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 evolution, yet many relationships remain controversial despite whole-genome sequence data. These controversies are due in part to complex patterns of phylogenetic and non-phylogenetic signal coming from regions of the genome experiencing distinct evolutionary forces, which can be difficult to disentangle. Here we profile the amounts and proportions of phylogenetic and non-phylogenetic signal derived from loci spread across mammalian genomes. We identified orthologous sequences from primates, rodents, and pecora, annotated sites as one or more of nine locus types (e.g. coding, intronic, intergenic), and profiled the phylogenetic information contained within locus types across evolutionary timescales associated with each clade. In all cases, non-coding loci provided more overall signal and a higher proportion of phylogenetic signal compared to coding loci. This suggests potential benefits if perspectives are altered away from primarily targeting genes or coding regions for phylogenetic studies, particularly in this era of accessible whole genome sequence data. In contrast to long-held assumptions about the phylogenetic utility of more variable genomic regions, most locus types provided relatively consistent phylogenetic information across timescales, although we find evidence that coding and intronic regions may, respectively and to a limited degree, inform disproportionately about older and younger splits. As part of this work we also validate the SISRS pipeline as an annotation-free ortholog discovery pipeline capable of identifying millions of phylogenetically informative sites directly from raw 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). Molecular phylogenetics utilizes

sequence variation among orthologous loci to infer historical diversification events. When comparing orthologs among species, sites may be said to carry 'phylogenetic signal' when the sequence variation reflects a genetic synapomorphy that 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, but rather some other phenomenon, as in cases of convergent evolution, homoplasy, or incomplete lineage sorting (Song et al. 2012; Rokas and Carroll 2008; Li, Gojobori, and Nei 1981). As a means to increase the amount and diversity of available phylogenetic signal, more accessible next-generation sequencing (NGS) technology is facilitating a discipline-wide shift away from resolving species trees using small sets of typically genic 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, due to the fact that non-phylogenetic signal is scaled up in parallel with phylogenetic signal as datasets expand, many species and clade relationships remain unresolved despite genome-scale efforts (Cao et al. 1994; Hillis and Huelsenbeck 1992; Reddy et al. 2017; Philippe et al. 2011). As the use of genome-scale data transitions from cutting edge to commonplace, our inability to easily scale up phylogenetic signal without carrying along non-phylogenetic signal hinders our understanding of broader evolutionary processes and necessitates closer inspection (Reddy et al. 2017; Brown and Thomson 2017; Dornburg, Su, and Townsend 2019).

For decades, attempts to resolve more recent splits (e.g. those within a genus or family) have utilized sequence data from stereotypically fast-evolving loci such as intronic or intergenic regions, as these loci are thought to accrue substitutions rapidly enough to generate sufficient phylogenetic signal at shorter timescales (Debry and Seshadri 2001; Omland 2003; Baldwin and Markos 1998; Shaw et al. 2005). Conversely, work to resolve deep-time splits has typically used more slowly-evolving protein-coding sequences (CDS) or other more constrained genomic subsets as a way to limit the impact of overwriting mutations (i.e. 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, the relationship between rate and phylogenetic utility is often complicated by interacting factors including complex patterns and constraints in character evolution, tree structure, and taxon sampling (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). Even at genomic scales, phylogenetic tree inference using data derived from loci experiencing distinct evolutionary forces continues to generate topologies that are incompatible, yet often well-supported (Jarvis et al. 2014; Sharma et al. 2014; Rokas et al. 2003; Nosenko et al. 2013; Kumar et al. 2012). When species relationships are unknown, conflicting trees inferred from different locus types are difficult to reconcile (Reddy et al. 2017). However, when a robust underlying tree has been resolved for a set of species, interpreting how sequence variation informs (or misinforms) about splits in that tree provides an opportunity to explicitly contrast the relative amounts of phylogenetic and non-phylogenetic signal coming from different regions of the genome. These results may be used to help disentangle topological incongruence in unresolved clades by evaluating the source of support for different hypotheses (Graybeal 1994; Russo, Takezaki, and Nei 1996; Townsend 2007).

Studies utilizing locus-wide substitution rates as the primary predictor of informativeness support the idea that some fast-evolving loci may provide more information for more recent splits while slowly-evolving loci can provide more support for ancient speciation events (Townsend, López-Giráldez, and Friedman 2008; Fong and Fujita 2011). While these locus-scale methods have been successful at resolving problematic phylogenetic questions (Dornburg et al. 2014; Granados Mendoza et al. 2013), whether a locus has a high average informativeness value does not always correlate with ultimate phylogenetic utility (Klopfstein, Kropf, and Quicke 2010; Aguileta et al. 2008; Moeller and Townsend 2011). Recent studies suggest that informativeness profiling can be improved when considering individual site rate variation within a locus, rather than averaging rates across sites (Klopfstein, Massingham, and Goldman 2017; Dornburg, Su, and Townsend 2019). Whether informativeness profiling is done at the locus- or site-scale, methods that rely on explicit rate estimates also depend on

accurate ortholog identification, alignment, and model specification (Doyle et al. 2015; Klopfstein, Massingham, and Goldman 2017). As such, informativeness profiles have historically been performed on modest numbers of loci that have already been well-characterized and annotated. This strategy prioritizes the characterization of known loci over the discovery of those that may in fact be more useful. Based in part on locus length and variability, individual loci also contain limited amounts of phylogenetic signal (Shen, Hittinger, and Rokas 2017; Klopfstein, Massingham, and Goldman 2017); this too restricts the types of questions and broader extrapolations that can be drawn from small locus sets.

