Genome-scale profiling reveals higher proportions of phylogenetic signal in non-coding data

Robert Literman and Rachel S. Schwartz

Abstract

Accurate estimates of species relationships are integral to our understanding of the evolution of traits and lineages, yet many relationships remain controversial despite whole genome sequence data. These controversies are due, at least in part, to complex patterns of phylogenetic and non-phylogenetic signal coming from regions of the genome experiencing different evolutionary forces. Here we profile the phylogenetic informativeness of loci from across mammalian genomes. We identified orthologous sequences from primates, rodents, and pecora, and annotated millions of sites as one or more of nine locus types (e.g. coding, intronic, intergenic) and profiled the informativeness of different locus types across the evolutionary timescales of each clade. In all cases, non-coding loci provided more overall signal and a higher proportion of phylogenetic signal compared to coding loci. Most locus types provide relatively consistent phylogenetic information across timescales, although we find evidence that coding and intronic regions may inform disproportionately about older and younger splits, respectively, to a limited degree. We also validated the SISRS pipeline as an annotation-free ortholog discovery pipeline that identifies millions of phylogenetically informative sites directly from whole-genome sequencing reads.

Introduction

Accurate estimates of species relationships are integral to our understanding of the evolution of traits and lineages, from modeling the co-evolution of hosts and pathogens to conserving biodiversity (Buerki et al. 2015; Bentley 2016). The accessibility of next-generation sequencing technology, together with a broader availability of powerful computational resources, is facilitating a discipline-wide shift away from resolving species trees using a small set of markers and towards the analysis of thousands of loci from across the genome. There was early optimism that increasing the size of datasets and selecting markers from a more evolutionarily-diverse locus pool would lead to the swift resolution of some of the biggest and

most problematic questions in evolutionary biology (Gee 2003). However, despite these genome-scale efforts many species and clade relationships remain unresolved. Phylogenies inferred from different locus types, even at genomic scales, continue to generate alternative topologies (Jarvis et al. 2014; Sharma et al. 2014; Rokas et al. 2003; Nosenko et al. 2013). As the concept of using genome-scale data transitions from cutting edge to commonplace, these conflicting results among genomic subsets hinder our understanding of broader evolutionary processes; thus, they deserve explicit inspection (Reddy et al. 2017).

When comparing orthologous sequences among species, sites may be said to carry 'phylogenetic signal' when the sequence variation separates species or clades in a manner that reflects their evolutionary history. Conversely, 'non-phylogenetic signal' describes sequence variation that does not reflect underlying species relationships as in cases of convergent evolution, homoplasy, or incomplete lineage sorting (Song et al. 2012; Rokas and Carroll 2008; Li, Gojobori, and Nei 1981). Using sets of phylogenetic markers that contain higher ratios of non-phylogenetic to phylogenetic signal can lead to incorrect tree reconstructions even at the genomic scale, as non-phylogenetic signal is scaled up along with phylogenetic signal (Cao et al. 1994; Hillis and Huelsenbeck 1992; Reddy et al. 2017; Philippe et al. 2011).

For decades, attempts to resolve more recent splits (e.g. those within a genus or family) have used sequence data from classically fast-evolving loci, such as intronic (Debry and Seshadri 2001; Omland 2003) or intergenic regions (Baldwin and Markos 1998; Shaw et al. 2005). These sites are generally thought to accrue substitutions rapidly enough to generate sufficient phylogenetic signal at shorter timescales. Conversely, resolution of deep-time splits often use more slow-evolving protein-coding sequences (CDS) or other more constrained genomic subsets as a way to limit the impact of overwriting mutations (e.g. homoplasy) over longer evolutionary timescales (Managadze et al. 2011; Goldman and Yang 1994; Hughes and Yeager 1998; Bromham, Rambaut, and Harvey 1996; Ren et al. 2016; Salichos and Rokas 2013; Dornburg et al. 2014). While this marker selection strategy appears reasonable, studies of the relationship between rate and phylogenetic utility are often complicated by interacting factors (Aguileta et al. 2008; Townsend and Leuenberger 2011; Dornburg, Su, and Townsend 2019; Steel and Leuenberger 2017; Heath et al. 2008; Su and Townsend 2015).

Profiling the phylogenetic informativeness of loci when a robustly supported tree is available provides an opportunity to contrast the relative amounts and sources of phylogenetic and non-phylogenetic signal coming from different subsets of the data (Graybeal 1994; Russo, Takezaki, and Nei 1996; Townsend 2007). Results from informativeness studies have found support for the idea that some fast-evolving loci providing more information for more recent splits while slow-evolving loci provide more support for ancient speciation events (Townsend, López-Giráldez, and Friedman 2008; Fong and Fujita 2011). However, the relative difficulty of processing non-coding data (e.g. annotation and alignment) has led to a majority of quantitative studies of informativeness focusing on comparisons among CDS types (Fong and Fujita 2011; Russo, Takezaki, and Nei 1996; Townsend, López-Giráldez, and Friedman 2008; Moeller and Townsend 2011; Graybeal 1994). Understanding the distribution of phylogenetic and non-phylogenetic signal across the genome relies on the ability to identify and compare orthologs beyond genes and coding regions, which can be challenging in clades lacking high-level genomic resources. However, explicit informativeness assessments for non-coding locus types are critical for contextualizing what we understand about coding regions (Granados Mendoza et al. 2013; Small et al. 1998). In fact, one such study suggests that CDS sites may carry a higher proportion of non-phylogenetic signal than intronic regions (Chen, Liang, and Zhang 2017). These results are especially notable given that CDS markers alone are often used to infer trees. However, genic regions still account for only a modest percentage of the total information content of the genome (Sims et al. 2009) and may accrue mutations at a fundamentally different rate than functionally-distinct locus types. Until these comparisons of phylogenetic utility are made at the genome scale, the relative informativeness of vast swaths of the genome will remain poorly understood which limits our ability to fully resolve interspecies relationships.

In this study, we compare the distribution of phylogenetic and non-phylogenetic signal among nine different mammalian locus types including coding, intronic, and intergenic regions. In the absence of reference genomes or annotation information, the SISRS bioinformatics pipeline (Schwartz et al. 2015) facilitates genome-wide ortholog comparison directly from unannotated whole-genome sequencing (WGS) reads, enabling genome-scale informativeness

comparisons among locus types. Using WGS data from focal sets of primates, rodents, and pecora, we identify millions of potentially informative sites in the absence of reference information. After post-hoc annotation, we find that the vast majority of these sites come from intronic, long non-coding RNA (lncRNA), and intergenic regions, highlighting the need for phylogenetic signal profiling in these overlooked but information-rich genomic subsets. While the majority of sites from this study carry phylogenetic signal, CDS-derived sites were found to carry the highest proportions of non-phylogenetic signal in each dataset. This is concerning given that coding sequences are often targeted for use in phylogenetics studies. At the evolutionary timescales associated with this study, we find only sporadic support for the differential phylogenetic utility of classically fast- and slow-evolving locus types to resolve nodes of different ages, with most subsets actually informing broadly.

