelihoods (Lee and Hugall 2003; Shen et al. 2017) or marginal likelihoods (Brown and Thomson 2017) can be compared across the gene or partition of interest. This approach provides a gauge of the variation in support among loci and has proven effective in determining the homogeneity of the conflict, and is especially useful at identifying overly impactful outliers (Brown and Thomson 2017; Springer and Gatesy 2018). The site concordance factor (sCF) (Minh et al. 2020a) provides a quartet-based support measure in a parsimony framework, and can be used both on the entire dataset and on a locus-by-locus basis. However, despite the availability of the aforementioned techniques, for many empirical datasets and contentious phylogenetic relationships the extent and potential causes of the conflicts remain elusive.

The position of the treeshrews (Scandentia) represents one of the glaring uncertainties in the mammal tree of life (Hallström and Janke 2010; Murphy et al. 2021). Treeshrews are part of the Euarchontoglires, a lineage that also includes rodents, lagomorphs, primates, and colugos (Dermoptera). Historically, morphological analyses supported grouping treeshrews with Dermoptera (O'Leary et al. 2013); however, genomic data now overwhelmingly supports the

sister-group relationship between Dermoptera and Primates (Meredith et al. 2011; Esselstyn et al. 2017). One of the currently prevailing hypotheses is that of treeshrews as the sister to the flying lemurs (Dermoptera) and Primates (Primatomorpha; Fig. 1) (Hallström and Janke 2008; McCormack et al. 2012; Zoonomia Consortium 2020). This placement has been recovered by numerous analyses, although typically with relatively low confidence. Alternatively, several phylogenomic analyses have supported the sister-group relationship of treeshrews with rodents and lagomorphs (Glires; Fig. 1) (Tarver et al. 2016; Liu et al. 2017; Du et al. 2019). Finally, one of the most recent molecular analyses (Fig. 1) found that Primatomorpha and Glires themselves form a monophyletic clade, with treeshrews equidistantly related to both (Esselstyn et al. 2017). Analyses recovering each of the topologies have varied in their sampling of species and higher order taxa, the number of sites analyzed, and methodology (Song et al. 2012; Esselstyn et al. 2017; Vachaspati and Warnow 2018), meaning that both biological and methodological factors may be at play when considering the conflicting taxonomic groupings. Thus, with copious genomic data across mammals and a limited set of hypotheses, the placement of treeshrews presents an ideal case study for phylogenetic conflict interrogation (Zoonomia Consortium 2020).

Here we examine the support for the alternative hypotheses of the placement of treeshrews on the mammal tree of life. Using millions of variable orthologous sites spread across mammal genomes, we assessed the support for each hypothesis in both parsimony and maximum likelihood frameworks. While the majority of loci were only weakly informative with respect to the phylogenetic placement of treeshrews, the overall dataset supports the hypothesis that treeshrews are most closely related to Glires. These results suggest that treeshrews, Glires, and Primatomorpha might have diversified in rapid succession, but our large dataset allows us to

distinguish the order of splitting. However, when examining loci that strongly favored primates as monophyletic, we found support for the Primatomorpha-sister hypothesis. This highlights concerns around the marker selection process even for very large datasets, as loci may be selected on their ability to reproduce an established phylogeny (Faircloth 2017). Our approach can be extended to other contentious taxa to examine support for alternative hypotheses on a locus-by-locus basis and gauge the amount of phylogenetic and non-phylogenetic signal, and its potential sources.

Material and Methods

Taxon Selection

In order to comprehensively interrogate the conflict in the phylogenetic placement of the treeshrews, we sampled taxa from this group and all major lineages surrounding its most likely position (Primates, Dermoptera, Rodentia, Lagomorpha), as well as five outgroup taxa from Laurasiatheria (Fig. 1). We hypothesized that sampling as many diverse representatives of each group as the publicly available data allows would help to more accurately detect saturating substitutions by breaking long branches and thus improve the signal-to-noise ratio (Baurain et al. 2007; Philippe et al. 2011). Thus we attempted to sample every family with publicly available short read whole genome sequencing data. A total of 50 taxa were included in the dataset (Fig. 1, Table S1).

Raw Reads

Raw reads were downloaded from the European Nucleotide Archive FTP server (accession numbers are listed in Table S1). The mean number of paired-end (PE) reads per sample was 955,859,108 (for details see the info in Table S1). Read quality was assessed with FastQC (Andrews 2010).

Obtaining Orthologous Loci

Orthologous loci were obtained through the SISRS v2.1 pipeline (Schwartz et al. 2015; Literman and Schwartz 2021). Briefly, reads were trimmed using BBDuk (Bushnell 2021) and subsampled using BBMap v38.81 suite scripts (Bushnell 2021) to 10X (0.2X each) assuming an average taxon genome size of 3.5 Gbps. Subsampled reads were assembled in Ray v2.3.1 (Boisvert et al. 2010) to produce a composite genome as the basis of identifying orthologs. QUAST v5.0.2 (Gurevich et al. 2013) was used to assess the basic properties of the assembly. The composite assembly was used as a reference for mapping all taxon reads with Bowtie 2 v2.3.5.1 (Langmead and Salzberg 2012; Langmead et al. 2019). We used bcftools v1.12 (Danecek et al. 2021) to call variants and produce a consensus sequence for each locus for each taxon. In order to avoid composite-assembly bases replacing missing data in the consensus sequence, sites with no coverage in a given taxon were first masked using BEDtools v2.30.0 (Quinlan and Hall 2010). The loci were filtered to those containing at least 50% of the taxa, at least one taxon from each of the five major lineages and the outgroup, and to have no more than two thirds of the sequence missing for each taxon. After filtering, the loci were aligned using MAFFT v7.475 (Katoh and Standley 2013) with the --auto flag.

Annotation

We hypothesized that locus type (e.g. whether sites fall within coding sequences, introns, etc.) may be associated with the phylogenetic signal for each of the alternative hypotheses, as a similar association has been shown for the overall strength of the signal (Literman and Schwartz 2021). Thus, we annotated the obtained SISRS loci using the human genome as a reference. Human sequences were extracted from the above generated SISRS loci and BLASTN 2.9.0+ (Altschul et al. 1990; Camacho et al. 2009) was used to align these sequences to the human

genome assembly GRCh38.p13 (Schneider et al. 2017). A custom script was used to filter the BLASTN output table and convert it into a BED file. Coordinates of the resulting alignments on the human genome chromosomes were intersected (Quinlan and Hall 2010) with the GFF file containing annotations. Another custom script was used to process the annotated BED file and produce a table with a proportion of each type of genomic feature for each locus. In this paper, we used the following feature types: CDS (coding sequence), UTR (untranslated region, defined as a combination of gene, mRNA, and exon, but no CDS), intron (defined as a combination of gene and mRNA, but no exon), lncRNA (long non-coding RNA), pseudogene, other (other features), and unannotated sequence. For the purposes of subsequent analyses, each SISRS locus was assigned to only one of the types of features based on which type covered the highest proportion of the locus.

Assessment of Locus Properties

We hypothesized that in addition to the annotation type, other properties of loci may be correlated with the support for particular phylogenetic hypotheses (Mongiardino Koch 2021). For each locus, we scored the number of taxa with data, locus length, proportion of missing data, the proportion of variable and parsimony informative sites, and GC content using AMAS (Borowiec 2016). The number of taxa and proportion of missing data quantify different levels of missing data, shown to be impactful when non-randomly distributed within phylogenetic datasets (Xi et al. 2016). The locus length and proportions of variable and parsimony informative sites are associated with the number of sites available for model estimates and for resolving relationships (Shen et al. 2016; Steenwyk et al. 2020). Quantifying locus GC content is helpful at determining skewed base composition which might be indicative of GC-biased gene conversion (Duret and Galtier 2009), shown to impact the phylogenetic signal (Romiguier et al. 2013). Good fit of a

substitution model allows more accurate discrimination between synapomorphic and convergent substitutions (Philippe et al. 2011), thus we assessed the absolute fit of a GTR+G model to each locus using PhyloMAd (Duchêne et al. 2017, 2018a, 2018b). The GTR+G model was chosen for all loci as it is one of the most generic substitution models and could be fit to a wide diversity of loci (Sumner et al. 2012; Abadi et al. 2019). As a metric of the adequacy of the model, we used the Mahalanobis distance (Mahalanobis 1936; Drummond and Suchard 2008), which serves as a summary statistic in PhyloMAd. Specifically, we recorded the number of standard deviations (SD) of the predictive distribution between the mean of the distribution and the test statistic for the empirical data. Since minimization of the substitution saturation levels was shown to improve the phylogenetic signal in the past (Philippe et al. 2011), we assessed saturation in our dataset. Following (Borowiec et al. 2015), we compared corrected and uncorrected pairwise distances in R 4.1.0 (R Core Team 2021) with the package ape 5.5 (Paradis and Schliep 2019). We used TN93 instead of gene tree-based GTR distances because gene trees were not computed due to SISRS loci generally being too short for that purpose. A linear model was fit to the data to estimate the slope of the regression line, with higher values corresponding to lower relative saturation. We performed a principal component analysis on the locus features using R with the ade4 v1.7-17 package (Thioulouse et al. 1997). The results of the analysis were used to assess the correlation between different locus properties and ensure the independence of the chosen metrics.