Understanding the distribution of phylogenetic and non-phylogenetic signal across the genome relies on the ability to identify and compare orthologs across locus types, but acquiring these data can be especially challenging in clades lacking high-level genomic resources. The relative ease of acquiring, annotating, and aligning CDS data (e.g. through transcriptome sequencing or exome capture) has led to a majority of quantitative studies of informativeness focusing on comparisons among CDS types or other historically popular markers (Fong and Fujita 2011; Russo, Takezaki, and Nei 1996; Townsend, López-Giráldez, and Friedman 2008; Moeller and Townsend 2011; Graybeal 1994). However, explicit phylogenetic utility 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). Analyses of ultraconserved elements have found that net informativeness of more variable flanking regions can rival that of protein-coding gene sequences (Faircloth et al. 2012; Gilbert et al. 2015), while non-coding introns have been shown to provide a more robust and consistent phylogenetic signal relative to coding loci (Chen, Liang, and Zhang 2017). These results provide a tantalizing glimpse into the genome-wide informativeness landscape; yet, many clades lack the high-level genomic resources required to generate such comparisons, restricting the scope of most studies to a select handful of loci containing limited signal.

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 ortholog identification 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 using site-rates derived without the need for traditional sequence alignment and associated substitution model correction. 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 to consider the phylogenetic value of these information-rich genomic subsets, which have received historically limited focus. While the majority of sites from this study carry phylogenetic signal, CDS-derived sites 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. Surprisingly, we find only limited support for the differential phylogenetic utility of fast- and slow-evolving locus types to resolve nodes of different ages, at least at the evolutionary timescale of this study. Thus, here we leverage our ability to identify millions of informative sites across the genome to recontextualize our understanding of phylogenetic utility among different genomic subsets. Looking forward, we predict that such genome-wide utility profiling will play an important role in dramatically expanding the pool of potential markers available to tackle recalcitrant issues in phylogenetics.

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 phylogenetic utility of genomic loci relies on having available data for species that are related under a robustly-supported underlying evolutionary hypothesis. To that end, we identified three mammalian clades with well-established evolutionary relationships (Figure 1) and sufficient whole-genome sequencing (WGS) data: catarrhine primates, murid rodents, and members of the infraorder Pecora (Reis et al. 2018; Zurano et al. 2019; Steppan and Schenk 2017). For each clade, we obtained paired-end Illumina reads from the European Nucleotide

Archive (Leinonen et al. 2011) for ten focal taxa and two outgroup taxa. 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 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 was 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 lncRNA 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 using parsimony (i.e. without evolutionary model correction). 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 signal support by 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 2, 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 2, 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 (all $p < 2.3E^{-7}$, Table S7).

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 2, 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 2, 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 2, Table 1, Table S8).

Most genomic subsets show no change in phylogenetic utility over focal timescales

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 - 46.7 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 1.4% - 2.3% per million years in the Pecora, Rodent, and Combined datasets ($p = 2.46E^{-3}$, 3.48^{-3},$2.00E^{-3}$ 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 0.24% per million years, respectively ($p = 6.04E^{-4}$, $2.00E^{-3}$, Figure 3, Table S10).

Discussion

The increasing availability of publicly archived genomic data and lower costs associated with high-throughput sequencing are driving the field of phylogenetics towards genome-scale analyses. However, due in part to conflicting signal, these scaled-up datasets have not yet facilitated the resolution of many controversial relationships (Jarvis et al. 2014; Sharma et al. 2014; Rokas et al. 2003; Nosenko et al. 2013). Here we provide a genome-scale perspective on how phylogenetic and non-phylogenetic signal is distributed across a diverse set of locus types in three mammal clades, and how subset-specific signal informs about splits over time. While pooling and analyzing sites within locus types restricts what we can infer about the signal contained in particular locus, it importantly yields datasets that are large enough to enable robust scrutinization of long-held beliefs about the phylogenetic utility of different genomic subsets. Even when categorized very broadly, clear trends emerge that suggest certain subsets of the data may be prone to carrying a higher proportion of non-phylogenetic signal that mislead our understanding of species diversification, and these results provide targets for more focused analyses.

<u>Non-coding loci carry summatively and proportionally more phylogenetic signal</u>

Coding sequences (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. However, non-coding sites provided more total phylogenetic signal and a higher proportion of phylogenetic signal relative

to CDS-derived sites in all datasets. Specifically, of the 2.4M signal-bearing sites identified in the rodent dataset, 92% were annotated as either intronic, lncRNA, or intergenic (i.e. unannotated), and this ratio rose to 98% in the larger primate and pecora datasets. Notably, given that the cow genome lacked lncRNA annotations, this high ratio in pecora suggests that the SISRS annotation-free ortholog discovery pipeline was likely able to assemble and query lncRNA loci, but that we lacked the reference information to properly identify them; hence, they were rolled into the intergenic fraction.

Given their relative abundances in the genome, that non-coding sites vastly outnumber coding sites in any truly genome-scale dataset is hardly surprising; however, CDS also 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). While significantly enriched for non-phylogenetic signal in all three focal clades, CDS carried the highest proportion of non-phylogenetic signal in the rodent dataset where nearly one in every three CDS sites provided support for a non-phylogenetic split. This trend was exaggerated in the combined dataset where close to every other CDS site carried non-phylogenetic signal, thus providing a quantitative demonstration as to why indiscriminately scaling up the size of datasets, especially CDS-biased datasets, is unlikely to be a cure-all towards resolving phylogenetic incongruence. Additionally, CDS in both the rodent and combined datasets contained higher proportions of sites evolving at a rate determined by SISRS to be putatively informative, providing further support to the idea that rate alone is likely a poor predictor of ultimate phylogenetic performance (Aguileta et al. 2008; Klopfstein, Kropf, and Quicke 2010).