Methods

All associated scripts and relevant output, can be found in the companion GitHub repository: https://github.com/BobLiterman/PhyloSignal_MS

Raw Data Processing

Assessing the phylogenetic utility of genomic data relies on a robustly-supported underlying evolutionary hypothesis. To that end we obtained paired-end Illumina whole-genome sequencing (WGS) reads from the European Nucleotide Archive (Leinonen et al. 2011) for three well-supported mammal clades: catarrhine primates, murid rodents, and members of the infraorder Pecora. Each dataset included ten focal taxa as well as two outgroup taxa, all with well-supported evolutionary relationships (Figure 1) (Reis et al. 2018; Zurano et al. 2019; Steppan and Schenk 2017). To enable downstream ortholog annotation, each focal dataset contained one species with a well-assembled and well-annotated reference genome (Primates: *Homo sapiens*, Rodents: *Mus musculus*, Pecora: *Bos taurus*). We also ran a combined analysis with all taxa that we annotated using the *H. sapiens* reference genome. We assessed read data quality before and after trimming using FastQC v0.11.5 (S. Andrews -

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and raw reads were trimmed using BBDuk v.37.41 (B. Bushnell - sourceforge.net/projects/bbmap/).

Extracting phylogenetic data from WGS reads

For each dataset (Primates, Rodents, Pecora, and Combined), we used the SISRS pipeline to identify putative orthologs and phylogenetically informative sites (Schwartz et al. 2015). SISRS identifies orthologous loci by assembling a "composite genome" from a subsample of reads pooled across all species. By subsampling reads prior to assembly, regions of relatively high conservation will have sufficient depth for assembly while taxon-specific or poorly conserved regions will fail to assemble. With a genome size estimate of 3.5Gb per dataset (Kapusta, Suh, and Feschotte 2017), reads were subsampled equally from each taxon so that the final assembly depth was ~10X genomic coverage (e.g. 35Gb per composite genome assembly). We used Ray v.2.3.2-devel to assemble the composite genome, with default parameters and a k-value of 31 (Boisvert, Laviolette, and Corbeil 2010). This assembly represents a conserved and "taxonomically-averaged" subset of the shared regions of the genome against which all taxa can be compared.

In order to generate species-specific ortholog sets, SISRS maps the trimmed WGS reads from each taxon against their respective composite genome. Reads that mapped to multiple composite scaffolds were removed from analysis prior to composite genome conversion. When two key conditions are met (sites must be covered by at least three reads and must not vary within the taxon), SISRS uses the mapping information from each species to replace bases in the composite genome with species-specific bases. Any sites with insufficient read coverage or within-taxon variation were denoted as 'N'. SISRS removes sites that lack interspecies variation (invariant sites), as these provide no direct topological support when inferring phylogenetic trees. For this study, we also removed alignment columns containing any missing data (e.g. Ns) as well as sites where the variation consisted of a gap and an otherwise invariant nucleotide. Finally, SISRS partitions the final alignment into overlapping subsets containing (1) all variable sites, (2) all variable sites with singletons removed (parsimony informative sites), and (3) parsimony informative sites with only two alleles (biallelic sites). We used the SISRS-filtered

dataset containing only gapless, biallelic sites where data is present for all taxa ('SISRS Sites') for subsequent phylogenetic analysis.

Data filtering and annotation

We obtained chromosome and mitochondrial scaffolds along with associated annotation data for *Homo sapiens, Mus musculus,* and *Bos taurus* from the Ensembl Build 92 database (Zerbino et al. 2018). For each reference species, we mapped their taxon-converted composite sequences onto the reference genome using Bowtie2 v.2.3.4 (Langmead and Salzberg 2012). We removed any contigs that either did not map or mapped equally well to multiple places in the reference genome, as this obscured their evolutionary origin. We also removed individual sites that displayed overlapping coverage from independent scaffolds to avoid biasing downstream results through redundant counting or by arbitrarily favoring alleles in one contig over another.

We scored each mapped composite genome site as one or more of the following eight annotation types: coding sequence (CDS, including all annotated transcript variants), untranslated regions (both 5'- and 3' UTRs), intronic (gene region minus CDS/UTR), long-noncoding RNA (lncRNA; none annotated in Pecora), noncoding gene (genes without annotated CDS; none annotated in Pecora), pseudogenes, or small RNAs (smRNA; miRNAs + ncRNAs + rRNAs + scRNAs + smRNAs + snoRNAs + snRNAs + tRNAs + vaultRNAs). Any reference genome position that was not annotated as one of the these locus types was denoted as intergenic/unannotated. In some cases an individual site may have multiple annotations, such as pseudogenes within introns, or alternative five-prime UTR regions overlapping CDS. SISRS composite sites were annotated using the reference annotation files, the output from the reference genome mapping, and the intersect function in BEDTools v.2.26 (Quinlan 2014).

Assessing annotation-specific biases in composite genome assembly and SISRS filtering

In this study, phylogenetic signal analysis was limited to (1) sites within loci that were assembled into the composite genome, and of those, (2) sites that passed through all SISRS filtration steps. For each annotation subset, we calculated the proportion of sites from the

reference genome that had been assembled as part of the composite genome. We compared the relative assembly percentages among annotation types using a two-tailed modified Z-score analysis, which is robust at detecting deviations within small sample sizes (Leys et al. 2013; Dobbie 1963). Based on the number of annotation subsets present in each dataset, critical Z-score values indicative of significant assembly biases were identified at a Bonferroni-corrected $\alpha = 0.05/9 = 5.56E^{-3}$ (Primates, Rodents, and Combined; $Z_{Critical} = 2.77$) or $\alpha = 0.05/7 = 7.14E^{-3}$ (Pecora; $Z_{Critical} = 2.69$).

Every site assembled into the composite genome was subjected to SISRS filtration, including the removal of hypervariable sites that vary within single taxa, invariant sites among taxa, and sites with more than two possible alleles. Locus types with exceptionally high or low filtration rates were identified using the modified Z-score analysis described above.

Assessing the phylogenetic signal among locus types from SISRS-filtered data

Biallelic sites partition a dataset into a pair of taxonomic groups, each defined by their unique fixed allele. Based on which groups of taxa were supported, the variation at each biallelic SISRS site was scored as carrying phylogenetic or non-phylogenetic signal relative to the reference topologies. We tallied support for each unique split signal found in the data, including all concordant and discordant splits. For each dataset, we assessed differences in concordant and discordant site support levels using a Wilcoxon test as implemented in R (R Core Team, 2013). For each annotation subset, we calculated the proportion of sites carrying phylogenetic signal and identified statistical outliers using the modified Z-score analysis.

We built phylogenies using data from the raw SISRS output (prior to reference genome mapping) as well as from each annotation-specific dataset. We inferred all trees with a maximum-likelihood approach using a Lewis-corrected GTR+GAMMA model (Lewis 2001), concatenated data, rapid bootstrap analysis, and autoMRE stopping criteria as implemented in RAxML v.8.2.11 (Stamatakis 2014). We visually assessed trees for concordance with the reference topology in Geneious v.9.1.8 (https://www.geneious.com).

