2-2021

# Genome-scale Profiling Reveals Noncoding Loci Carry Higher Proportions of Concordant Data

Robert Literman
*University of Rhode Island*, literman@uri.edu

Rachel Schwartz
*University of Rhode Island*, rsschwartz@uri.edu



## Citation/Publisher Attribution

Robert Literman, Rachel Schwartz, Genome-scale profiling reveals noncoding loci carry higher proportions of concordant data, *Molecular Biology and Evolution*, 2021;, msab026, https://doi.org/10.1093/molbev/msab026

**ARTICLE – DISCOVERIES (REVISED MBE-20-0898)**

**Title:** Genome-scale profiling reveals noncoding loci carry higher proportions of concordant data

**Authors:**

Dr. Robert Literman **(Corresponding Author)**
E-Mail: literman@uri.edu
University of Rhode Island (Site of all research)
Current Affiliation: ORISE Fellow – FDA (CFSAN), USA

Dr. Rachel Schwartz
Current Affiliation: University of Rhode Island, USA

**Abstract**

Many evolutionary relationships remain controversial despite whole-genome sequencing data. These controversies arise in part due to challenges associated with accurately modeling the complex phylogenetic signal coming from genomic regions experiencing distinct evolutionary forces. Here we examine how different regions of the genome support or contradict well-established hypotheses among three mammal groups using millions of orthologous parsimony-informative biallelic sites [PIBS] distributed across primate, rodent, and Pecora genomes. We compared PIBS concordance percentages among locus types (e.g. coding sequences, introns, intergenic regions), and contrasted PIBS utility over evolutionary timescales. Sites derived from noncoding sequences provided more data and proportionally more concordant sites compared with those from coding sequences [CDS] in all clades. CDS PIBS were also predominant drivers of tree incongruence in two cases of topological conflict. PIBS derived from most locus types provided surprisingly consistent support for splitting events spread across the timescales we examined, although we find evidence that CDS and intronic PIBS may, respectively and to a limited degree, inform disproportionately about older and younger splits. In this era of accessible whole genome sequence data, these results (1) suggest benefits to more intentionally focusing on noncoding loci as robust data for tree inference, and (2) reinforce the importance of accurate modeling, especially when using CDS data.

**Introduction**

Molecular systematics relies on genetic variation to infer the history of splitting events leading to contemporary patterns of diversity (e.g. ‘speciation events’ in the case of inferring species trees). The phylogenetic interpretation of some variation is straightforward: sites that mutate once over the history of a clade and fix in accordance with a split provide unambiguous phylogenetic signal. However, the evolutionary history of sites is often far more complex and avoiding error in phylogenetic inference can require complex modeling to account for factors such as substitution-rate biases and evolutionary processes like incomplete-lineage sorting [ILS] (Bleidorn, 2017; Rokas & Carroll, 2008; Song, Liu, Edwards, & Wu, 2012). The relative impact of these factors can be exacerbated in smaller datasets where variation is limited (Cao, Adachi, Janke, Pääbo, & Hasegawa, 1994), so it is not surprising that there was early optimism that increasing the size of datasets would lead to the swift resolution of some of the most challenging questions in systematics (Gee, 2003).

However, because genome-scale data derive from an exponentially larger sample of loci with substitutions governed by a broader spectrum of evolutionary forces, adequately

parameterizing evolutionary models presents both intellectual and computational challenges (Philippe et al., 2011; Reddy et al., 2017; Rodriguez-Ezpeleta et al., 2007). Critically, poor model fit can severely restrict the phylogenetic reliability of a dataset (Jeremy M Brown, 2014; Doyle, Young, Naylor, & Brown, 2015), and inferring trees using large numbers of loci experiencing disparate forces (e.g. genome-scale coding sequence [CDS], intron, and intergenic datasets) results in the generation of incompatible phylogenies (Jarvis et al., 2014; Nosenko et al., 2013; Rokas, King, Finnerty, & Carroll, 2003; Sharma et al., 2014). Traditional methods of assessing split support (e.g. phylogenetic bootstrapping) become artificially inflated as datasets expand; thus, the resulting phylogenies may all appear to be well-supported (Kumar, Filipski, Battistuzzi, Kosakovsky Pond, & Tamura, 2012; Salichos & Rokas, 2013). The number of evolutionary relationships that remain unresolved in the face of whole-genome data suggests the need to examine the phylogenetic reliability of different subsets of genomic data; in addition to providing valuable context for interpreting phylogenetic discordance among data subsets, these types of analyses can also identify data partitions where accurate phylogenetic interpretation is more robust to model misspecification.