It should be noted that of the low concordance rates among CDS sites may be explained in part by the fact that we were unable to distinguish between the first, second, and third codon positions in this analysis. Third codon positions typically have a higher relative substitution rate, making them prone to homoplasy and a common target of phylogenetic site filtration (Breinholt and Kawahara 2013), although the ultimate impact of such filtration is contentious (Simmons et al. 2006). Nevertheless, it should be noted that SISRS removes sites

that vary within a species as well as those that vary highly among species (i.e. more than two alleles), and as such, these types of sites likely had little impact on the final results.

Phylogenetic informativeness of most locus types is stable across evolutionary timescales

In contrast to expectations of how markers from different genomic subsets are thought to be 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 50 million years of evolution associated with these taxa. CDS sites provided disproportionate support to older nodes in the pecora, rodent, and the combined datasets, and this trend 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). However, recall that these sites also carry the highest proportions of non-phylogenetic signal of all analyzed loci, and that this trend appears to be exacerbated in older, deeper trees; thus, this will affect tree inference negatively. Notably, the genomic sites expected to mutate at the most unconstrained rates (i.e. those within intergenic or pseudogenic regions) showed no significant deterioration in signal over evolutionary time in any dataset. When they can be identified and compared, these results show that these often overlooked portions of the genome may hold valuable phylogenetic signal across evolutionary timescales.

With respect to the focal primates, rodents, and pecora, the lineage diversification timescales were similar, with each group comprising around 20 - 25MY of mammal evolution. None of the trends in age-dependent phylogenetic utility held across all three mammal groups, which may in part due to difference in generation times amongst clades (Sims et al. 2009). Rodents, and to a lesser extent pecora, will go through more generations over the same evolutionary timescale when compared to most primates, and each ‘extra’ generation provides more opportunities for new phylogenetic signal to be written, and for existing phylogenetic signal to be overwritten. With mutations being passed down more quickly in rodents and pecora, it is not surprising that regions of the genome less prone to mutation (i.e. functionally-constrained CDS) may provide a higher proportion of signal for older nodes in the

tree, as these sites would be the most immune to homoplasy. The trend of rodent introns to provide more signal for younger nodes was not seen in pecora, and in primates, neither CDS or intronic sites displayed a significant change in relative utility over time. This suggests that in lieu of a broad, 'one size fits all' approach regarding which locus types will provide time-sensitive phylogenetic signal, that those decisions should be based on an understanding of other biological factors which may affect how those loci evolve among focal species.

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, and this limits the utility of reference-guided (i.e. alignment- or mapping-based) phylogenetic marker discovery pipelines. SISRS bypasses the need for reference annotation data by identifying orthologs directly from WGS reads. In fact, millions of the signal-bearing sites identified in this study derived from completely unannotated genomic regions that would have been excluded from any reference-based ortholog discovery pipeline. Additionally, the SISRS composite genome assembly process requires only a small fraction of per-taxon sequence data as would be required in the *de novo* development of a UCE-type dataset, making it a tractable and affordable alternative to UCEs that can be performed using previously published WGS data, as was the case for this study.

Site-based informativeness profiling has been shown to outperform locus-scale methods, and it uses the substitution rate for each site in a multiple-sequence alignment to describe the potential informativeness of different loci (Klopfstein, Massingham, and Goldman 2017; Dornburg, Su, and Townsend 2019). In much the same way, SISRS too filters sites based on their pattern and tempo of sequence variation; however, it does so without the need for *a priori* ortholog identification, sequence alignment, established taxonomic relationships, or model-based rate calculations. The SISRS filtration process is not one that actively selects sites based on a rate, but rather, it excludes sites that would be unsuitable for reliable topological inference (i.e. poorly-supported, invariant, or hypervariable sites). By not considering composite genome sites that appear to be evolving too slowly or quickly to be reliably informative for resolving splits in the focal clade, SISRS analyzes only those sites that appear to

have mutated once during the evolutionary history in question. Among the focal clades, this process resulted in datasets ranging from 3M - 11M sites, and only 10% - 25% of sites in each dataset carried non-phylogenetic signal.

With 300K sites identified and 32% of those providing non-phylogenetic signal, the combined dataset that included all 36 species yielded the least total data and the lowest overall concordance rate among datasets. While this 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. Research is ongoing to determine whether the composite genome assembly process can be adjusted to account for higher amounts of sequence variation among focal taxa, which would be beneficial if SISRS was to be used to robustly resolve deep-time relationships.

Conclusions

In this study, we provide a perspective on the phylogenetic utility of nine different locus types as they apply to resolving species relationships among three mammal clades. Non-coding sites provided higher proportions and amounts of phylogenetic signal compared to CDS sites in all datasets. This suggests potential benefits if perspectives are altered away from primarily targeting genes or coding regions for phylogenetic studies, particularly in this era of accessible whole genome sequence data. 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 varied in their utility and that these patterns were both subtle and clade-specific. As these results appear to fly in the face of long-held beliefs about the phylogenetic utility of different genomic subsets, we hope our findings will provide the motivation to expand the search for phylogenetic markers that will be needed to resolve some of the more recalcitrant relationships in evolutionary biology. Additionally, we validate SISRS as a mechanism to easily identify millions of sites across the genome containing valid phylogenetic signal.