Phylogenetic Signal Assessment

We used two metrics as proxies for phylogenetic signals: a log-likelihood difference (Lee and Hugall 2003; Shen et al. 2017) between competing hypotheses (delta log-likelihood, $\Delta \ln L$), and a difference in sCF (Minh et al. 2020a) between competing hypotheses. In addition to

measuring the phylogenetic signal for the competing hypotheses of the treeshrew placement, we also assessed the signal for the monophyly of primates in order to compare the support for a contentious node with the support for a well-supported split of similar age.

Because assessment of both ΔlnL and sCF required predetermined topologies, we used the mammalian phylogenetic tree of (Armstrong et al. 2020), which was based on the taxa with genomic data available through the Zoonomia project (Zoonomia Consortium 2020) and included 49 of the 50 species we sampled. We pruned the tree to retain only the taxa that we sampled and, since *Tupaia belangeri* was not included in the original tree, we grafted this species as the sister to *Tupaia chinensis* based on (Fan et al. 2013). This topology is well-supported and uncontroversial with regard to the major lineages (Kumar et al. 2017). Because the resulting tree represented only one possible treeshrew placement (Scandentia+Primatomorpha), we obtained the two alternative placements using the phangorn package v2.7.0 (Schliep 2011) in R. These three topologies were used for subsequent evaluations (Fig. 1). To examine the phylogenetic signal for primates, we used the original tree (Fig. 1) with monophyletic primates and two trees with the alternative topologies around the primates branch (i.e. Haplorhini or Strepsirrhini sister to Dermoptera).

Each locus was fit to each phylogenetic hypothesis using IQ-TREE v2.1.2 (Minh et al. 2020b) and log-likelihoods were recorded. The approach of fitting loci individually was used to minimize the probability of a single underlying tree assumption violation, which is very likely in the concatenation framework (Rannala et al. 2020). We scaled locus log-likelihood by locus length to estimate a mean per site log-likelihood and to avoid bias associated with locus length (Walker et al. 2020). Because the loci were on average very short, we hypothesized that the estimates of model parameters may be too inaccurate compared to model estimates for longer

alignments (Xia 2020). Thus we also assessed the ΔlnL by fitting concatenated partitioned alignments to the alternative topologies, a more common approach for ΔlnL-based conflict interrogation (Shen et al. 2017). Because the total amount of sequence data exceeded the capabilities of the compute cluster system (total length over 73Mbp, see the results), we made the concatenated alignment computationally tractable: loci were arbitrarily binned into groups of 1000, and loci of each group were concatenated. Subsequent processing of these concatenated alignments was done as above for individual loci except that per site log-likelihoods were summed according to the original individual loci (Lee and Hugall 2003; Shen et al. 2017).

To assess the overall support across the entire dataset for each topology we summed log-likelihoods per locus to produce total log-likelihood in each case. On a locus by locus basis, the strength of the likelihood support is typically calculated between two competing hypotheses as the difference in log-likelihood (ΔlnL), with positive and negative values representing the support for one or another topology (Shen et al. 2017). Because we considered three competing hypotheses, per site log-likelihood values for each topology were standardized to be at 0 for a hypothesis with the lowest likelihood. Positive values for a given hypothesis thus indicate the ΔlnL-based support for it compared to the least likely hypothesis. In addition to assessing the overall distribution of ΔlnL, we focused on the loci with larger differences in fit between the alternative topologies (ΔlnL > 0.03 - 0.1 depending on the dataset, equating to ~1000-2000 outlier loci in each case) and by extension with a stronger phylogenetic signal.

We then classified loci into three groups, based on topology, based on which loci had the highest log-likelihood. Loci having two best resolutions or equal likelihood between the alternative resolutions were excluded from this assessment. Groups of loci favoring one of the competing hypotheses were used to obtain outlier loci with a strong signal for a particular

hypothesis (ΔlnL > 0.03 - 0.1 as above), as well as to examine properties of loci belonging to each group. We further analyzed the loci strongly supporting and rejecting the primate monophyly regarding their signal for each of the alternative treeshrew hypotheses. We calculated the mean ΔlnL in the two groups of loci when fit to each of the treeshrew topologies and compared it with the mean ΔlnL of a random sample of loci from the dataset.

We used IQ-TREE v2.1.2 to calculate sCF (Minh et al. 2020a). For the Scandentia+Primatomorpha and Scandentia+Glires hypotheses we recorded the sCF of the branch subtending the corresponding clades, while for the Scandentia+(Primatomorpha+Glires) hypothesis we recorded the sCF for the branch subtending the Primatomorpha+Glires clade. In addition, we assessed the correlation between sCF and ΔlnL on a locus-by-locus basis to assess a connection between parsimony-based and model-based phylogenetic signals.

We assessed the significance of the difference in total log-likelihood across the dataset for all competing hypotheses using the approximately unbiased (AU) test (Shimodaira 2002) implemented in the R package scaleboot (Shimodaira 2008). Significance of the differences in loci numbers favoring each topological resolution was assessed using Pearson's chi-squared test (Pearson 1900; Agresti 2018), while repeated-measures ANOVA (Girden 1992) followed by pairwise paired t-tests with Bonferroni correction were used to compare the mean per locus ΔlnL and sCF between the competing topologies.

Interrelationship Between Phylogenetic Signal and Loci Properties

We tested for a correlation between locus property values and the phylogenetic signal towards particular topological resolution of the treeshrews. For this question, we analyzed the distribution of ΔlnL in groups of loci favoring each topology, as well as the distribution of previously assessed locus properties in each group. We used ANOVA followed by pairwise t-

tests with Bonferroni correction to compare the means of the properties of the loci favoring competing hypotheses, except for the locus type proportions, differences in which were assessed using Pearson's chi-squared test. All scripts for data analysis are available at: https://github.com/AlexKnyshov/TreeshrewProject

Results

Orthologous Loci Generation

On average, 573,949,175 reads per sample (min 90,024,816, max 1,868,772,590, median 576,806,927) passed the trimming and were used for the SISRS pipeline. A total of 11,997,130 SISRS contigs containing 2,208,378,641bp were identified. The largest contig spanned 11,534bp, N50 was 189, L50 was 4,693,728, GC content was 40.23%. After filtering using several missing data criteria (see methods), a total of 386,935 loci containing 73,014,086bp were retained for subsequent analyses.

Properties of Selected Loci

Results of the evaluations are listed in Table S2 and in Figure S1. The number of taxa varied from 25 (minimum passing filter) to 50, with a median of 39. The average locus length was 189bp. Proportions of variable and parsimony informative sites as well as percent missing were nearly normally distributed with means of 0.53, 0.29, and 26% respectively. Mean GC content was 41%, although significant differences in GC content were observed between some locus types (Fig. S2). On average, the estimated saturation of the SISRS loci was low, with the mean slope of saturation curve 0.8. The number of SDs between the mean of the distribution and the test statistic for the empirical data formed a bimodal distribution, with a smaller peak around 11 SD and a much larger peak around 6 SD. The results of PCA among different locus properties are shown in Figure S3. Among the loci successfully mapped to the reference and passed the

filtering, types of genomic features were annotated in the following proportions: CDS 17.06%, intron 34.13%, lncRNA 11.62%, pseudogene 1.48%, UTR 5.74%, other 0.53%, and unannotated 29.43%.

Phylogenetic Signal in the Treeshrew case

Summed log-likelihood of all loci individually fit to each treeshrew placement topology was as follows: -554,389,692 for the Primatomorpha sister, -554,284,571 for the Glires sister, and -554,358,628 for the Primatomorpha+Glires sister. Summed log-likelihood of loci fit to each topology in concatenated alignments was -562,408,325 for Primatomorpha sister, -562,381,959 for Glires sister, and -562,408,865 for Primatomorpha+Glires sister. Thus in both cases the Glires sister hypothesis was the most likely one (AU test p-value = 0). Most loci demonstrated only minor differences in the likelihood of the fit to the alternative treeshrew placements, both when fit individually (mean per site ΔlnL = 0.006, Figs 2a, S4a) and when fit in concatenation (mean per site ΔlnL = 0.003, Figs 2b, S4b). The variance of ΔlnL, however, was considerably smaller when loci were fit in concatenated alignments compared to individually (SD ΔlnL = 0.004 and 0.013 respectively). The numbers of loci favoring each treeshrew placement were relatively similar (although significantly different, Figs 3a, 3b) both when fit individually and when fit in concatenation. The numbers of strong outliers (per site ΔlnL > 0.1 for individually fit loci or per-site ΔlnL > 0.03 for concatenated loci) varied more drastically (Figs 3a, 3b), with many more loci favoring the Glires sister hypothesis both when fit individually (Fig. 3a, n=162 for the Primatomorpha sister, n=595 for the Glires sister, and n=238 for the Primatomorpha+Glires sister) and when fit in concatenation (Fig. 3b, n=136 for the Primatomorpha sister, n=709 for the Glires sister, and n=49 for the Primatomorpha+Glires sister).