Detecting changes in phylogenetic utility over evolutionary time among locus types

To detect changes in the relative phylogenetic utility of different genomic subsets over evolutionary time, we dated each split in the reference topologies using alignment data from SISRS orthologs. Briefly, for each dataset we sorted the assembled composite orthologs based on their final SISRS site counts. The SISRS composite genome contains many contigs that are highly conserved or even completely invariant among species, so sorting by variable SISRS sites ensured that each analyzed ortholog had variation for node dating. Using mafft v.7.310 (Katoh 2002), we aligned the 50,000 most SISRS-rich orthologs and used sequence variation among these loci to estimate branch lengths on the fixed reference topology under a GTR+GAMMA model as implemented in RAxML (Schwartz and Mueller 2010). With these branch lengths, we applied a correlated substitution rate model to estimate node ages on each reference topology in R using the *chronos* function as implemented in the package *ape* v.5.3 (Paradis and Schliep 2019)**.** To convert relative split times into absolute divergence time estimates, we calibrated specific nodes in the reference topologies using divergence time information from the TimeTree database (Kumar et al. 2017). Focal groups were calibrated at the root node using the TimeTree divergence time confidence intervals as minimum and maximum bound estimates. The Combined topology was calibrated in the same way at the base of the tree as well as with the focal group calibrations. Due to stochasticity in the split time estimation process, we inferred each node age 1000 times and used the median value in all downstream analyses.

For each split in the reference topologies, we broke down the percent support for each locus type. We used linear models in R to assess whether the relative phylogenetic utility of different locus types changed over evolutionary time. Statistical significance was interpreted at Bonferroni-corrected α values based on the number of annotation types per dataset.

Results

Processing of WGS reads into annotated composite genomes

Based on a genome size estimate of 3.5Gb, post-trimming read depths ranged from 10X - 38X among species (Table S1). The SISRS composite genomes, which we assembled using reads pooled across taxa, contained 2M - 6M contigs ranging from 123bp - 18Kb in length with an average N50 value of 150bp (Table S2). Through reference genome mapping, we were able

to identify and annotate 32% - 79% of these composite sites (103M - 422M sites per dataset, Table 1, Tables S2-3). For each dataset and among locus types, we compared the proportion of sites from each reference annotation type that had been assembled and annotated in the SISRS composite genome (Figure 2A, Table 1, Table S3). Reference loci annotated as pseudogenic were significantly less likely to be assembled into the composite genome of Rodents (Z_{MOD} = 4.72 ; p = $2.3E^{-6}$) and Pecora (Z_{MOD} = 2.75 ; p = $5.9E^{-3}$), while CDS loci were assembled with a higher probability relative to other locus types in the Combined analysis (Z_{MOD} = 4.29 ; p=$1.8E^{-5}$).

Clades vary in which locus types are selected as potentially informative (or non-informative)

SISRS called the taxon-specific base for every position in the composite genome, provided there was sufficient read support for a fixed allele (Table S4). Then from these sites SISRS identified 337K - 11.8M sites per dataset with (1) sequence data for all species and (2) biallelic variation among taxa (Table S5). After reference genome mapping, we were able to annotate 300K - 11.5M of these biallelic sites (82% - 97% of SISRS sites) for use in downstream phylogenetic analyses (Table 1, Table S5). We compared the relative proportions of different annotation types that passed through the SISRS filtering process (Figure 2B, Table 1, Table S6). Among Pecora locus types, SISRS filtered out pseudogenic and smRNA sites at a higher rate than other subsets (Z_{MOD} = 3.17, 3.00; p = $1.5E^{-3}$, $2.7E^{-3}$). Conversely, CDS sites in the Combined dataset successfully passed through SISRS filtration significantly more often than other annotation types (Z_{MOD} = 8.16 ; p = $3.3E^{-16}$). The same CDS trend was even more pronounced in Rodents (Z_{MOD} = 21.1 ; p = $4.1E^{-99}$), where the five- and three-prime UTR also showed higher rates of SISRS site selection (Z_{MOD} = 4.60, 6.15; p = $4.3E^{-6}$, $7.6E^{-10}$).

SISRS identifies phylogenetically useful sites, but non-coding sites provide higher proportions of phylogenetic signal

We scored the variation at each SISRS site as carrying either phylogenetic or non-phylogenetic signal; among datasets 68% - 90% of SISRS sites supported a split in the reference topology (Table S5). From each set of SISRS sites, concordant splits had statistically higher support than non-reference splits ($p < 2.3E^{-7}$, Table S7, Figure S1).

Phylogenies inferred for each focal group from the raw, unannotated SISRS output as well as from each locus type separately were concordant with the reference topology with high bootstrap support (all BS > 80). In the Combined analysis, the 763 noncoding gene sites were sufficient to resolve all but one node in the tree with high bootstrap support, failing to resolve the grouping of hippo and whale within the pecora. Similarly, the 139 identified smRNA sites left many relationships poorly resolved, although did structure major groups and many sub-groups. Alignments and trees are available from the companion GitHub repository.

Concordance percentages varied among locus types (Figure 2C, Table 1, Table S8). In each dataset we found that the CDS sites carried a significantly higher proportion of non-phylogenetic signal compared to most other locus types (Pecora: $Z = 2.98$, $p = 2.9E^{-3}$; Primates: $Z=11.8$, $p = 4.9E^{-32}$; Rodents: $Z = 16.7$, $p = 2.4E^{-62}$; Combined: $Z=4.70$; $p = 2.6E^{-6}$). Concordance percentages for CDS SISRS sites (56.6% - 88.2%) were 2.2% - 19.8% lower than the median non-coding percentage (70.6% - 90.2%) (Figure 2C, Table 1, Table S8). In addition to CDS sites, smRNA sites also carried a significantly higher proportion of non-phylogenetic signal in Primates relative to other subsets ($Z = 3.00$, $p = 2.7E^{-3}$). No locus type in any dataset displayed a statistically higher percentage of site concordance (Figure 2C, Table 1, Table S8).

<u>Most genomic subsets show no change in phylogenetic utility over focal timescales</u>

Estimated split times among focal groups ranged from 2.1 MY - 25.1 MY, while node ages in the Combined dataset were estimated between 1.4 MY - 46.6 MY (Table S9). Using linear models to determine whether the relative proportions of phylogenetic signal coming from different locus types changed significantly over time (Figure 3, Table S10), we detected no significant association between relative phylogenetic informativeness and node age for seven of the nine locus types (five-prime UTR, intergenic, lncRNA, ncGenes, pseudogenes, smRNA, and three-prime UTR) (Table S10). As node age increased (i.e. for older splits) the proportion of node support coming from CDS sites increased 1.3% - 2.3% per million years in the Pecora, Rodent, and Combined datasets ($p = 2.46E^{-3}$, 3.40^{-3},$1.40E^{-3}$ respectively; Figure 3, Table S10). Conversely, the relative proportion of node support coming from intronic sites in Rodents and

the Combined datasets decreased by 0.29% and 0.32% per million years, respectively ($p < 1.14E^{-3}$, Figure 3, Table S10).