Figure 1

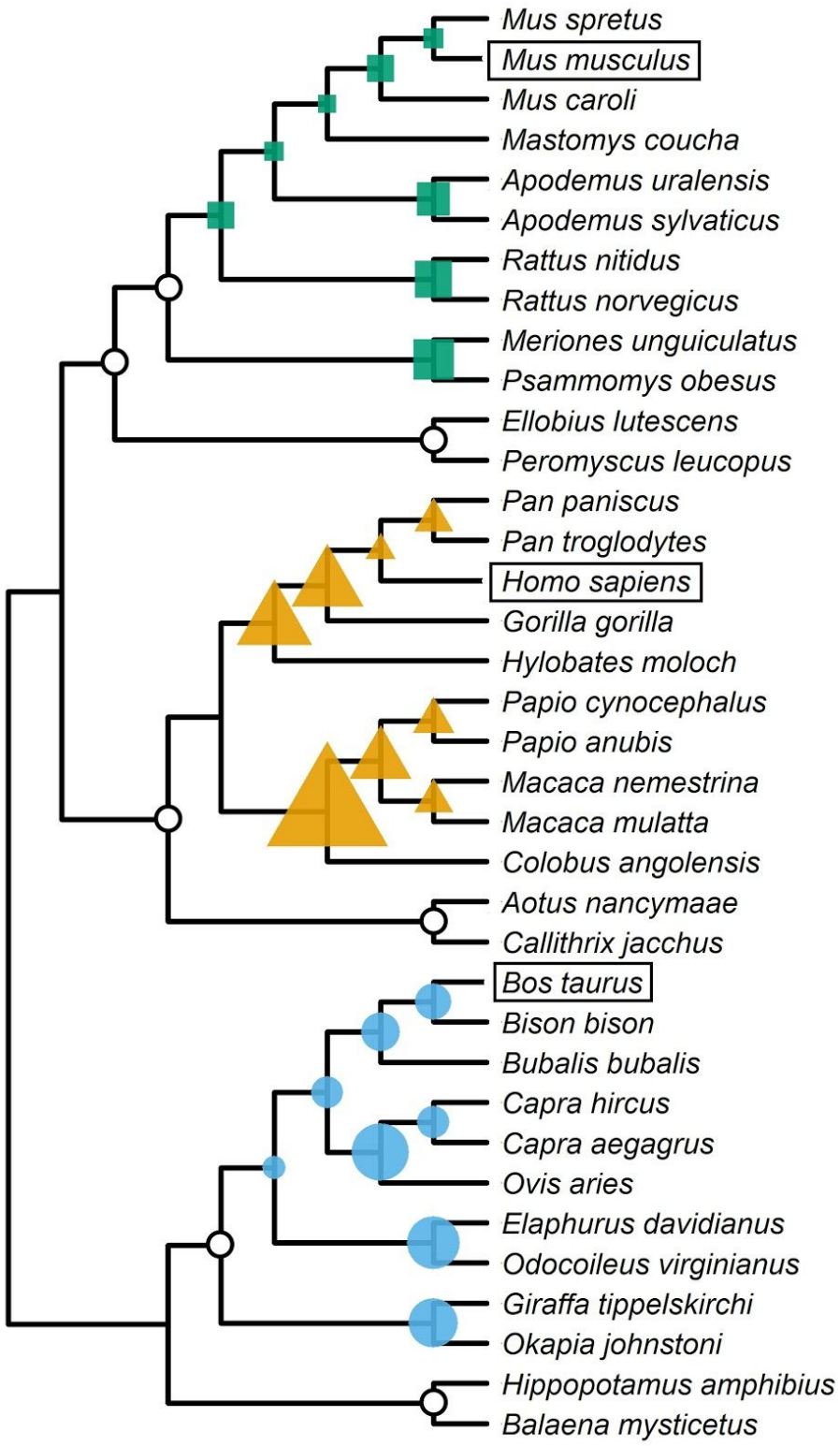

Mus spretus
Mus musculus
Mus caroli
Mastomys coucha
Apodemus uralensis
Apodemus sylvaticus
Rattus nitidus
Rattus norvegicus
Meriones unguiculatus
Psammomys obesus
Ellobius lutescens
Peromyscus leucopus
Pan paniscus
Pan troglodytes
Homo sapiens
Gorilla gorilla
Hylobates moloch
Papio cynocephalus
Papio 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