Differences in sCF of alternative treeshrew topologies were also minor (sCF=26.52±11.98 for the Primatomorpha sister, sCF=28.0±13.05 for the Glires sister, and sCF=26.47±12.56 for the Primatomorpha+Glires sister) (Figs 3c, 3d), although the Glires sister had a significantly higher average sCF (repeated measures ANOVA, F-statistic 1633.604, p-value = 0). The relationship between ΔlnL and sCF for particular loci was positive but very weak (Fig. S6) both in the per locus fit approach (R-squared 0.04-0.09, p-value < 2.2e-16) and in the concatenation approach (R-squared 0.1-0.15, p-value < 2.2e-16). The same relationship was weak and negative for the outlier loci independent of the ΔlnL assessment approach (R-squared 0.004 - 0.07, for p-values see Table S3).

Phylogenetic signal in the Primates case

Summed log-likelihood of all loci fit individually was -554,395,578 for the control case of monophyletic primates, compared to -554,680,853 and -554,663,509 for the two alternative NNIs. Summed log-likelihood of all loci in concatenation was -567,221,958 for the monophyletic primates, compared to -567,473,217 and -567,470,884 for the two alternative NNIs. Thus both methods of assessing the likelihood difference converged on the monophyletic primates as the most likely hypothesis (AU p-value = 0). The mean differences in fit of loci to the three primate alternatives were higher than those in the treeshrew evaluations, both when fit individually (mean per site ΔlnL = 0.009, Figs 4a, S5a) and when fit in concatenation (mean per site ΔlnL = 0.007, Figs 4b, S5b). The proportion of loci favoring the monophyletic primates was much higher than the number of loci favoring the alternatives both when fit individually (51%, Fig. 5a) and when fit in concatenation (54%, Fig. 5b). The number of strong outliers favoring monophyletic primates in the individual fit approach (per site ΔlnL > 0.1) was similar to the

overall loci (55%, Fig. 5a). However, in the concatenation approach, the overwhelming majority (98%) of outliers (per site ΔlnL > 0.05) favored the monophyly of primates (Fig. 5b).

Differences in sCF between the monophyletic primates and two alternative NNIs were higher than between any treeshrew placement topologies (sCF=32.9±18.68 for the monophyletic primates, sCF=24.96±14.43 and sCF=23.66±14.26 for the two alternative NNIs) (Figs 5c, 5d). There was a strong positive relationship between ΔlnL and sCF across all loci for the case of monophyletic primates (standardized slope 0.47-0.58, R-squared 0.22-0.33, p-value < 2.2e-16). The same relationship for the two alternative NNIs across all loci, or for all topologies across the strong outliers was weak (R-squared 0.0001-0.09, for p-values see Table S3).

Support for Treeshrew Relationships Using Loci Supporting Monophyletic Primates

The loci that exhibited strong support for the monophyly of primates when fit individually (ΔlnL > 0.1, N=698) had high support for the Primatomorpha-sister hypothesis of the treeshrew placement in the individual-fit approach (mean ΔlnL = 0.0048). The same loci had lower mean ΔlnL = 0.0024 and 0.0027 when fit to the Glires-sister and Primatomorpha+Glires sister topologies. Loci that strongly rejected the monophyly of primates (ΔlnL > 0.1, N=583) had higher support for the Glires-sister (mean ΔlnL = 0.0118) and the Primatomorpha+Glires sister (mean ΔlnL = 0.0106) hypotheses compared to the Primatomorpha-sister (mean ΔlnL = 0.0042). As expected, a randomly selected set of loci (N=698) reflected the signal of the entire dataset and had significantly higher support for the Glires-sister (mean ΔlnL = 0.0034) compared to the alternatives (mean ΔlnL = 0.0017 and 0.0018).

However, when we examined the fit of the same loci (698 loci strongly supporting Primate monophyly and 583 loci strongly rejecting it) to the treeshrew topologies in the concatenation framework, the results were drastically different. All three groups of loci (favoring

Primate monophyly, rejecting it, as well as a random set of loci) had the highest fit to the Glires-sister topology (mean ΔlnL = 0.0013-0.0022) compared to the alternatives (mean ΔlnL = 0.0009-0.0013).

Interrelationship Between Phylogenetic Signal and Loci Properties

Loci favoring each of the treeshrew placements had similar means for values of the properties we assessed, although most of the differences were significant (see Fig. S7, Table S3). When we looked at the loci strongly favoring each of the competing topologies in the individual fit approach (ΔlnL > 0.1), most comparisons yielded no significant differences (Fig. 6, Table S3). The exceptions were differences in taxon composition, GC content, and saturation slope. In particular, the fraction of Primatomorpha and Glires were inversely related to the ΔlnL of the loci favoring the Primatomorpha-sister and the Glires-sister hypotheses respectively (p-value = 1.38E-19).

Discussion

Overall, the SISRS loci contained strong phylogenetic signal, as measured on an uncontroversial split (Primates). Both metrics used as proxies for phylogenetic signal - ΔlnL and sCF - showed that the overall dataset clearly favored the topology with monophyletic primates over the two alternative NNIs. While the overall modest per-site ΔlnL-based signal suggests that the majority of genomic markers contain only weak phylogenetic signal for deep splits, and some contain non-phylogenetic signal, a plurality of data supports for the hypothesis of Glires as sister to treeshrews.

Phylogenetic Signal and the Competing Treeshrew Hypotheses

Overall, there was greater support for the Glires-sister hypothesis over the alternative hypotheses in our data. Prevalence of the Glires-sister signal was evident both in the parsimony

(sCF-based) and the model (ΔlnL-based) frameworks. In the latter, the Glires-sister hypothesis had the highest support in both types of assessment (individually-fit and concatenated loci) and was detectable in the entire dataset as well as in the strong ΔlnL-based outliers. The magnitude of the signal difference in favor of the Glires-sister was considerably smaller than the magnitude of the signal difference in the case of primate monophyly, but this is consistent with a notion of rapid radiation between the major Euarchontoglires lineages (Zhou et al. 2015; Esselstyn et al. 2017). The magnitude of support for monophyletic primates was higher both in the sCF-based and ΔlnL-based assessments, however the support was never maximal, thus showing detectable conflict even in the uncontroversial split. Similar results were shown in Esselstyn et al. (2017) when examining the likelihood support for a well-established monophyly of Glires. We thus argue that smaller differences in support of alternative topological placements of treeshrews does not necessarily invalidate a hypothesis of a bifurcating species tree with the Glires-sister placement of Scandentia, since to some extent the conflict might be pervasive throughout the mammal phylogeny and evident even in well-supported nodes.

Interestingly, the loci that strongly supported the monophyly of primates favored the Primatomorpha-sister hypothesis, which is contrary to the overall support of our dataset. However, analysis of concatenated loci that strongly favored primate monophyly showed an overall preference for the Glires-sister hypothesis, which is consistent with the complete dataset. We hypothesize that the results from individual loci may stem from the challenges of accurately fitting models to short low-information loci, and discuss this further below.

This conflict among subsets of loci in our datasets suggests caution when considering how to filter phylogenomic datasets based on apparent phylogenetic signal. For many phylogenomic subsampling methods, loci are selected based on alignments between a handful of

reference genomes (Faircloth et al. 2012; Lemmon et al. 2012), and are typically checked to reproduce a well-supported phylogeny as a control measure in the design phase (Faircloth 2017). Our results show that such (non-random) filtering could bias results away from the signal across the genome. However, while the phylogenetic signal shows statistically significant support for the Glires-sister over the alternate hypotheses, we also recognize that the statistical results here might reflect the enormous size of the dataset and do not preclude the presence of systematic errors biasing our results (Kumar et al. 2012). Thus, an alternate interpretation may be that the loci supporting the monophyly of primates hold more accurate signal compared to the larger dataset. We discuss these alternate possibilities below.

Concatenation and Impacts on Phylogenetic Signal

The modest signal of individual loci was ameliorated when loci were concatenated, leading to nearly all loci strongly supporting primates as monophyletic. We hypothesize that poor model fit for short single-locus alignments may be one reason that support from concatenated loci was so much stronger than for loci fit individually (Xia 2020). The SISRS loci in our dataset had a mean length of about 200bp, which for conserved loci is likely insufficient to adequately estimate substitution model parameters. Thus the parameter estimates were likely to have random noise, which might increase the likelihood for the alternative topologies and decrease the likelihood for the true tree. The use of short loci has been criticized previously on the grounds of poor model fit and resulting poor gene tree estimates (Bryant and Hahn 2020; Rannala et al. 2020). While normally model choice and fit does not impact the resulting topology as much as it impacts branch length estimates, model accuracy could be critical for nodes that are difficult to resolve (Kelchner and Thomas 2007). This is well illustrated in the case of the order of divergence of Porifera and Ctenophora, where using different types of models results in

reconstruction of two alternative topological hypotheses (Li et al. 2021). In our study, error in model parameter estimation thus could explain why the loci supporting primate monophyly led to support for Primatomorpha-sister when analyzed individually, while these same loci supported Glires-sister when concatenated.