Information in phylogenetic data

Ultimately, the phylogenetic utility of a dataset (i.e. how broadly, deeply, and reliably it informs on queried relationships, if appropriately modeled) depends on (1) the rates and timescales associated with focal clade diversification, and (2) the amount and proportion of sites evolving under a range of substitution rates commensurate with the generation and maintenance of relevant phylogenetic information (Dornburg, Su, & Townsend, 2019; Doyle et al., 2015; Graybeal, 1994; Townsend, 2007). The resolution of relatively recent splitting events requires sites that have experienced substitutions recently enough to have generated sufficient phylogenetic signal. In contrast, the accurate reconstruction of relatively older bifurcations requires that putatively informative sites have avoided rampant overwriting of their phylogenetic signal. While the ultimate impact of more moderate levels of homoplasy on ancient split resolution is disputed (Müller, Borsch, & Hilu, 2006), proper phylogenetic interpretation of data with any significant rate of overwriting substitutions would only come at the cost of additional modeling (Philippe et al., 2011).

Studies of phylogenetic informativeness [PI] have quantified the relative power of loci (or other data subsets) to resolve specific evolutionary relationships by integrating substitution rate information with tree topology data, sometimes calibrating the rates using *a priori* divergence time estimates (Dornburg et al., 2019; Klopfstein, Massingham, & Goldman, 2017; Moeller &

Townsend, 2011; Townsend, 2007). This work has supported the prediction that relatively slowly-evolving loci can provide disproportionately more phylogenetic information for older splits, and vice versa for younger splits and relatively faster-evolving loci (Fong & Fujita, 2011; Townsend, López-Giráldez, & Friedman, 2008). However, direct correlations between substitution rate and phylogenetic utility are complicated by interacting factors including complex patterns and constraints in character evolution, model fit, tree topology, and taxon sampling (Aguileta et al., 2008; Dornburg et al., 2019; Heath, Zwickl, Kim, & Hillis, 2008; Klopfstein et al., 2017; Steel & Leuenberger, 2017; Su & Townsend, 2015; Townsend & Leuenberger, 2011). Additionally, because accurate estimates of substitution rate are key to most PI assessments, and these rely on well-fitting evolutionary models, the challenges associated with accurately modeling big data often limit these analyses to moderate numbers of loci.

Ortholog data

Molecular phylogenetics relies on orthologous DNA sites for comparison. Many studies target CDS for use in phylogenetics due to their straightforward amplification (e.g. through total RNA sequencing), identification, and alignment (in addition to general interest in protein-coding mutations) (Ishiwata, Sasaki, Ogawa, Miyata, & Su, 2011; Regier et al., 2010; Russo, Takezaki, & Nei, 1996). However, the phylogenetic reliability of CDS can be severely diminished in the absence of adequate evolutionary modeling (Chen, Liang, & Zhang, 2017; Reddy et al., 2017). Modeling CDS can be especially challenging due to a lack of clock-like evolution and poor model fit related to variable levels of selective constraint (Keightley, Eory, Halligan, & Kirkpatrick, 2011) and factors like codon usage bias (Galtier et al., 2018). While accurately modeling these processes is critical for the phylogenetic interpretation of CDS data, the computational requirements to model them scale up with dataset size (Philippe et al., 2011; Phillips, Delsuc, & Penny, 2004). Furthermore, long-standing biases in marker selection towards using CDS mean that less is known about the relative importance of such models when interpreting phylogenetic information from large amounts of noncoding (or non-genic) data.