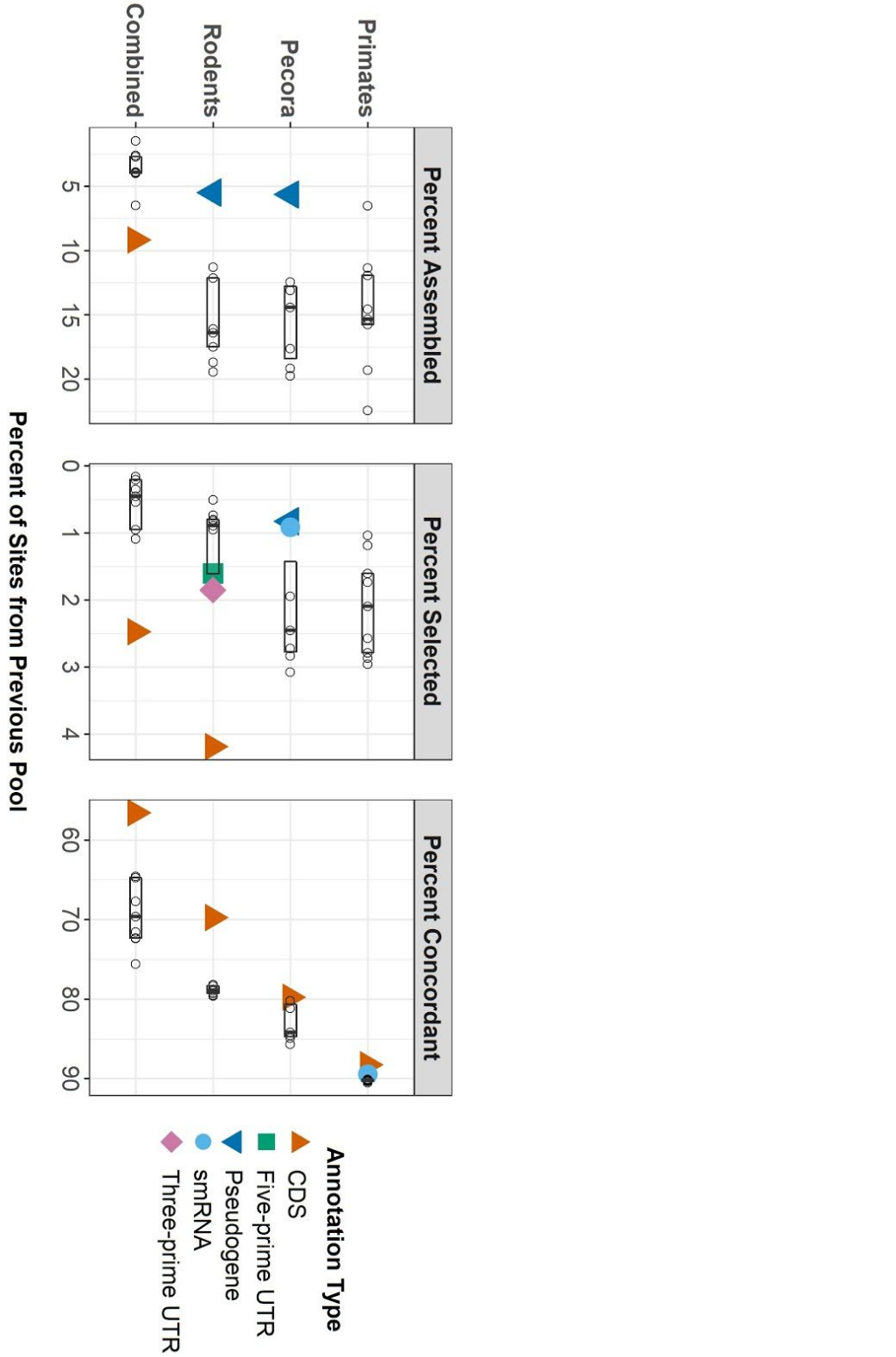

Percent Assembled
Percent Selected
Percent Concordant
Primates
Pecora
Rodents
Combined
5
10
15
20
0
1
2
3
4
60
70
80
90
Percent of Sites from Previous Pool
Annotation Type
CDS
Five-prime UTR
Pseudogene
smRNA
Three-prime UTR

Figure 3

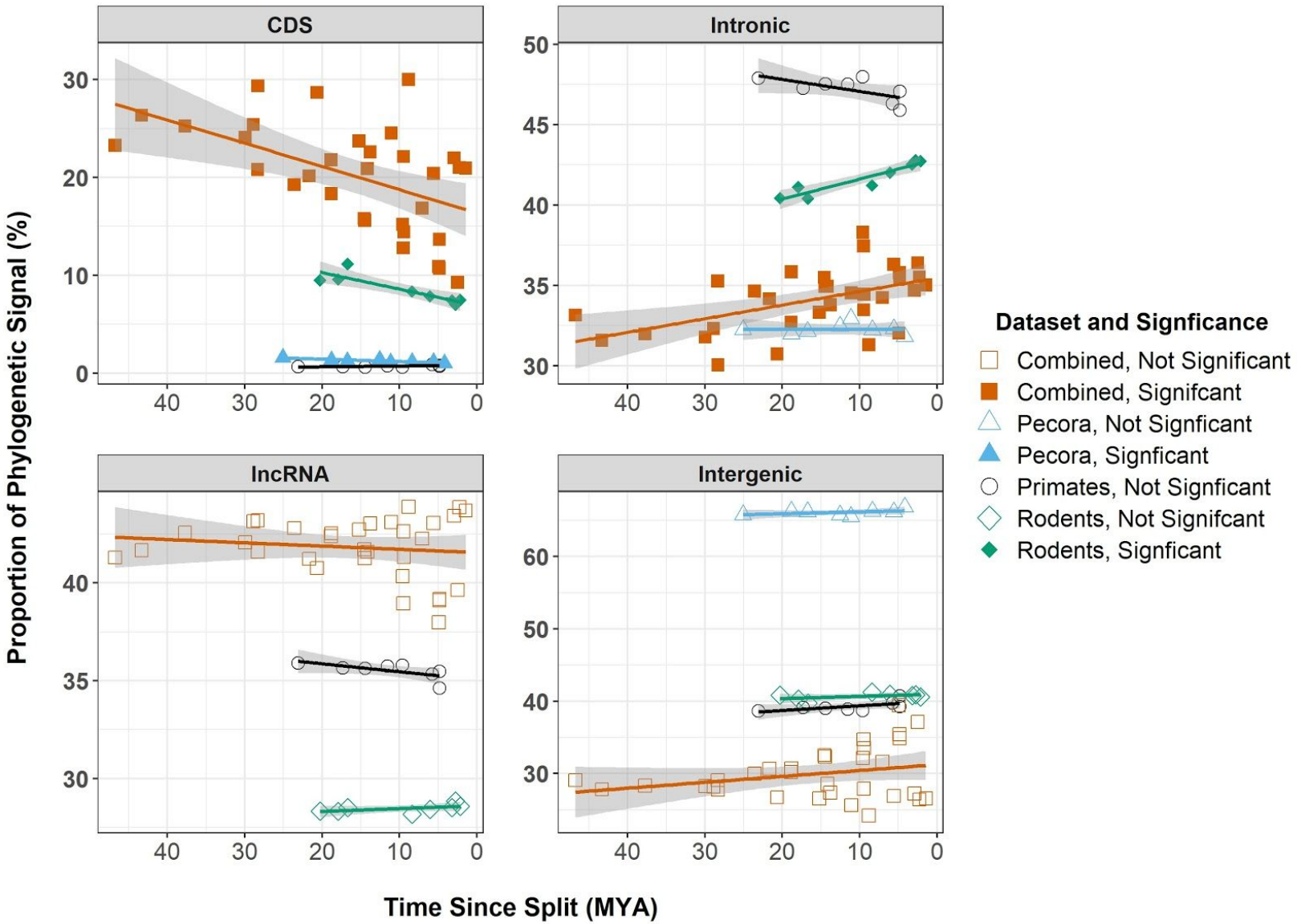

CDS
Intronic
lncRNA
Intergenic
Proportion of Phylogenetic Signal (%)
Time Since Split (MYA)
Dataset and Signficance
Combined, Not Significant
Combined, Signifcant
Pecora, Not Signficant
Pecora, Signficant
Primates, Not Signficant
Rodents, Not Signifcant
Rodents, Signficant