Discussion

The increasing availability of publicly archived NGS data and lower costs associated with high-throughput sequencing are driving the field of phylogenetics towards genome-scale analyses. However, even genome scale data has not yet allowed us to resolve many controversial relationships due to conflicting signal in the data (Jarvis et al. 2014; Sharma et al. 2014; Rokas et al. 2003; Nosenko et al. 2013). Here we have taken a step toward addressing how we can look at the signal in the data across a diverse set of locus types. We see that even by broadly categorizing types of genetic data that some types of data carry more information, while others carry a higher proportion of non-phylogenetic signal that misleads our understanding of species diversification.

Non-coding loci carry a higher proportion of phylogenetic signal

In each dataset, non-coding sites provided more total phylogenetic signal and a higher proportion of phylogenetic signal relative to CDS-derived sites. CDS carried the highest proportion of non-phylogenetic signal of any locus type in each dataset, suggesting that the same characteristics that make coding sequences an attractive target for marker development (e.g. constrained evolution) may be linked with higher relative amounts of non-phylogenetic signal (Chen, Liang, and Zhang 2017).

CDS are often used for phylogenetic analysis due in part to the relative ease of generating sequence data (e.g. transcriptomics), combined with low mutation rates that facilitate straightforward locus annotation and alignment. In the taxonomically-diverse combined dataset, high levels of sequence conservation among coding regions likely explains the observation that CDS were assembled proportionally more often than any other locus type. Meanwhile in the rodent and combined datasets, CDS sites more commonly passed through SISRS filtration steps, which indicates a phylogenetically-useful substitution rate pattern. However, the lower proportions of phylogenetic signal among CDS in all clades provides further

evidence that mutation rate may not be the best overall predictor of phylogenetic informativeness, and underlies the necessity for broader characterizations of informativeness to provide greater context.

Phylogenetic informativeness of most locus types is stable across evolutionary timescales

In contrast to expectations of how markers from different genomic subsets are useful in phylogenetics studies (Dornburg, Su, and Townsend 2019; Steel and Leuenberger 2017; Phillips, Delsuc, and Penny 2004), seven of the nine locus types showed broad phylogenetic utility over the timescales associated with this study. Intronic sites were moderately more informative for younger nodes in the rodent and combined datasets, while CDS sites had a stronger trend towards informativeness in older nodes among pecora, rodents, and the combined dataset. That these trends did not hold across all clades suggests that choosing phylogenetic markers to resolve nodes of a certain age should be based on an understanding of how those loci evolve among focal species rather than with a broad, 'one size fits all' approach. The trend of CDS sites to provide more information for older splits may imply that they would be of significantly greater utility for deep-time phylogenetics as often suggested (Graybeal 1994; Ren et al. 2016; Salichos and Rokas 2013); yet, we these sites carry the highest amount of non-phylogenetic signal of all analyzed locus types, which will affect tree inference negatively.

SISRS isolates phylogenetically useful orthologs directly from WGS reads

Well-assembled and annotated reference genomes currently exist for only a small fraction of life on Earth. Reference-guided phylogenetic marker development is limited by the availability of these resources. However, SISRS bypasses the need for reference annotation by identifying orthologs directly from WGS reads. SISRS filtering yields millions of sites per focal clade, the vast majority of which carried accurate phylogenetic signal. SISRS data derives from a variety of annotation types including both genic regions and non-genic regions, permitting genome-scale comparisons of relative phylogenetic informativeness. Less total data and the lowest overall concordance rate in the combined dataset may reflect the expected deterioration of phylogenetic signal over longer stretches of evolutionary time (Salichos and

Rokas 2013; Pisani et al. 2012; Townsend, Su, and Tekle 2012). It is also important to note that the sequence diversity among all 36 species is higher than any focal subset of 12, and this likely had an impact on the composite genome assembly process, which impacts sites available to SISRS.

Importantly, SISRS provides a marker identification pipeline that is immune to missing, incomplete, or incorrect reference information. For example, sites derived from lncRNA made up approximately a quarter of all assembled sites among primates and rodents, suggesting that this data type is useful for phylogenetics. However, in pecora, lncRNA loci were not annotated; thus this important data type could not be targeted for marker development in reference-based pipelines. However, intergenic sites account for nearly two-thirds of the pecora data (compared to one-third of sites in rodents and primates), suggesting that lncRNA sites are in fact included in the pecora analysis.

Conclusions

In this study, we establish broad phylogenetic informativeness profiles for nine different locus types as they apply to resolving species relationships among three mammal clades. Across 50MY of mammal evolution, we find that, in contrast to prior work suggesting that the phylogenetic utility of data types likely varies over time, only CDS and intronic regions vary, and these patterns are both subtle and clade-specific. Non-coding sites providing more total signal and CDS sites providing the highest proportion of non-phylogenetic signal in all datasets. Additionally, we validate SISRS as a mechanism to easily identify millions of sites across the genome containing valid phylogenetic signal.

Figure 1

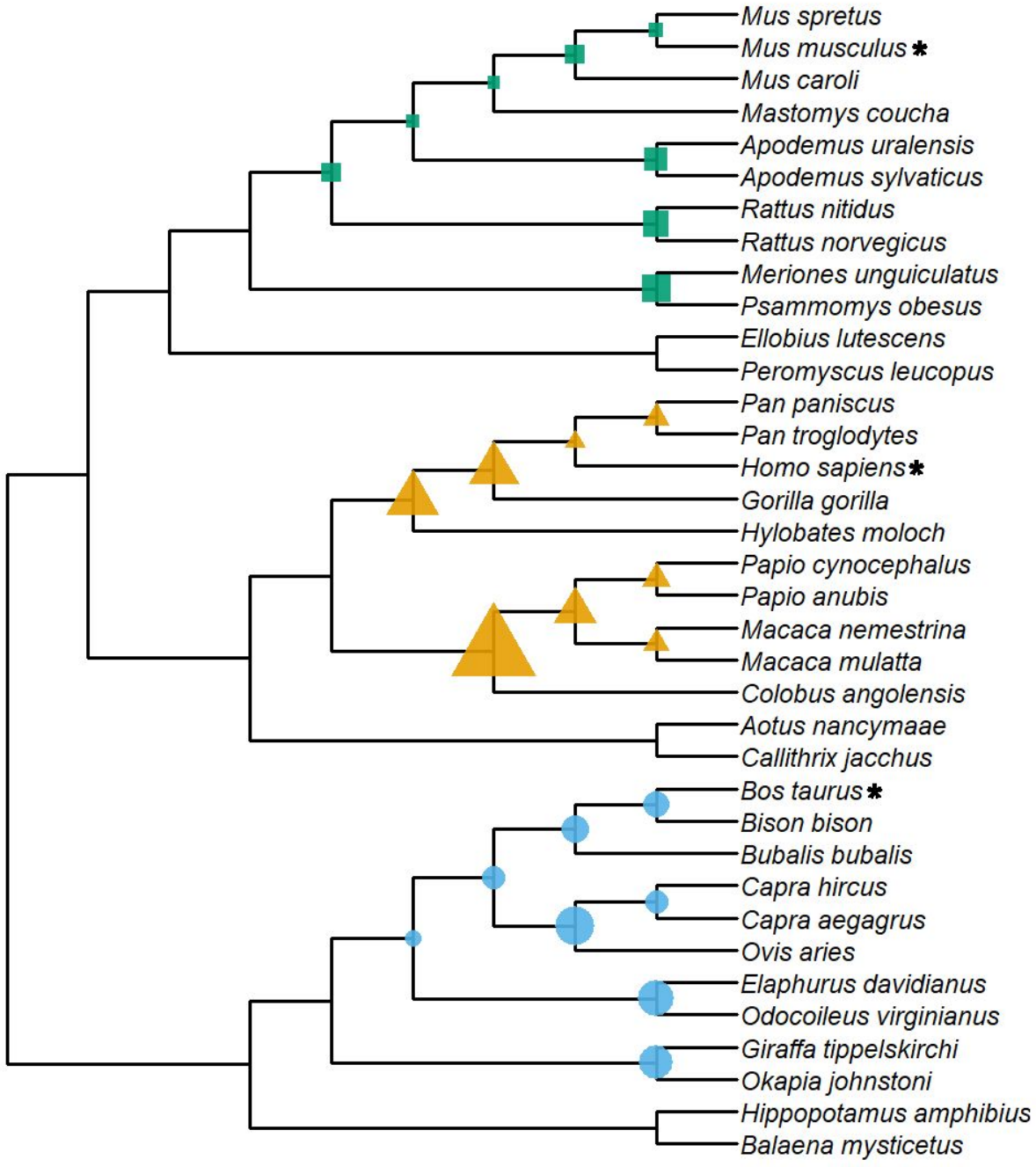
Mus spretus
Mus musculus*
Mus caroli
Mastomys coucha
Apodemus uralensis
Apodemus sylvaticus
Rattus nitidus
Rattus norvegicus
Meriones unguiculatus
Psammomys obesus
Ellobius lutescens
Peromyscus leucopus
Pan paniscus
Pan troglodytes
Homo sapiens*
Gorilla gorilla
Hylobates moloch
Papio cynocephalus
Papio anubis
Macaca nemestrina
Macaca mulatta
Colobus angolensis
Aotus nancymaae
Callithrix jacchus
Bos taurus*
Bison bison
Bubalis bubalis
Capra hircus
Capra aegagrus
Ovis aries
Elaphurus davidianus
Odocoileus virginianus
Giraffa tippelskirchi
Okapia johnstoni
Hippopotamus amphibius
Balaena mysticetus