In clades where multiple genomes have been well-assembled, the development of ultra-conserved element [UCE] datasets have provided one route towards expanding ortholog pools beyond mainly CDS (Bejerano, 2004; Faircloth et al., 2012; McCormack et al., 2012). UCEs are identified through whole-genome alignments, by first identifying regions of relatively high conservation (independent of locus type) and then designing ‘bait probes’ to isolate both the conserved core sequence and more variable flanking regions from all focal taxa (Bejerano, 2004; Faircloth et al., 2012; McCormack et al., 2012). Data from the flanking regions of UCEs

perform well in phylogenetic analyses (Faircloth et al., 2012; Gilbert et al., 2015); however, for clades that currently lack the genomic resources required to make use of UCE pipelines, developing *de novo* UCE datasets requires (1) generating reasonable genome assemblies for two (or ideally more) taxa, along with (2) the bioinformatics and laboratory steps associated with the probe design and bait-capture sequencing.

Alternatively, pipelines like SISRS (Schwartz, Harkins, Stone, & Cartwright, 2015) generate orthologous sequence data in an automated fashion, without the need for high-level genome assembly, locus annotation data, or reduced-representation sequencing. SISRS creates a *de novo* pan genome for the clade of interest (i.e. a 'composite genome' containing genomic regions that are conserved among focal taxa) using whole-genome sequencing [WGS] data pooled across all focal taxa. This effectively results in custom-tailored orthologs for use in the clade of interest, and because they are generated in the absence of genome assembly or annotation data, these data can be generated for clades with no pre-existing genomic resources. SISRS focuses on biallelic single-nucleotide polymorphisms [SNPs], which are known to be effective markers to resolve relationships among prokaryotic, eukaryotic, and viral groups (Gardner & Slezak, 2010; Girault, Blouin, Vergnaud, & Derzelle, 2014; McCue et al., 2012). SNPs where the variant is present in only one taxon (i.e. singletons) provide little to no topological support when inferring trees; removal of these sites from a SNP dataset yields parsimony-informative biallelic sites [PIBS]. Phylogenies can be inferred from PIBS data using multiple methods: (1) Under maximum-likelihood on concatenated PIBS or locus-partitioned datasets, employing ascertainment bias correction to correct for the lack of invariant sites (Massatti, Reznicek, & Knowles, 2016); (2) with Bayesian methods, which are typically thought to parameterize models with better fit (albeit with high computational requirements) (Rannala & Yang, 2017) and (3) quartet-based methods (i.e. sampling and analyzing four species at a time over many iterations), which have been gaining in popularity due to their ease of use, moderate memory requirements, and flexibility regarding common confounding attributes of many genome-scale datasets: ILS and large amounts of missing data (Chifman & Kubatko, 2014).

PIBS can be extracted directly from multiple-sequence alignment data in the absence of substitution rate estimates (and therefore evolutionary modeling), and the binary nature of PIBS (i.e. under parsimony, biallelic sites are either in 100% agreement or disagreement with reference topology) means that the phylogenetic site concordance (i.e. whether the two alleles reflect an accepted splitting event) can easily be calculated and compared among different PIBS groups containing millions of sites. This type of parsimony-based analysis of split support assumes no underlying evolutionary model; therefore, discordant PIBS (i.e. biallelic sites where

neither set of taxa is monophyletic in a reference tree) are not ‘phylogenetic noise’, but rather they reflect sites where accurate evolutionary modeling would be required for proper phylogenetic interpretation. Thus, significant differences in concordance rates among PIBS groups can provide a partial glimpse into the relative importance of precise and accurate modeling when using certain data subsets. Computational burdens and restrictions on dataset size are alleviated when using such model-free methods of phylogenetic data interrogation, and these larger datasets provide more opportunities for exploration and partitioning when investigating how particular subsetting strategies influence phylogenetic estimates (Jeremy M. Brown & Thomson, 2016). This allows us to redirect some of our prior focus on maximizing signal from limited variation towards strategies for sorting, binning, and filtering larger datasets down to predictively-informative subsets (Dornburg et al., 2019; Graybeal, 1994; Klopfstein et al., 2017; Townsend, 2007).