Figure 1: Evolutionary relationships among study species. The size of filled node icons is scaled proportionally to the number of SISRS sites supporting that split. Split support ranges for each focal group were as follows: Pecora (green squares): 172K - 1.04M sites; Primates (orange triangles): 148K - 1.87M; Rodents (blue circles): 27.4K - 600K. Open circles denote nodes included in the combined analysis that were excluded from focal analyses, and are not scaled to support size (Combined support range: 487 - 34.0K sites). Reference annotation species are boxed.

Figure 2: Proportion of sites passing through various steps in the SISRS pipeline. Panel 1: Based on all annotated sites in the reference genome, the percent of sites assembled into the composite genome. Panel 2: The percent of composite genome sites that were selected by SISRS as having a putatively informative substitution rate. Panel 3: The percent of SISRS sites carrying 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, even though CDS sites were highly selected by SISRS based on their substitution rates in rodents and in the combined dataset, CDS sites carried the lowest proportion of phylogenetic signal in all datasets.

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. lncRNA sites are likely rolled into the intergenic fraction of pecora, as the reference genome lacks lncRNA annotation.

Tables

	Annotated SISRS Sites				Percent Concordant Sites			
Annotation	**Pecora**	**Primates**	**Rodents**	**Combined**	**Pecora**	**Primates**	**Rodents**	**Combined**
CDS	156K	94.3K	**285K (+)**	**80.6K (+)**	**79.7 % (-)**	**88.2 % (-)**	**69.7 % (-)**	**56.6 % (-)**
Five-Prime UTR	5.32K	24.2K	**19.2K (+)**	4.40K	85.6 %	90.5 %	79.2 %	64.6 %
Intergenic	6.63M	4.47M	1.26M	80.4K	84.6 %	90.0 %	79.5 %	72.3 %
Intronic	3.25M	5.48M	1.29M	96.3K	84.1 %	90.1 %	79.2 %	71.5 %
Long-noncoding RNA	-	4.12M	886K	128K	-	90.2 %	78.3 %	67.7 %
Noncoding Genes	-	18.7K	14.1K	763	-	90.4 %	78.7 %	72.3 %
Pseudogenes	**318 (-)**	64.9K	7.48K	1.33K	80.2 %	90.3 %	78.9 %	69.6 %
small RNAs	**508 (-)**	1.17K	690	139	81.1 %	**89.4 % (-)**	79.6 %	75.5 %
Three-Prime UTR	38.8K	177K	**129K (+)**	26.9K	84.9 %	90.3 %	78.1 %	64.72

Table 1: Annotation breakdown of sites selected by SISRS as having a potentially informative substitution rate and the percent carrying phylogenetic signal. Bold cells indicate annotation-specific deviations from the data-wide median, and whether the subset overperforms (+) or underperforms (-) is indicated in parentheses. 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.

Breinholt, Jesse W., and Akito Y. Kawahara. 2013. "Phylotranscriptomics: Saturated Third Codon Positions Radically Influence the Estimation of Trees Based on next-Gen Data." *Genome Biology and Evolution* 5 (11): 2082–92.

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.

Brown, Jeremy M., and Robert C. Thomson. 2017. "Bayes Factors Unmask Highly Variable Information Content, Bias, and Extreme Influence in Phylogenomic Analyses." *Systematic Biology* 66 (4): 517–30.

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.

Doyle, Vinson P., Randee E. Young, Gavin J. P. Naylor, and Jeremy M. Brown. 2015. "Can We Identify Genes with Increased Phylogenetic Reliability?" *Systematic Biology* 64 (5): 824–37.

Faircloth, Brant C., John E. McCormack, Nicholas G. Crawford, Michael G. Harvey, Robb T. Brumfield, and Travis C. Glenn. 2012. "Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales." *Systematic Biology* 61 (5): 717–26.

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.

Gilbert, Princess S., Jonathan Chang, Calvin Pan, Eric M. Sobel, Janet S. Sinsheimer, Brant C. Faircloth, and Michael E. Alfaro. 2015. "Genome-Wide Ultraconserved Elements Exhibit Higher Phylogenetic Informativeness than Traditional Gene Markers in Percomorph Fishes." *Molecular Phylogenetics and Evolution* 92 (November): 140–46.

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.

Klopfstein, Seraina, Christian Kropf, and Donald L. J. Quicke. 2010. "An Evaluation of Phylogenetic Informativeness Profiles and the Molecular Phylogeny of Diplazontinae (Hymenoptera, Ichneumonidae)." *Systematic Biology* 59 (2): 226–41.

Klopfstein, Seraina, Tim Massingham, and Nick Goldman. 2017. "More on the Best Evolutionary

Rate for Phylogenetic Analysis." *Systematic Biology* 66 (5): 769–85.
Kumar, Sudhir, Alan J. Filipski, Fabia U. Battistuzzi, Sergei L. Kosakovsky Pond, and Koichiro Tamura. 2012. "Statistics and Truth in Phylogenomics." *Molecular Biology and Evolution* 29 (2): 457–72.
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.