Figure 2

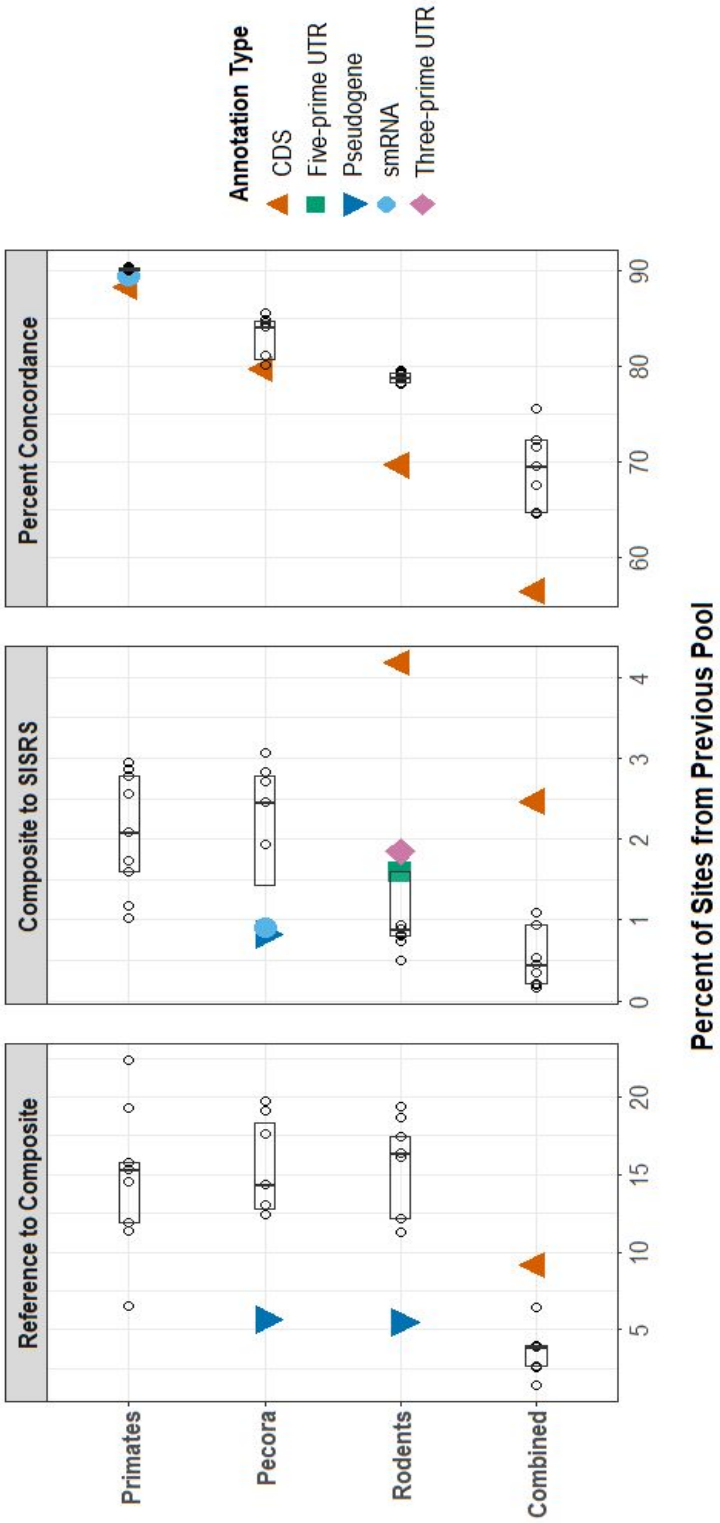
Reference to Composite
Composite to SISRS
Percent Concordance
Primates
Pecora
Rodents
Combined
5
10
15
20
0
1
2
3
4
60
70
80
90
Percent of Sites from Previous Pool
Annotation Type
CDS
Five-prime UTR
Pseudogene
smRNA
Three-prime UTR