However, it is also possible that model parameters may be estimated incorrectly due to concatenating loci. This case would lead us to believe the analysis of individual loci in the primate-monophyly dataset supporting Primatomorpha-sister over the concatenated results (although keeping in mind that for the full dataset both individual and concatenated analyses supported Glires-sister). Model parameter estimation for concatenated datasets is commonly done with the help of the tree reconstructed on the entire matrix (in the case of IQ-TREE, it is a parsimony-based tree [Minh et al. 2020b]). Such a procedure would violate the assumption of multispecies coalescence (MSC) and potentially bias individual loci to support the predominant topology, which they would otherwise reject (Walker et al. 2020). Longer concatenated loci have been noted as violating the assumptions of MSC and loci across our large dataset are practically guaranteed to violate the assumption of a single generating tree (Rannala et al. 2020). Recent work has noted discordance between reconstructed gene trees for individual loci and the tree supported by the same loci when part of a larger supermatrix (Walker et al. 2020). Thus, signal supporting alternate hypotheses may be lost in concatenation.

Overall, it is worth keeping in mind both the issue of model fit in short loci leading to erroneous support, as well as the issues of both model fit and lost signal in concatenated datasets. We conclude that both approaches should be used when interrogating phylogenetic conflicts.

Loci Properties and Signal for Treeshrew Placement

It has been shown that locus type (Literman and Schwartz 2021) and locus properties (Mongiardino Koch 2021) are related to the overall amount of phylogenetic signal. However, we did not observe relationships between locus properties and a particular conflicting topology. While minor but significant differences of some properties could be predictive of the amount of signal and noise in the data, more sensitive methods would be needed to tease this out. It is also possible that the signal for each of the topologies is randomly distributed throughout the genomes independent of the locus properties, and any difference found between groups of loci favoring competing hypotheses would bear no implication across different datasets and phylogenetic problems.

Signal Metrics Discordance on a Locus-by-locus Basis

Despite the congruence between mean ΔlnL and mean sCF in the entire dataset, there was no strong correlation between the two metrics on a locus-by-locus basis. This is not surprising given that sCF is computed in a parsimony framework and is not impacted by model parameter misspecification. On the other hand, sCF is a quartet-based metric and can be stochastic when the conservative nature of SISRS loci results in a small number of informative quartets (Minh et al. 2020a). Interestingly, there was a relatively strong positive correlation between ΔlnL and sCF for the loci supporting the monophyletic primates. We hypothesize that these loci, even despite their short size, have enough signal to overcome stochasticity in model fit and/or quartet sampling, and thus bring the two metrics to congruence. Consequently, that would mean that in the cases with a strong ΔlnL-sCF signal discordance loci might have elevated levels of noise compared to the phylogenetic signal. We hypothesize that model fit accuracy might explain some of the differences between the two types of ΔlnL-based signal and the sCF-based signal assessments.

Taxon Sampling and Signal for Treeshrew Placement

Taxon sampling has been shown previously to affect phylogenetic estimates (Rosenberg and Kumar 2003; Nabhan and Sarkar 2012). In this study, Dermoptera, Lagomorpha, and treeshrews were represented by only a few species each and were constrained to have at least one species present in each locus. Thus the sampling of the diverse Primates and rodents primarily determined the overall variance in taxon sampling, as well as the fractions of Primatomorpha and Glires. While our taxon set was relatively skewed towards a high fraction of Glires (0.55), similar fractions of Glires were included in prior analyses supporting all alternative hypotheses of treeshrew placement. In the datasets recovering the Glires sister, the fraction of Glires ranged between 0.39 and 0.44 (Song et al. 2012; Romiguier et al. 2013; Tarver et al. 2016; Liu et al. 2017), while in the datasets finding the Primatomorpha sister this fraction ranged between 0.31 and 0.5 (Hallström and Janke 2008; McCormack et al. 2012; Kumar et al. 2013; Esselstyn et al. 2017). Thus, although the high Glires sampling might be seen as a concern, our results are consistent with results from lower Glires sampled datasets. However, it is possible that more intricate properties of taxon sampling (such as presence or absence of particular species) are at play here.

SISRS Data and Conflict Interrogation

Our results also contain several practical outcomes with regard to SISRS-based data. The recovered SISRS loci in our case are relatively conserved, which is unsurprising given the methodology of the pipeline. However, the proportion of parsimony informative sites in our dataset was in the same range as in some of the mammal UCE studies (McCormack et al. 2012; Esselstyn et al. 2017), although SISRS alignments were considerably shorter (~200bp vs ~700bp [Esselstyn et al. 2017]). Given that SISRS loci contain considerable proportions of both coding

and regulatory elements, as well as a large amount of normally not used unannotated (intergenic) loci (Literman and Schwartz 2021), SISRS datasets offer a good diversity of locus types. In contrast, most other methods focus on subsets of particular locus types (i.e., UCE, protein-coding sequences, introns) or small subsets of randomly targeted markers (i.e., RADSeq methods), which have the potential to bias results (Literman and Schwartz 2021).

Conclusion

Overall, the SISRS loci used here provided a wealth of phylogenetic information, enabling us to observe support for alternate hypotheses for the placement of treeshrews. The Glires-sister is the most likely hypothesis when considering the entire dataset. While the magnitude of the difference with respect to the alternatives is small, we are able to support this hypothesis due to our large dataset. While loci that strongly support our “control tree” (primates) have a higher likelihood for the Primatomorpha-sister hypothesis, we believe this is likely due to poor model fit, especially as these data supported Glires-sister when concatenated. Conflicting phylogenetic signals between subsets of loci thus suggest filtering loci (or selecting particular loci to sequence) should be done with caution due to the potential to estimate a phylogeny that is not concordant with an estimate based on a more complete dataset.

Our work also highlights the importance of per-locus phylogenetic signal interrogation (Smith et al. 2015; Brown and Thomson 2017; Shen et al. 2017) as a valuable tool to examine the conflicting phylogenetic relationships. The methodology for conflict interrogation utilized in this study allowed us to ascertain the overall magnitude of the conflict and the amount of signal supporting each conflicting topology in a locus-by-locus framework. The techniques allowed us to conclude that the support for the Glires-sister hypothesis is uniformly distributed throughout the dataset, while the support for the Primatomorpha-sister is restricted to a small subset of loci.

Using a locus-by-locus interrogation we were also able to find a discordance between the model-based and parsimony-based methods which we in part attributed to the adequacy of model fit to the data.

Acknowledgments

This research was funded by a grant to R. Schwartz from the National Science Foundation (DBI-1942273). We thank members of the Schwartz lab for valuable comments and suggestions. The High Performance Research Computing facility at URI is acknowledged for providing the computational resources for the analyses in this manuscript.

References

Abadi S., Azouri D., Pupko T., Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. Nat. Commun. 10:934.

Agresti A. 2018. An Introduction to Categorical Data Analysis. John Wiley & Sons.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Andrews S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Armstrong J., Hickey G., Diekhans M., Fiddes I.T., Novak A.M., Deran A., Fang Q., Xie D., Feng S., Stiller J., Genereux D., Johnson J., Marinescu V.D., Alföldi J., Harris R.S., Lindblad-Toh K., Haussler D., Karlsson E., Jarvis E.D., Zhang G., Paten B. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 587:246–251.

Baurain D., Brinkmann H., Philippe H. 2007. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? Mol. Biol. Evol. 24:6–9.

Boisvert S., Laviolette F., Corbeil J. 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J. Comput. Biol. 17:1519–1533.

Borowiec M.L. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ. 4:e1660.

Borowiec M.L., Lee E.K., Chiu J.C., Plachetzki D.C. 2015. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics. 16:987.

Braun E.L., Cracraft J., Houde P. 2019. Resolving the Avian Tree of Life from Top to Bottom: The Promise and Potential Boundaries of the Phylogenomic Era. In: Kraus R.H.S., editor. Avian Genomics in Ecology and Evolution: From the Lab into the Wild. Cham: Springer International Publishing. p. 151–210.

Brown J.M., Thomson R.C. 2017. Bayes Factors Unmask Highly Variable Information Content, Bias, and Extreme Influence in Phylogenomic Analyses. Syst. Biol. 66:517–530.

Bryant D., Hahn M.W. 2020. The Concatenation Question. In: Scornavacca, C., Delsuc, F., and Galtier, N., editor. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 3.4:1–3.4:23.

Bushnell B. 2021. BBTools. BBMap https://sourceforge. net/projects/bbmap/.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421.

Chakrabarty P., Faircloth B.C., Alda F., Ludt W.B., Mcmahan C.D., Near T.J., Dornburg A., Albert J.S., Arroyave J., Stiassny M.L.J., Sorenson L., Alfaro M.E. 2017. Phylogenomic Systematics of Ostariophysan Fishes: Ultraconserved Elements Support the Surprising Non-Monophyly of Characiformes. Syst. Biol. 66:881–895.

Danecek P., Bonfield J.K., Liddle J., Marshall J., Ohan V., Pollard M.O., Whitwham A., Keane T., McCarthy S.A., Davies R.M., Li H. 2021. Twelve years of SAMtools and BCFtools. Gigascience. 10.

Drummond A.J., Suchard M.A. 2008. Fully Bayesian tests of neutrality using genealogical summary statistics. BMC Genet. 9:68.