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.

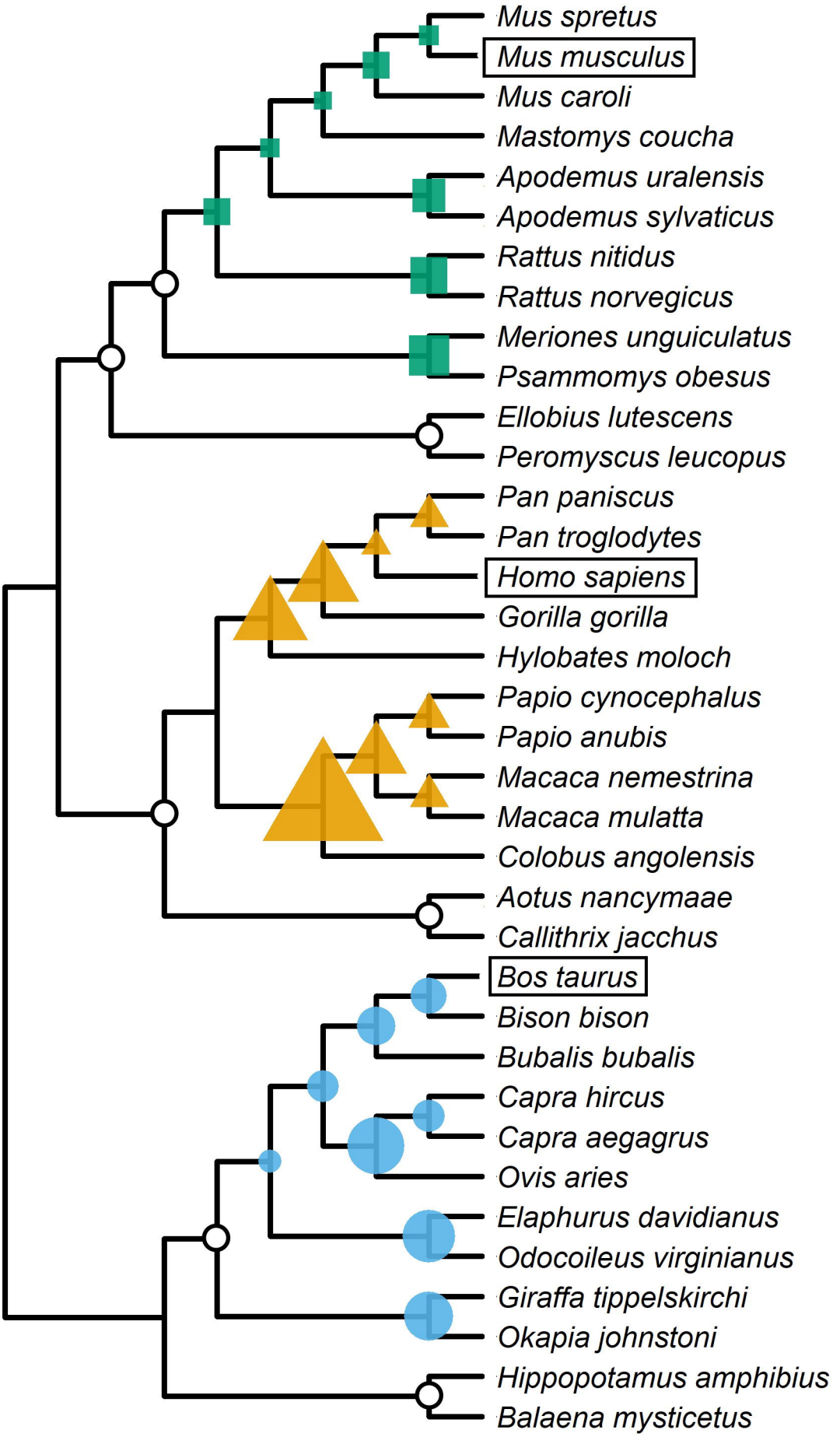
Shen, Xing-Xing, Chris Todd Hittinger, and Antonis Rokas. 2017. "Contentious Relationships in Phylogenomic Studies Can Be Driven by a Handful of Genes." *Nature Ecology & Evolution* 1 (5): 126.

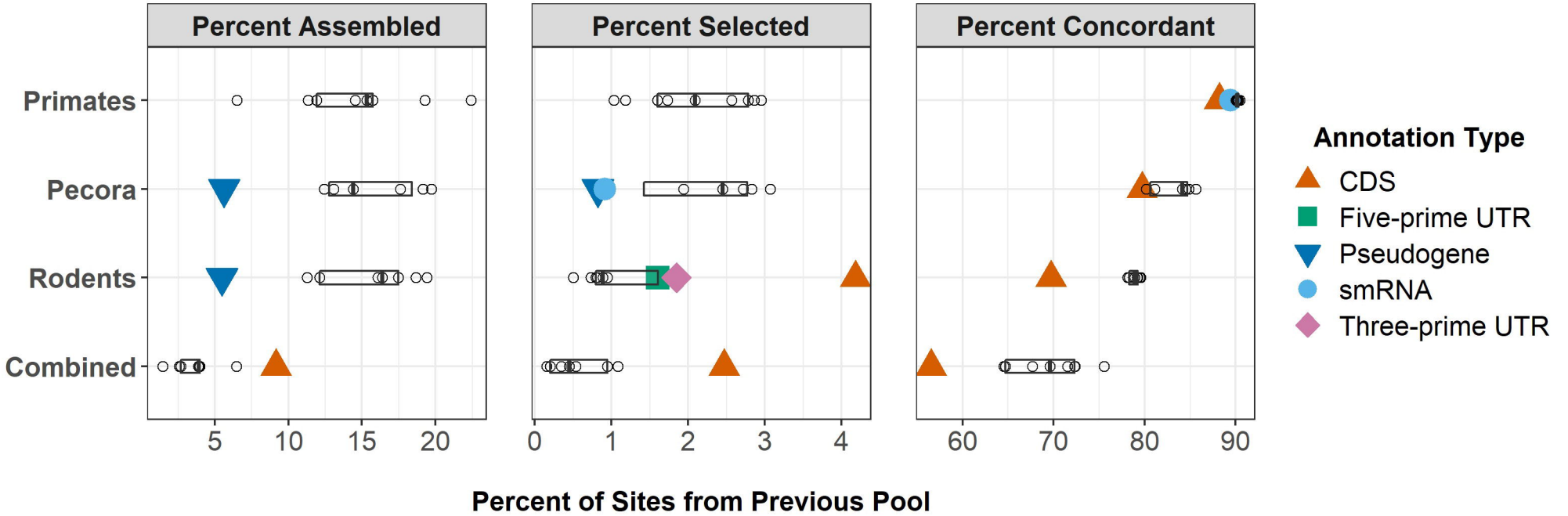
Simmons, Mark P., Li-Bing Zhang, Colleen T. Webb, and Aaron Reeves. 2006. "How Can Third Codon Positions Outperform First and Second Codon Positions in Phylogenetic Inference? An Empirical Example from the Seed Plants." *Systematic Biology* 55 (2): 245–58.