Figure 3

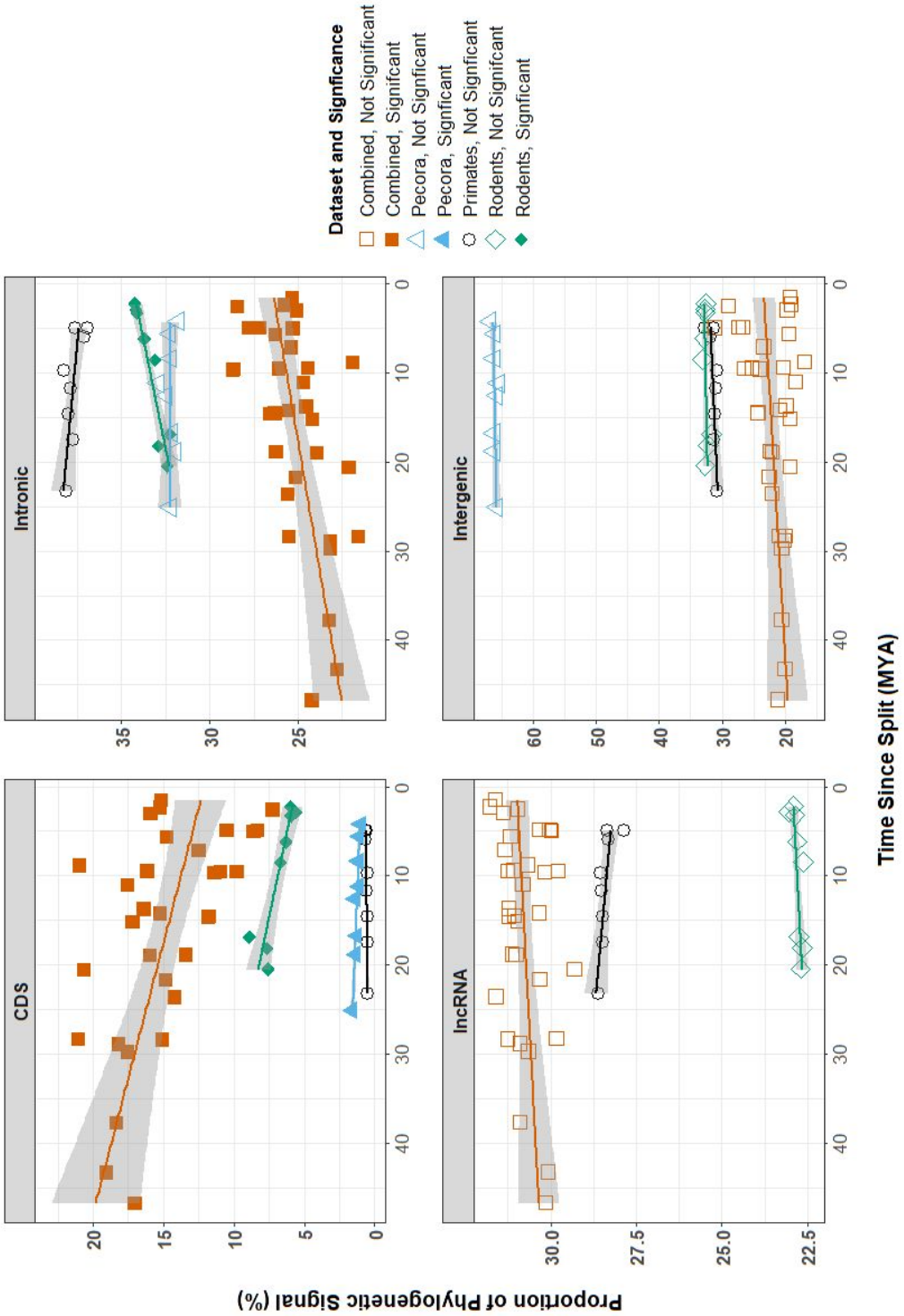
Dataset and Significance
Combined, Not Significant
Combined, Signifcant
Pecora, Not Signficant
Pecora, Signficant
Primates, Not Signficant
Rodents, Not Signifcant
Rodents, Signficant
Intronic
Intergenic
CDS
lncRNA
Proportion of Phylogenetic Signal (%)
Time Since Split (MYA)

Figure 1: Evolutionary relationships among study species. Reference annotation species are denoted with an asterisk. Node icon size is proportional to the number of SISRS sites supporting that split.

Figure 2: Proportion of sites from each locus type that (A) were assembled into the composite genome, (B) were selected by SISRS as potentially informative, and (C) carried phylogenetic signal. Filled shapes indicate locus types with a significant deviation from the data-wide median as determined by a modified Z-score analysis. Notably, whether an annotation type is disproportionately selected by SISRS for analysis is not indicative of higher ultimate phylogenetic performance, but rather a reflection of how SISRS interprets relative evolutionary rates.

Figure 3: Changes in the relative phylogenetic contribution of coding sequence (CDS), intronic, long non-coding RNA (lncRNA), and intergenic sites over time. Filled shapes indicate a significant correlation between the proportion of signal coming from a locus type, and the age of the node. CDS sites were marginally more informative for older nodes in rodents, pecora, and the combined dataset, although CDS had a smaller contribution to the overall phylogenetic signal in the data. Signal from intronic regions was proportionally higher among younger nodes in rodents and the combined dataset. Sites from lncRNA and intergenic regions showed no significant change in relative utility over time.

Tables

	Annotated Composite Sites				Annotated SISRS Sites				Percent Concordant Sites			
Annotation	**Pecora**	**Primates**	**Rodents**	**Combined**	**Pecora**	**Primates**	**Rodents**	**Combined**	**Pecora**	**Primates**	**Rodents**	**Combined**
CDS	6.37M	7.99M	6.81M	**3.27M**	156K	94.3K	**285K**	**80.6K**	**79.7 %**	**88.2 %**	**69.7 %**	**56.6 %**
Five-Prime UTR	275K	1.51M	1.19M	40.6K	5.32K	24.2K	**19.2K**	4.40K	85.6 %	90.5 %	79.2 %	64.6 %
Intergenic	234M	174M	171M	39.3M	6.63M	4.47M	1.26M	80.4K	84.6 %	90.0 %	79.5 %	72.3 %
Intronic	106M	185M	158M	46.9M	3.25M	5.48M	1.29M	96.3K	84.1 %	90.1 %	79.2 %	71.5 %
Long-noncoding RNA	-	144M	93.7M	36.9M	-	4.12M	886K	128K	-	90.2 %	78.3 %	67.7 %
Noncoding Genes	-	671K	1.58M	169K	-	18.7K	14.1K	763	-	90.4 %	78.7 %	72.3 %
Pseudogenes	**38.5K**	3.74M	**1.50M**	8.30K	**318**	64.9K	7.48K	1.33K	80.2 %	90.3 %	78.9 %	69.6 %
small RNAs	55.7K	114K	86.6K	25.9K	**508**	1.17K	690	139	81.1 %	**89.4 %**	79.6 %	75.5 %
Three-Prime UTR	1.42M	8.43M	6.98M	2.84M	38.8K	177K	**129K**	26.9K	84.9 %	90.3 %	78.1 %	64.72

Table 1: Annotation breakdown of assembled composite genome sites, sites selected by SISRS as potentially informative, and the percent of sites from each locus type carrying phylogenetic signal. Colored cells indicate annotation-specific deviations from the data-wide median, with green cells indicating over-performance and red cells indicating under-performance. CDS were disproportionately included by SISRS in rodents and the combined dataset, yet yielded the lowest percent concordance among all data subsets.

LITERATURE CITED

Aguileta, G., S. Marthey, H. Chiapello, M-H Lebrun, F. Rodolphe, E. Fournier, A. Gendrault-Jacquemard, and T. Giraud. 2008. "Assessing the Performance of Single-Copy Genes for Recovering Robust Phylogenies." *Systematic Biology* 57 (4): 613–27.

Baldwin, Bruce G., and Staci Markos. 1998. "Phylogenetic Utility of the External Transcribed Spacer (ETS) of 18S–26S rDNA: Congruence of ETS and ITS Trees ofCalycadenia(Compositae)." *Molecular Phylogenetics and Evolution*. https://doi.org/10.1006/mpev.1998.0545.