Duchêne D.A., Duchêne S., Ho S.Y.W. 2017. New Statistical Criteria Detect Phylogenetic Bias Caused by Compositional Heterogeneity. Mol. Biol. Evol. 34:1529–1534.

Duchêne D.A., Duchêne S., Ho S.Y.W. 2018a. PhyloMAd: efficient assessment of phylogenomic model adequacy. Bioinformatics. 34:2300–2301.

Duchêne D.A., Duchêne S., Ho S.Y.W. 2018b. Differences in Performance among Test Statistics for Assessing Phylogenomic Model Adequacy. Genome Biol. Evol. 10:1375–1388.

Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M.V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. 2008. Broad

phylogenomic sampling improves resolution of the animal tree of life. Nature. 452:745–749.

Duret L., Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. 10:285–311.

Du Y., Wu S., Edwards S.V., Liu L. 2019. The effect of alignment uncertainty, substitution models and priors in building and dating the mammal tree of life. BMC Evol. Biol. 19:203.

Esselstyn J.A., Oliveros C.H., Swanson M.T., Faircloth B.C. 2017. Investigating Difficult Nodes in the Placental Mammal Tree with Expanded Taxon Sampling and Thousands of Ultraconserved Elements. Genome Biol. Evol. 9:2308–2321.

Faircloth B.C. 2017. Identifying conserved genomic elements and designing universal bait sets to enrich them. Methods Ecol. Evol. 8:1103–1112.

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61:717–726.

Fan Y., Huang Z.-Y., Cao C.-C., Chen C.-S., Chen Y.-X., Fan D.-D., He J., Hou H.-L., Hu L., Hu X.-T., Jiang X.-T., Lai R., Lang Y.-S., Liang B., Liao S.-G., Mu D., Ma Y.-Y., Niu Y.-Y., Sun X.-Q., Xia J.-Q., Xiao J., Xiong Z.-Q., Xu L., Yang L., Zhang Y., Zhao W., Zhao X.-D., Zheng Y.-T., Zhou J.-M., Zhu Y.-B., Zhang G.-J., Wang J., Yao Y.-G. 2013. Genome of the Chinese tree shrew. Nat. Commun. 4:1426.

Girden E.R. 1992. ANOVA: Repeated Measures. SAGE.

Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Gurevich A., Saveliev V., Vyahhi N., Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29:1072–1075.

Hallström B.M., Janke A. 2008. Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations. BMC Evol. Biol. 8:162.

Hallström B.M., Janke A. 2010. Mammalian evolution may not be strictly bifurcating. Mol. Biol. Evol. 27:2804–2816.

Jékely G., Budd G.E. 2021. Animal Phylogeny: Resolving the Slugfest of Ctenophores, Sponges and Acoels? Curr. Biol. 31:R202–R204.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kelchner S.A., Thomas M.A. 2007. Model use in phylogenetics: nine key questions. Trends Ecol. Evol. 22:87–94.

Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky Pond S.L., Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457–472.

Kumar S., Stecher G., Suleski M., Hedges S.B. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. 34:1812–1819.

Kumar V., Hallström B.M., Janke A. 2013. Coalescent-based genome analyses resolve the early branches of the euarchontoglires. PLoS One. 8:e60019.

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9:357–359.

Langmead B., Wilks C., Antonescu V., Charles R. 2019. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics. 35:421–432.

Lee M.S.Y., Hugall A.F. 2003. Partitioned likelihood support and the evaluation of data set conflict. Syst. Biol. 52:15–22.

Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61:727–744.

Literman R., Schwartz R. 2021. Genome-scale profiling reveals noncoding loci carry higher proportions of concordant data. Mol. Biol. Evol. 38:2306–2318.

Liu L., Zhang J., Rheindt F.E., Lei F., Qu Y., Wang Y., Zhang Y., Sullivan C., Nie W., Wang J., Yang F., Chen J., Edwards S.V., Meng J., Wu S. 2017. Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. Proc. Natl. Acad. Sci. U. S. A. 114:E7282–E7290.

Li Y., Shen X.-X., Evans B., Dunn C.W., Rokas A. 2021. Rooting the Animal Tree of Life. Evolution. 38:4322–4333.

Mahalanobis P.C. 1936. On the generalized distance in statistics. Proceedings of the National Institute of Science of India, 2:49–55.

McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22:746–754.

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

Minh B.Q., Hahn M.W., Lanfear R. 2020a. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. Mol. Biol. Evol. 37:2727–2733.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. 2020b. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37:1530–1534.

Mongiardino Koch N. 2021. Phylogenomic subsampling and the search for phylogenetically reliable loci. Mol. Biol. Evol. 38:4025–4038.

Murphy W.J., Foley N.M., Bredemeyer K.R., Gatesy J., Springer M.S. 2021. Phylogenomics and the Genetic Architecture of the Placental Mammal Radiation. Annu Rev Anim Biosci. 9:29–53.

Nabhan A.R., Sarkar I.N. 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Brief. Bioinform. 13:122–134.

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

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.

Pearson K. 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Lond. Edinb. Dublin Philos. Mag. J. Sci. 50:157–175.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Quinlan A.R., Hall I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842.

Rannala B., Edwards S.V., Leaché A., Yang Z. 2020. The Multi-species Coalescent Model and Species Tree Inference. In: Scornavacca, C., Delsuc, F., and Galtier, N., editor. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 3.3:1–3.3:21.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. .

Reddy S., Kimball R.T., Pandey A., Hosner P.A., Braun M.J., Hackett S.J., Han K.-L., Harshman J., Huddleston C.J., Kingston S., Marks B.D., Miglia K.J., Moore W.S., Sheldon F.H., Witt C.C., Yuri T., Braun E.L. 2017. Why Do Phylogenomic Data Sets Yield Conflicting Trees? Data Type Influences the Avian Tree of Life more than Taxon Sampling. Syst. Biol. 66:857–879.

Rodríguez-Ezpeleta N., Brinkmann H., Roure B., Lartillot N., Lang B.F., Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56:389–399.

Romiguier J., Ranwez V., Delsuc F., Galtier N., Douzery E.J.P. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol. Biol. Evol. 30:2134–2144.

Rosenberg M.S., Kumar S. 2003. Taxon sampling, bioinformatics, and phylogenomics. Syst. Biol. 52:119–124.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497:327–331.

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

Schneider V.A., Graves-Lindsay T., Howe K., Bouk N., Chen H.-C., Kitts P.A., Murphy T.D., Pruitt K.D., Thibaud-Nissen F., Albracht D., Fulton R.S., Kremitzki M., Magrini V., Markovic C., McGrath S., Steinberg K.M., Auger K., Chow W., Collins J., Harden G., Hubbard T., Pelan S., Simpson J.T., Threadgold G., Torrance J., Wood J.M.,

Clarke L., Koren S., Boitano M., Peluso P., Li H., Chin C.-S., Phillippy A.M., Durbin R., Wilson R.K., Flicek P., Eichler E.E., Church D.M. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27:849–864.

Schwartz R.S., Harkins K.M., Stone A.C., Cartwright R.A. 2015. A composite genome approach to identify phylogenetically informative data from next-generation sequencing. BMC Bioinformatics. 16:193.

Shen X.-X., Hittinger C.T., Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat Ecol Evol. 1:126.

Shen X.-X., Salichos L., Rokas A. 2016. A Genome-Scale Investigation of How Sequence, Function, and Tree-Based Gene Properties Influence Phylogenetic Inference. Genome Biol. Evol. 8:2565–2580.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.

Shimodaira H. 2008. Testing regions with nonsmooth boundaries via multiscale bootstrap. J. Stat. Plan. Inference. 138:1227–1241.

Simion P., Delsuc F., Philippe H. 2020. To What Extent Current Limits of Phylogenomics Can Be Overcome? In: Scornavacca, C., Delsuc, F., and Galtier, N., editor. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 2.1:1–2.1:34.

Simon C. 2020. An Evolving View of Phylogenetic Support. Syst. Biol. 0:1–8.

Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol. 15:150.

Song S., Liu L., Edwards S.V., Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proceedings of the National Academy of Sciences. 109:14942–14947.

Springer M.S., Gatesy J. 2018. Pinniped Diphyly and Bat Triphyly: More Homology Errors Drive Conflicts in the Mammalian Tree. J. Hered. 109:297–307.

Steenwyk J.L., Buida T.J. 3rd, Li Y., Shen X.-X., Rokas A. 2020. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol. 18:e3001007.

Sumner J.G., Jarvis P.D., Fernández-Sánchez J., Kaine B.T., Woodhams M.D., Holland B.R. 2012. Is the general time-reversible model bad for molecular phylogenetics? Syst. Biol. 61:1069–1074.

Tarver J.E., Dos Reis M., Mirarab S., Moran R.J., Parker S., O'Reilly J.E., King B.L., O'Connell M.J., Asher R.J., Warnow T., Peterson K.J., Donoghue P.C.J., Pisani D. 2016. The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference. Genome Biol. Evol. 8:330–344.

Thioulouse J., Chessel D., Doledec S., Olivier J.-M. 1997. ADE-4: a multivariate analysis and graphical display software. Stat. Comput. 7:75–83.