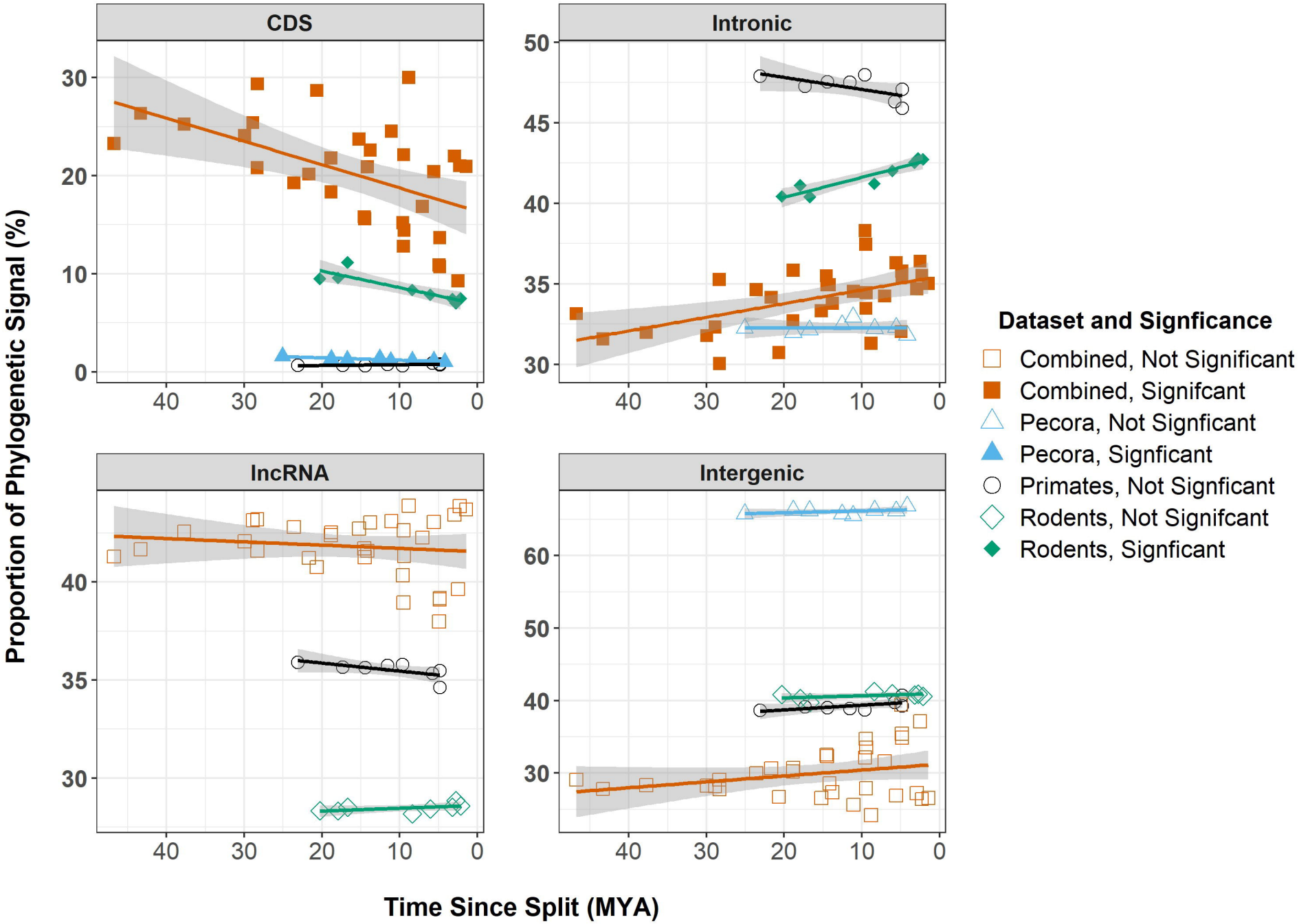
Sims, Gregory E., Se-Ran Jun, Guohong Albert Wu, and Sung-Hou Kim. 2009. "Whole-Genome Phylogeny of Mammals: Evolutionary Information in Genic and Nongenic Regions." *Proceedings of the National Academy of Sciences of the United States of America* 106 (40):

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/.


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


Percent Assembled
Percent Selected
Percent Concordant
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


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