Bentley, Gillian R. 2016. "Applying Evolutionary Thinking in Medicine: An Introduction." *Evolutionary Thinking in Medicine*. https://doi.org/10.1007/978-3-319-29716-3_1.

Boisvert, Sébastien, François Laviolette, and Jacques Corbeil. 2010. "Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 17 (11): 1519–33.

Bromham, L., A. Rambaut, and P. H. Harvey. 1996. "Determinants of Rate Variation in Mammalian DNA Sequence Evolution." *Journal of Molecular Evolution* 43 (6): 610–21.

Buerki, S., M. W. Callmander, S. Bachman, J. Moat, J. -N. Labat, and F. Forest. 2015. "Incorporating Evolutionary History into Conservation Planning in Biodiversity Hotspots." *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2014.0014.

Cao, Y., J. Adachi, A. Janke, S. Pääbo, and M. Hasegawa. 1994. "Phylogenetic Relationships among Eutherian Orders Estimated from Inferred Sequences of Mitochondrial Proteins: Instability of a Tree Based on a Single Gene." *Journal of Molecular Evolution* 39 (5): 519–27.

Chen, Meng-Yun, Dan Liang, and Peng Zhang. 2017. "Phylogenomic Resolution of the Phylogeny of Laurasiatherian Mammals: Exploring Phylogenetic Signals within Coding and Noncoding Sequences." *Genome Biology and Evolution* 9 (8): 1998–2012.

Debry, R. W., and S. Seshadri. 2001. "Nuclear Intron Sequences for Phylogenetics of Closely Related Mammals: An Example Using the Phylogeny of Mus." *Journal of Mammalogy*. https://doi.org/10.1093/jmammal/82.2.280.

Dobbie, James M. 1963. "THE NEGATIVE BINOMIAL DISTRIBUTION: COMPUTATION OF THE MEDIAN AND THE MEAN ABSOLUTE DEVIATION." https://doi.org/10.21236/ad0638740.

Dornburg, Alex, Zhuo Su, and Jeffrey P. Townsend. 2019. "Optimal Rates for Phylogenetic Inference and Experimental Design in the Era of Genome-Scale Data Sets." *Systematic Biology* 68 (1): 145–56.

Dornburg, Alex, Jeffrey P. Townsend, Matt Friedman, and Thomas J. Near. 2014. "Phylogenetic Informativeness Reconciles Ray-Finned Fish Molecular Divergence Times." *BMC Evolutionary Biology* 14 (August): 169.

Fong, Jonathan J., and Matthew K. Fujita. 2011. "Evaluating Phylogenetic Informativeness and Data-Type Usage for New Protein-Coding Genes across Vertebrata." *Molecular Phylogenetics and Evolution*. https://doi.org/10.1016/j.ympev.2011.06.016.

Gee, Henry. 2003. "Evolution: Ending Incongruence." *Nature* 425 (6960): 782.

Goldman, N., and Z. Yang. 1994. "A Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA Sequences." *Molecular Biology and Evolution* 11 (5): 725–36.

Granados Mendoza, Carolina, Stefan Wanke, Karsten Salomo, Paul Goetghebeur, and Marie-Stéphanie Samain. 2013. “Application of the Phylogenetic Informativeness Method to Chloroplast Markers: A Test Case of Closely Related Species in Tribe Hydrangeeae (Hydrangeaceae).” *Molecular Phylogenetics and Evolution* 66 (1): 233–42.

Graybeal, Anna. 1994. “Evaluating the Phylogenetic Utility of Genes: A Search for Genes Informative About Deep Divergences Among Vertebrates.” *Systematic Biology*. https://doi.org/10.2307/2413460.

Heath, Tracy A., Derrick J. Zwickl, Junhyong Kim, and David M. Hillis. 2008. “Taxon Sampling Affects Inferences of Macroevolutionary Processes from Phylogenetic Trees.” *Systematic Biology*. https://doi.org/10.1080/10635150701884640.

Hillis, D. M., and J. P. Huelsenbeck. 1992. “Signal, Noise, and Reliability in Molecular Phylogenetic Analyses.” *The Journal of Heredity* 83 (3): 189–95.

Hughes, A. L., and M. Yeager. 1998. “Comparative Evolutionary Rates of Introns and Exons in Murine Rodents.” *Journal of Molecular Evolution* 46 (4): 497.

Jarvis, Erich D., Siavash Mirarab, Andre J. Aberer, Bo Li, Peter Houde, Cai Li, Simon Y. W. Ho, et al. 2014. “Whole-Genome Analyses Resolve Early Branches in the Tree of Life of Modern Birds.” *Science* 346 (6215): 1320–31.

Kapusta, Aurélie, Alexander Suh, and Cédric Feschotte. 2017. “Dynamics of Genome Size Evolution in Birds and Mammals.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (8): E1460–69.

Katoh, K. 2002. “MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkf436.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. “TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.” *Molecular Biology and Evolution* 34 (7): 1812–19.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Leinonen, Rasko, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, et al. 2011. “The European Nucleotide Archive.” *Nucleic Acids Research* 39 (Database issue): D28–31.

Lewis, Paul O. 2001. “A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data.” *Systematic Biology*. https://doi.org/10.1080/106351501753462876.

Leys, Christophe, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. “Detecting Outliers: Do Not Use Standard Deviation around the Mean, Use Absolute Deviation around the Median.” *Journal of Experimental Social Psychology*. https://doi.org/10.1016/j.jesp.2013.03.013.

Li, W. H., T. Gojobori, and M. Nei. 1981. “Pseudogenes as a Paradigm of Neutral Evolution.” *Nature* 292 (5820): 237–39.

Managadze, David, Igor B. Rogozin, Diana Chernikova, Svetlana A. Shabalina, and Eugene V. Koonin. 2011. “Negative Correlation between Expression Level and Evolutionary Rate of Long Intergenic Noncoding RNAs.” *Genome Biology and Evolution* 3 (November): 1390–1404.

Moeller, Andrew H., and Jeffrey P. Townsend. 2011. “Phylogenetic Informativeness Profiling of 12 Genes for 28 Vertebrate Taxa without Divergence Dates.” *Molecular Phylogenetics and*