Vachaspati P., Warnow T. 2018. SVDquest: Improving SVDquartets species tree

estimation using exact optimization within a constrained search space. Mol. Phylogenet. Evol. 124:122–136.

Walker J.F., Shen X.-X., Rokas A., Smith S.A., Moyroud E. 2020. Disentangling biological and analytical factors that give rise to outlier genes in phylogenomic matrices. bioRxiv.:2020.04.20.049999.

Xia X. 2020. A Mathematical Primer of Molecular Phylogenetics. CRC Press.

Xi Z., Liu L., Davis C.C. 2016. The Impact of Missing Data on Species Tree Estimation. Mol. Biol. Evol. 33:838–860.

Zhou X., Sun F., Xu S., Yang G., Li M. 2015. The position of tree shrews in the mammalian tree: Comparing multi-gene analyses with phylogenomic results leaves monophyly of Euarchonta doubtful. Integr. Zool. 10:186–198.

Zoonomia Consortium. 2020. A comparative genomics multitool for scientific discovery and conservation. Nature. 587:240–245.

Figure legends

Fig. 1. The taxon set and the phylogeny used for evaluations. Different resolutions of node 1 are shown on the right. Node 2 indicates the most recent common ancestor of Primates, this node was used as a control node and was compared to the two alternative nearest neighbor interchanges around the subtending branch.

Fig. 2. Likelihood support for the three competing hypotheses of treeshrew placement, with loci fit individually (a) or in concatenation (b) suggest that treeshrews (Scandentia) are sister to Glires. Each plot shows the difference in likelihood (ΔlnL), obtained via fitting each locus to the three competing hypotheses, and rescaled by the locus length. If a point is at the center or close to it, there is no or little difference in likelihood between the topologies for a given locus thus low-likelihood data was omitted. Points plotted further away from the center correspond to loci having higher likelihood under one or two topologies. Each locus point is color coded by density value in that area of the plot.

Fig. 3. Counts of loci favoring one of the competing hypotheses of treeshrew placement based on the phylogenetic signal metrics demonstrate slight prevalence of loci favoring the Glires sister hypothesis. a) The numbers of loci favoring one topology over other two based on ΔlnL in the overall dataset and among the strong outliers, when assessed using individually fit loci. b) The numbers of loci favoring one topology over other two based on ΔlnL in the overall dataset and among the strong outliers, when assessed in concatenation. c) The overall distribution of site concordance factors (sCF), black horizontal bar representing the mean, dashed line shows the majority threshold of sCF (Minh et al 2020). f) The numbers of loci favoring one topology over the other two based on sCF.

Fig. 4. Likelihood support for monophyly of Primates shows weak support with loci fit individually (a) and strong support when loci are fit in concatenation (b). Each plot shows the difference in likelihood (ΔlnL), obtained via fitting each locus to the three competing hypotheses, and rescaled by the locus length. If a point is at the center or close to it, there is no or little difference in likelihood between the topologies for a given locus thus low-likelihood data was omitted. Points plotted further away from the center correspond to loci having higher likelihood under one or two topologies. Each locus point is color coded by density value in that area of the plot.

Fig. 5. Counts of loci favoring one of the competing hypotheses around the Primates branch based on the phylogenetic signal metrics demonstrate strong prevalence of loci favoring Primates monophyly. a) The numbers of loci favoring one topology over other two based on ΔlnL in the overall dataset and among the strong outliers, when assessed using individually fit loci. b) The numbers of loci favoring one topology over other two based on ΔlnL in the overall dataset and among the strong outliers, when assessed in concatenation. c) The overall distribution of sCF, black horizontal bar representing the mean, dashed line shows the majority threshold of sCF (Minh et al 2020). f) The numbers of loci favoring one topology over the other two based on sCF.

Fig. 6. Properties of the loci strongly favoring each of the treeshrew phylogenetic hypotheses (ΔlnL > 0.1).

Supplementary information

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.[NNNN]

Supplementary Figures

Fig. S1. Properties of the evaluated SISRS loci. A set of properties on the left (blue) contains size and information related features, a set of properties on the right (pink) contains compositional features, the bottom center section shows counts of different locus types.

Fig. S2. GC content of different locus types in the SISRS dataset.

Fig. S3. Interrelationship between evaluated loci properties, assessed via a principal component analysis. Only the first two PCA components are shown.

Fig. S4. ΔlnL-based signal for the three competing hypotheses of treeshrew placement, assessed a) using per locus fit, or b) in concatenation, dashed lines represent cutoffs for strong outliers.

Fig. S5. ΔlnL-based signal for monophyly of Primates, assessed a) using per locus fit, or b) in concatenation, dashed lines represent cutoffs for strong outliers.

Fig. S6. Interrelationship between ΔlnL and sCF for individual loci.

Fig. S7. Properties of the loci favoring each of the treeshrew phylogenetic hypotheses (all loci).