We applied SISRS to WGS reads from primate, rodent, and Pecora species with well-established relationships and annotated reference genomes to generate annotated orthologs whose phylogenetic site concordance could be assessed accurately. *Post-hoc* annotation of these loci (which were assembled using no genomic resources) revealed that they derived from all commonly annotated locus types (e.g. CDS, intronic regions, pseudogenes) in addition to unannotated/intergenic regions, and covered over 10% of the reference genome assemblies for human, mouse, and cow. We analyzed the concordance of more than 25 million PIBS, finding that over two-thirds supported a true bifurcation and that all but the smallest datasets were sufficient for accurate inference of the reference topologies, indicating a high level of phylogenetic utility and reliability. Higher proportions and numbers of concordant PIBS (e.g. those that can be accurately interpreted without modeling) derived from intronic, long noncoding RNA [lncRNA], and intergenic (i.e. unannotated) regions highlighting the utility of locus types that have received comparatively less focus. In contrast and for all clades, CDS-derived PIBS contained fewer overall sites than noncoding subsets, while also displaying disproportionately low concordance relative to other locus types. Additionally, CDS PIBS were the most likely to support the incorrect topology in two cases of topological conflict among our focal taxa. These findings reinforce the importance of accurate evolutionary modeling, particularly when datasets contain mostly coding loci. Over the 50MY of evolution associated with the clades studied here, PIBS derived from most locus types provided consistent levels of split support over time. Taken together these results provide insight into both the phylogenetic utility and the relative modeling needs of data derived from different locus types, and thus, valuable context for resolving genome-scale conflicts.

## Results

Processing of WGS reads into mammalian ortholog sets

We used the SISRS method (Schwartz et al., 2015) to generate three sets of putative orthologs using Illumina short-read data pooled across 10 species of (1) catarrhine primate ('Primates'), (2) murid rodent ('Rodents'), and (3) Pecora, plus two outgroup species per dataset (Fig. 1; Table S1). Assessing phylogenetic site concordance relies on a reference topology; therefore, we chose clades and species with robustly supported relationships which we used as reference trees (dos Reis et al., 2018; Steppan & Schenk, 2017; Zurano et al., 2019).

SISRS generates orthologs through the assembly of a 'composite genome', using read data pooled across all focal species. The assembled contigs represent genomic loci that are (1) conserved enough among study taxa to be assembled in this atypical manner and (2) present in the WGS data for most taxa (i.e. contigs are 'tailored' to be relevant for the focal dataset). For each clade, SISRS generated 3M to 6M sequences totaling 500Mb – 1Gb, with contig sizes ranging from 123bp - 18Kb (Table S2). Using the Ensembl v98 genome builds for human, mouse, and cow (Zerbino et al., 2018) we were able to map 39% (Rodents) – 88% (Primates) of SISRS contigs, resulting in annotated ortholog datasets totaling over 300Mb per clade with each covering ~13% of their respective reference genome (Tables S2-3). Using SISRS to analyze the combined dataset (all 36 mammal species) resulted in 103Mb of ortholog data that we annotated using the human reference genome (Tables S2-3).

SISRS converts the composite ortholog sequences into species-specific sequences by mapping reads from each taxon individually onto the respective dataset and replacing bases with species-specific bases if two key conditions are met: (1) sites must have been covered by at least three reads, and (2) must not have variation within the taxon (i.e. only fixed alleles with 3X coverage). All other sites were denoted as 'N'. Using 3.5Gb as a shared genome size estimate (Kapusta, Suh, & Feschotte, 2017), trimmed taxon-specific read depths ranged from 10X – 38X (Table S1). In the focal clade datasets, 23% - 78% of composite genome sites (234Mb – 479Mb; Table S4) could be positively genotyped for any given taxon, while species-specific genotyping rates in the combined analysis ranged from 11% - 37% (36.7Mb – 119Mb; Table S4).