Evolution. https://doi.org/10.1016/j.ympev.2011.04.023.
Nosenko, Tetyana, Fabian Schreiber, Maja Adamska, Marcin Adamski, Michael Eitel, Jörg Hammel, Manuel Maldonado, et al. 2013. "Deep Metazoan Phylogeny: When Different Genes Tell Different Stories." *Molecular Phylogenetics and Evolution* 67 (1): 223–33.
Omland, Kevin E. 2003. "Novel Intron Phylogeny Supports Plumage Convergence in Orioles (Icterus)." *The Auk*. https://doi.org/10.2307/4090267.
Paradis, Emmanuel, and Klaus Schliep. 2019. "Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R." *Bioinformatics* 35 (3): 526–28.
Philippe, Hervé, Henner Brinkmann, Dennis V. Lavrov, D. Timothy J. Littlewood, Michael Manuel, Gert Wörheide, and Denis Baurain. 2011. "Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough." *PLoS Biology* 9 (3): e1000602.
Phillips, Matthew J., Frédéric Delsuc, and David Penny. 2004. "Genome-Scale Phylogeny and the Detection of Systematic Biases." *Molecular Biology and Evolution* 21 (7): 1455–58.
Pisani, Davide, Roberto Feuda, Kevin J. Peterson, and Andrew B. Smith. 2012. "Resolving Phylogenetic Signal from Noise When Divergence Is Rapid: A New Look at the Old Problem of Echinoderm Class Relationships." *Molecular Phylogenetics and Evolution*. https://doi.org/10.1016/j.ympev.2011.08.028.
Quinlan, Aaron R. 2014. "BEDTools: The Swiss-Army Tool for Genome Feature Analysis." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 47 (September): 11.12.1–34.
R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
Reddy, Sushma, Rebecca T. Kimball, Akanksha Pandey, Peter A. Hosner, Michael J. Braun, Shannon J. Hackett, Kin-Lan Han, et al. 2017. "Why Do Phylogenomic Data Sets Yield Conflicting Trees? Data Type Influences the Avian Tree of Life More than Taxon Sampling." *Systematic Biology* 66 (5): 857–79.
Reis, Mario Dos, Gregg F. Gunnell, Jose Barba-Montoya, Alex Wilkins, Ziheng Yang, and Anne D. Yoder. 2018. "Using Phylogenomic Data to Explore the Effects of Relaxed Clocks and Calibration Strategies on Divergence Time Estimation: Primates as a Test Case." *Systematic Biology* 67 (4): 594–615.
Ren, Ren, Yazhou Sun, Yue Zhao, David Geiser, Hong Ma, and Xiaofan Zhou. 2016. "Phylogenetic Resolution of Deep Eukaryotic and Fungal Relationships Using Highly Conserved Low-Copy Nuclear Genes." *Genome Biology and Evolution* 8 (9): 2683–2701.
Rokas, Antonis, and Sean B. Carroll. 2008. "Frequent and Widespread Parallel Evolution of Protein Sequences." *Molecular Biology and Evolution* 25 (9): 1943–53.
Rokas, Antonis, Nicole King, John Finnerty, and Sean B. Carroll. 2003. "Conflicting Phylogenetic Signals at the Base of the Metazoan Tree." *Evolution & Development* 5 (4): 346–59.
Russo, C. A., N. Takezaki, and M. Nei. 1996. "Efficiencies of Different Genes and Different Tree-Building Methods in Recovering a Known Vertebrate Phylogeny." *Molecular Biology and Evolution* 13 (3): 525–36.
Salichos, Leonidas, and Antonis Rokas. 2013. "Inferring Ancient Divergences Requires Genes with Strong Phylogenetic Signals." *Nature* 497 (7449): 327–31.
Schwartz, Rachel S., Kelly M. Harkins, Anne C. Stone, and Reed A. Cartwright. 2015. "A Composite Genome Approach to Identify Phylogenetically Informative Data from next-Generation Sequencing." *BMC Bioinformatics* 16 (June): 193.

Schwartz, Rachel S., and Rachel L. Mueller. 2010. “Branch Length Estimation and Divergence Dating: Estimates of Error in Bayesian and Maximum Likelihood Frameworks.” *BMC Evolutionary Biology* 10 (January): 5.

Sharma, Prashant P., Stefan T. Kaluziak, Alicia R. Pérez-Porro, Vanessa L. González, Gustavo Hormiga, Ward C. Wheeler, and Gonzalo Giribet. 2014. “Phylogenomic Interrogation of Arachnida Reveals Systemic Conflicts in Phylogenetic Signal.” *Molecular Biology and Evolution* 31 (11): 2963–84.

Shaw, Joey, Edgar B. Lickey, John T. Beck, Susan B. Farmer, Wusheng Liu, Jermey Miller, Kunsiri C. Siripun, Charles T. Winder, Edward E. Schilling, and Randall L. Small. 2005. “The Tortoise and the Hare II: Relative Utility of 21 Noncoding Chloroplast DNA Sequences for Phylogenetic Analysis.” *American Journal of Botany* 92 (1): 142–66.

Sims, Gregory E., Se-Ran Jun, Guohong Albert Wu, and Sung-Hou Kim. 2009. “Whole-Genome Phylogeny of Mammals: Evolutionary Information in Genic and Nongenic Regions.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (40): 17077–82.

Small, Randall L., Julie A. Ryburn, Richard C. Cronn, Tosak Seelanan, and Jonathan F. Wendel. 1998. “The Tortoise and the Hare: Choosing between Noncoding Plastome and nuclearAdhsequences for Phylogeny Reconstruction in a Recently Diverged Plant Group.” *American Journal of Botany*. https://doi.org/10.2307/2446640.

Song, Sen, Liang Liu, Scott V. Edwards, and Shaoyuan Wu. 2012. “Resolving Conflict in Eutherian Mammal Phylogeny Using Phylogenomics and the Multispecies Coalescent Model.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (37): 14942–47.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13.

Steel, Mike, and Christoph Leuenberger. 2017. “The Optimal Rate for Resolving a near-Polytomy in a Phylogeny.” *Journal of Theoretical Biology*. https://doi.org/10.1016/j.jtbi.2017.02.037.

Steppan, Scott J., and John J. Schenk. 2017. “Muroid Rodent Phylogenetics: 900-Species Tree Reveals Increasing Diversification Rates.” *PloS One* 12 (8): e0183070.

Su, Zhuo, and Jeffrey P. Townsend. 2015. “Utility of Characters Evolving at Diverse Rates of Evolution to Resolve Quartet Trees with Unequal Branch Lengths: Analytical Predictions of Long-Branch Effects.” *BMC Evolutionary Biology* 15 (May): 86.

Townsend, Jeffrey P. 2007. “Profiling Phylogenetic Informativeness.” *Systematic Biology* 56 (2): 222–31.

Townsend, Jeffrey P., and Christoph Leuenberger. 2011. “Taxon Sampling and the Optimal Rates of Evolution for Phylogenetic Inference.” *Systematic Biology* 60 (3): 358–65.

Townsend, Jeffrey P., Francesc López-Giráldez, and Robert Friedman. 2008. “The Phylogenetic Informativeness of Nucleotide and Amino Acid Sequences for Reconstructing the Vertebrate Tree.” *Journal of Molecular Evolution* 67 (5): 437–47.

Townsend, Jeffrey P., Zhuo Su, and Yonas I. Tekle. 2012. “Phylogenetic Signal and Noise: Predicting the Power of a Data Set to Resolve Phylogeny.” *Systematic Biology* 61 (5): 835–49.

Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2018. “Ensembl 2018.” *Nucleic Acids Research* 46 (D1): D754–61.

Zurano, Juan P., Felipe M. Magalhães, Ana E. Asato, Gabriel Silva, Claudio J. Bidau, Daniel O. Mesquita, and Gabriel C. Costa. 2019. "Cetartiodactyla: Updating a Time-Calibrated Molecular Phylogeny." *Molecular Phylogenetics and Evolution* 133 (April): 256–62.

R Core Team. 2013. "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.