Supplementary Tables

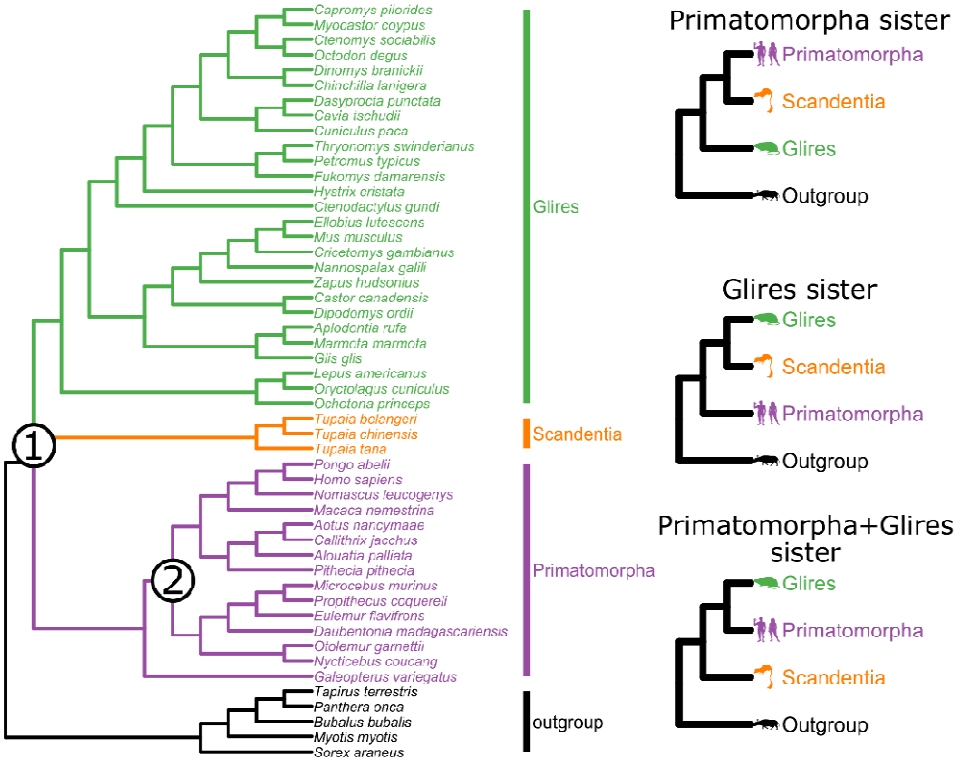
Tab. S1. Sample information

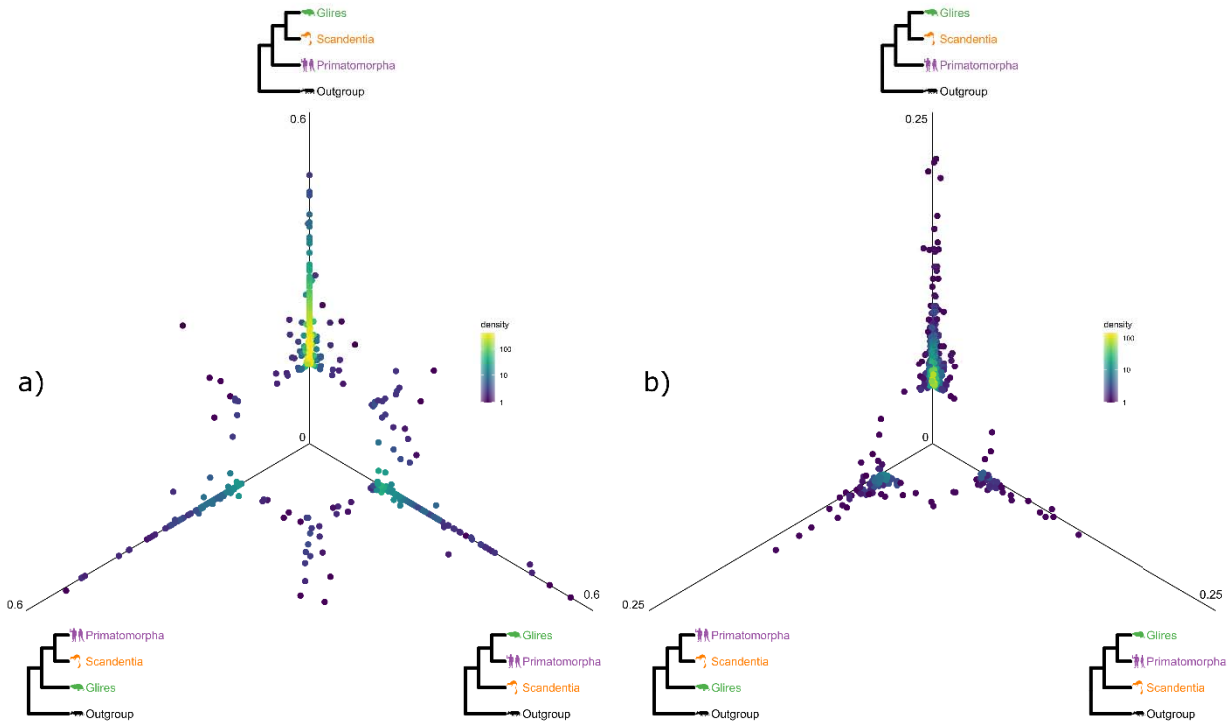
Tab. S2. Loci properties and signal

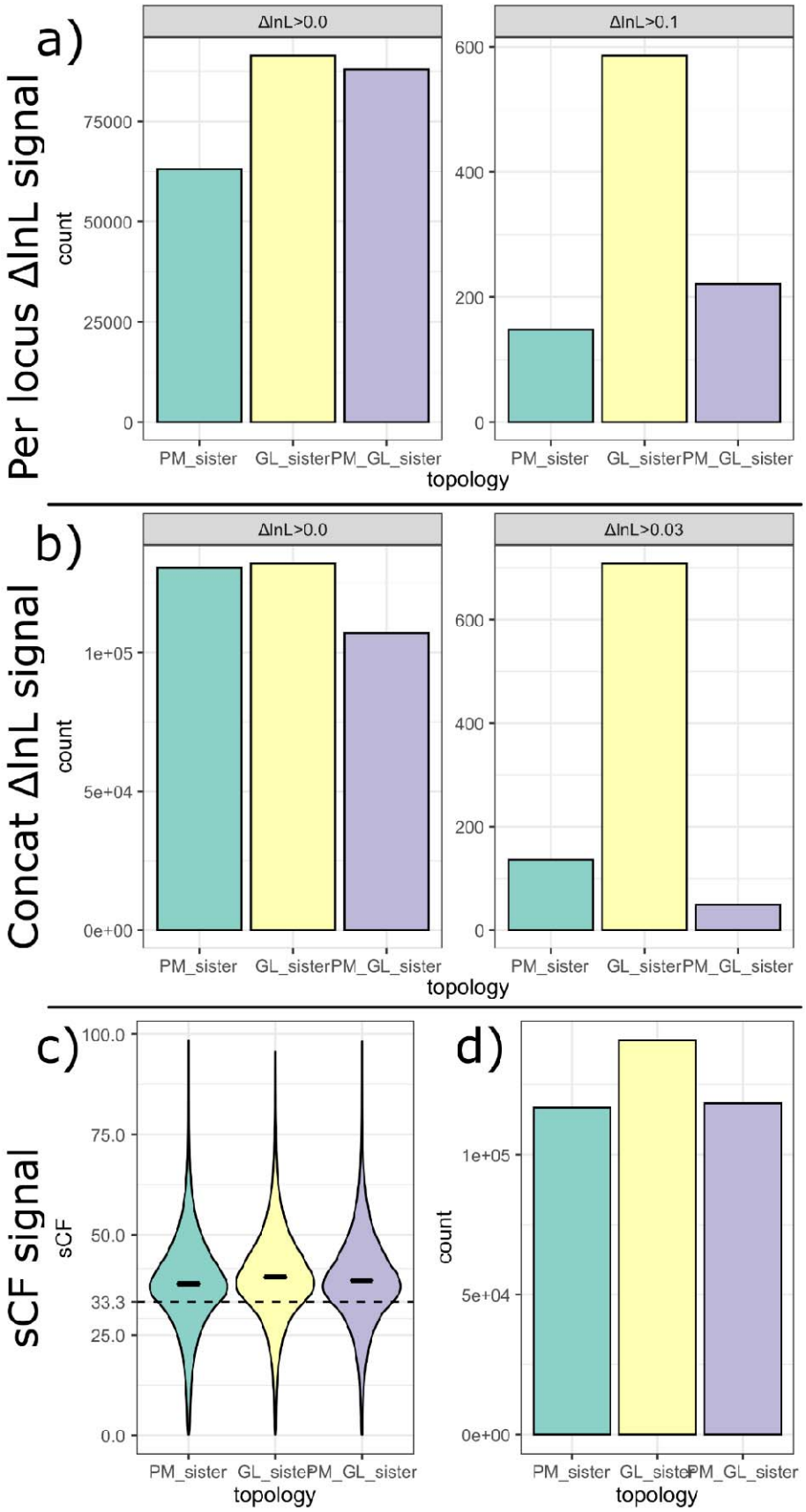
Tab. S3. Results of the statistical analyses

Miscellaneous

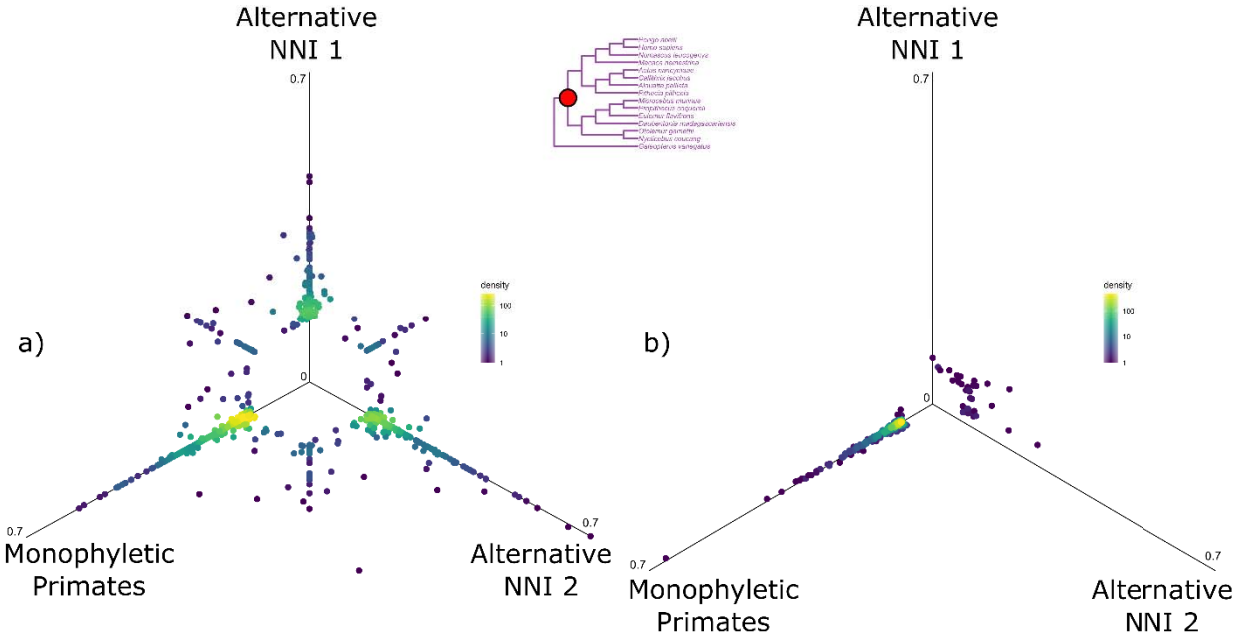
File. S1. Compressed sequences of SISRS loci