Because SISRS and UCE-type analyses both rely on sequence conservation to identify useful data, although in different ways (SISRS: composite genome assembly; UCE: whole-genome alignment), we checked the overlap between our *de novo* SISRS orthologs and a mammal UCE dataset containing just over 1,000 UCEs (McCormack et al., 2012). Of the

~1.2Mb of UCE data, 36% (Rodents) – 51% (Pecora) of sites were also included in the SISRS orthologs, and a quarter of UCE sites were present in the combined SISRS dataset (Table S5).

Extraction of parsimony-informative biallelic sites (PIBS)

The Pecora, primate, and rodent datasets yielded 10.4M, 11.7M, and 3.3M parsimony-informative sites respectively, while the combined analysis resulted in 330K parsimony-informative sites (Table S6). Parsimony-informative biallelic sites (PIBS) made up 90.9% - 97.7% of all parsimony-informative sites across datasets (300K [Combined] – 11.5M [Primates]; Table S6). Between 82% (Rodents) - 97% (Primates) of PIBS identified in this study were found on uniquely mapped orthologs and could be annotated (Table S6; Figure S1). While PIBS made up fewer than 1% of sites from most locus types in the Rodents and Combined datasets, loci annotated as CDS in these clades yielded significantly more PIBS-per-site than other locus types when compared to the median value using a modified Z-score test (Rodents: 4.18% of all CDS sites, $p = 4.31E^{-213}$; Combined: 2.47%, $p = 1.52E^{-68}$; Table S7). In order to most accurately gauge site concordance, we only profiled sites where there was data for all taxa. When we expanded the dataset to allow one taxon to have missing data PIBS counts rose by 54% (Primates) - 99% (Rodents) and allowing two missing taxa resulted in PIBS gains of 85% (Primates) to 233% (Rodents; Table S8).

Maximum-likelihood trees inferred using concatenated PIBS are concordant among locus types

Assessing site concordance relies on an underlying topology for proper interpretation. While we chose clades with well-resolved relationships, we also tested whether PIBS data alone were sufficient to resolve the relationships among focal taxa. We concatenated PIBS from each locus type together and used these alignments to generate trees under maximum-likelihood. Of the 39 trees inferred in this study (4 datasets, 9 – 10 locus types per dataset), all but the three smallest datasets resulted in trees that were fully resolved and agreed with topologies from the literature (Figure 1; Table S9). The 631 small RNA [smRNA] PIBS from Pecora yielded a poorly supported node grouping the clade of okapi + giraffe with the deer species, while trees inferred in the combined dataset using smRNA and noncoding gene PIBS (135 and 160 sites, respectively) broadly clustered the focal clades, but many within-clade relationships were incorrectly resolved or resolved with low support (Table S9). Alignments and trees are available from the companion GitHub repository.

Coding sequence PIBS carry a lower proportion of concordant phylogenetic sites

We assessed the proportion of PIBS from each locus type that supported a split from the reference topologies and calculated the median concordance rate among locus types for each dataset. These median concordance rates ranged from 69.5% (Combined) - 90.2% (Primates; Fig. 2a; Table S10). We identified locus types with significant deviations from these median values using a modified Z-score test. PIBS derived from CDS had the lowest concordance percentage of any locus type in all clades, with concordance rates 2.2% (Primates) - 18.7% (Combined) lower than the locus-wide median values (all $p <= 2.13E^{-7}$; Fig. 2a; Table S10). smRNA PIBS contained a lower percentage of concordant sites in the Pecora (-1.58%; $p = 4.48E^{-6}$; Fig. 2a; Table S10) and Primates datasets (-0.96%; $p = 5.64E^{-15}$; Fig. 2a; Table S10). PIBS from 3'-UTR were disproportionately discordant in the Rodents dataset (-1.39%; $p = 3.08E^{-3}$; Fig. 3a; Table S10), while pseudogenic PIBS contained a lower proportion of concordant sites in the Pecora dataset (-1.40%, $p = 5.19E^{-5}$; Fig. 2a; Table S10). The only locus type to display a significantly higher percent concordance was 5'-UTR PIBS in Pecora, with a concordance percentage 1.36% above the locus type median ($p = 5.19E^{-5}$; Fig. 2a; Table S10).