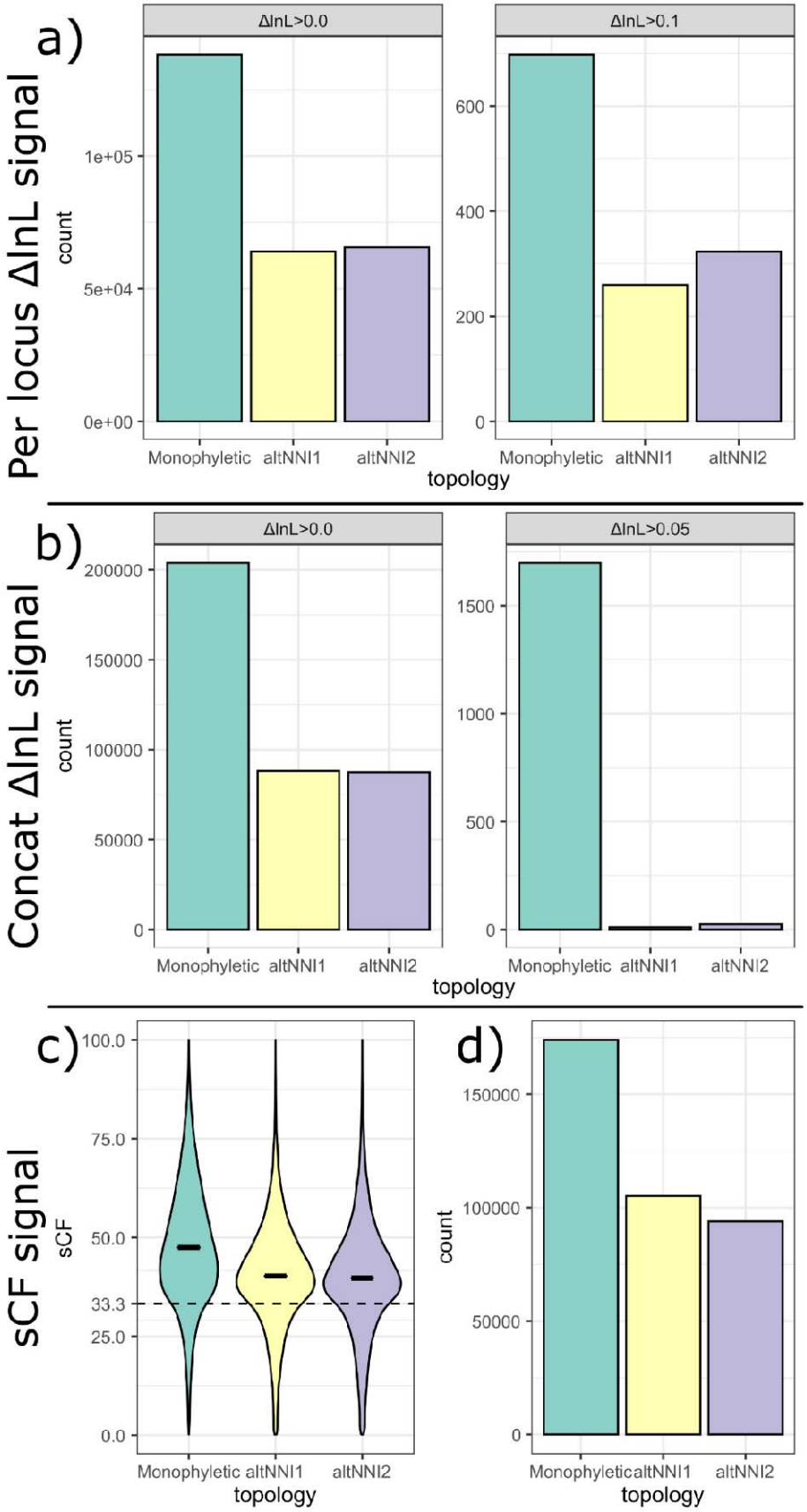

Capromys pilorides
Myocastor coypus
Ctenomys sociabilis
Octodon degus
Dinomys branickii
Chinchilla lanigera
Dasyprocta punctata
Cavia ischudii
Cuniculus paca
Thryonomys swinderianus
Petromus typicus
Fukomys damarensis
Hystrix cristata
Ctenodactylus gundi
Ellobius lutescens
Mus musculus
Cricetomys gambianus
Nannospalax galili
Zapus hudsonius
Castor canadensis
Dipodomys ordii
Aplodontia rufa
Marmota marmota
Gliis glis
Lepus americanus
Oryctolagus cuniculus
Ochotona princeps
Tupaia belangeri
Tupaia chinensis
Tupaia tana
Pongo abelii
Homo sapiens
Nomascus leucogenys
Macaca nemestrina
Aotus nancymaae
Callithrix jacchus
Alouatta palliata
Pithecia pithecia
Microcebus murinus
Propithecus coquereli
Eulemur flavifrons
Daubentonia madagascariensis
Otolemur garnettii
Nycticebus coucang
Galeopterus variegatus
Tapirus terrestris
Panthera onca
Bubalus bubalis
Myotis myotis
Sorex araneus
1
2
Glires
Scandentia
Primatomorpha
outgroup
Primatomorpha sister
Primatomorpha
Scandentia
Glires
Outgroup
Glires sister
Glires
Scandentia
Primatomorpha
Outgroup
Primatomorpha+Glires sister
Glires
Primatomorpha
Scandentia
Outgroup


a)
b)
Glires
Scandentia
Primatomorpha
Outgroup
0.6
0.25
0
density
100
10
1

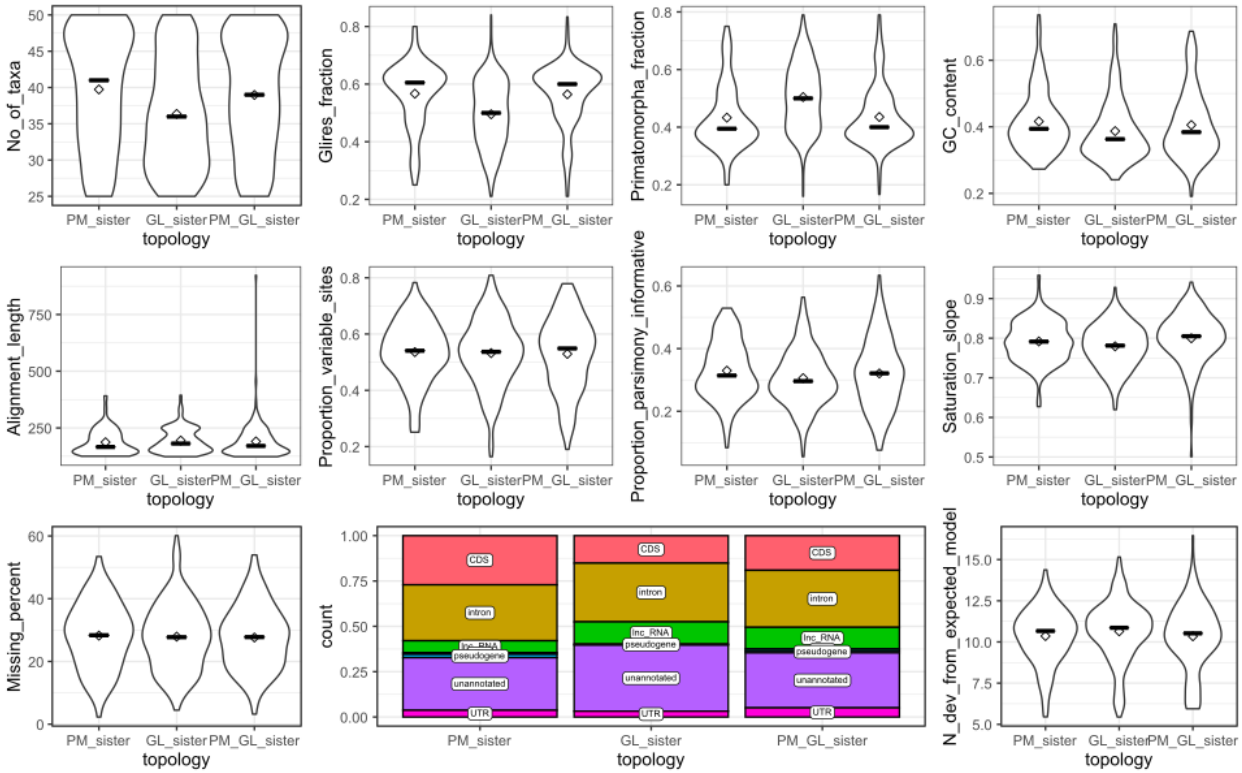

a)
Per locus ΔlnL signal
ΔlnL>0.0
ΔlnL>0.1
count
75000
50000
25000
0
600
400
200
0
PM_sister
GL_sister
PM_GL_sister
topology
b)
Concat ΔlnL signal
ΔlnL>0.0
ΔlnL>0.03
count
1e+05
5e+04
0e+00
600
400
200
0
PM_sister
GL_sister
PM_GL_sister
topology
c)
sCF signal
sCF
100.0
75.0
50.0
33.3
25.0
0.0
PM_sister
GL_sister
PM_GL_sister
topology
d)
count
1e+05
5e+04
0e+00
PM_sister
GL_sister
PM_GL_sister
topology

837


Alternative
NNI 1
0.7
0
0.7
Monophyletic
Primates
0.7
Alternative
NNI 2
a)
density
100
10
1
Homo sapiens
Tarsius syrichta
Nycticebus coucang
Macaca mulatta
Galeopterus variegatus
Alternative
NNI 1
0.7
0
0.7
Monophyletic
Primates
0.7
Alternative
NNI 2
b)
density
100
10
1


a)
Per locus ΔlnL signal
ΔlnL>0.0
ΔlnL>0.1
count
1e+05
5e+04
0e+00
600
400
200
0
Monophyletic
altNNI1
altNNI2
topology
b)
Concat ΔlnL signal
ΔlnL>0.0
ΔlnL>0.05
count
200000
150000
100000
50000
0
1500
1000
500
0
Monophyletic
altNNI1
altNNI2
topology
c)
sCF signal
sCF
100.0
75.0
50.0
33.3
25.0
0.0
Monophyletic
altNNI1
altNNI2
topology
d)
count
150000
100000
50000
0
Monophyletic
altNNI1
altNNI2
topology

841


No_of_taxa
Glires_fraction
Primatomorpha_fraction
GC_content
Alignment_length
Proportion_variable_sites
Proportion_parsimony_informative
Saturation_slope
Missing_percent
count
N_dev_from_expected_model
topology
PM_sister
GL_sister
PM_GL_sister
CDS
intron
lnc_RNA
pseudogene
unannotated
UTR