Coding sequence PIBS provide disproportionate support for controversial relationships

The TimeTree database (Kumar, Stecher, Suleski, & Hedges, 2017) presents an alternative topology for the rodents and Pecora, each effectively involving a single node swap relative to the reference trees (Fig. 1). For PIBS derived from each locus type, we compared the proportion of PIBS supporting the reference and TimeTree nodes and detected outlier proportions using the same modified Z-score test described above. Across locus types, the median proportions of PIBS that supported the TimeTree relationships were 34.8% (Rodents) and 25.4% (Pecora), yet CDS PIBS supported the TimeTree relationships at a rate of 36.8% in rodents (5.82% increase; $p = 1.07E^{-6}$) and 33.4% in Pecora (28.1% increase; $p < 1E^{-128}$; Fig.2b; Table S11). Conversely for the reference tree relationships, 5'-UTR PIBS provided proportionally more support for the reference nodes in both datasets (Rodents: 67.9% [4.27% increase], $p = 2.12E^{-11}$; Pecora: 78.2% [5.85% increase], $p = 1.14E^{-128}$; Fig. 2b; Table S11), as did lncRNA PIBS in the Pecora dataset (74.5% [0.81% increase], $p = 7.46E^{-4}$; Fig. 2b; Table S11).

PIBS derived from most locus types inform about splits consistently over focal timescales

The ability to resolve a complete phylogeny relies on having sites that support the oldest splitting event through to the most recent bifurcation among focal taxa. In order to determine whether PIBS from any locus type provided disproportionate support to older or more recent

splits, we broke down the PIBS support for each split in the reference trees by locus type (e.g. 5% of the PIBS support for 'Split A' came from CDS, 30% from intergenic, etc.) and used linear models to detect changes in PIBS support proportions over time. Two different sets of divergence times were used to date and analyze our reference trees to ensure robustness to potential discrepancies: (1) We extracted divergence times from the TimeTree database, and (2) due to the topological conflicts associated with the TimeTree phylogenies, we estimated divergence times directly from our data by estimating branch lengths using our complete ortholog alignments (i.e. all orthologous sites, not just PIBS) and implementing penalized-likelihood dating methods. To compare slopes between dating methods (i.e. using TimeTree dates versus data-derived dates impact slope estimation), we (1) ran linear models with and without an interaction term for the dating method and (2) used ANOVA analysis to determine whether the dating method significantly changed the interpretation of the regression.

Proportional PIBS support remained steady over evolutionary timescales for seven of the ten locus types analyzed in this study (5'-UTR, intergenic, lncRNA, noncoding genes, pseudogenes, smRNA, and 3'-UTR), and PIBS from all locus types provided consistent support to splits over time in primates and Pecora (Figure 3; Table S12). In rodents and for the combined dataset, CDS PIBS provided proportionally more support for older nodes with split support proportions rising at a rate of 0.15%/MY and 0.26%/MY, respectively (i.e. as nodes got older, a higher proportion of PIBS support was derived from CDS; $p = 2.66E^{-3}$, $5.34E^{-4}$; Figure 3; Table S12). Conversely and in the same groups, intronic PIBS provided a higher proportion of support to more recent splits, with support proportions falling at a rate of 0.094%/MY in rodents and 0.095%/MY in Pecora ($p = 2.34E^{-3}$, $2.08E^{-4}$; Figure 3; Table S12).

While absolute node age estimates differed between study-derived ages and those from TimeTree (Table S13), the significant time-dependent trends in CDS and intronic PIBS held under both dating methods (Table S12). The only difference in results between dating methods involved rodent PIBS that derive from genic regions not annotated as CDS, UTR, or intron [Genic 'Other']. Although the study-derived and TimeTree-derived slope values for change in proportional PIBS support over time were statistically indistinguishable (this study: -0.0014%/MY; TimeTree: -0.0011%/MY; $p_{Interaction} = 0.624$), the weak trend was significant when using study-derived dates ($p = 3.78E^{-3}$; Fig. 3; Table S12) but not significant when using the TimeTree dates ($p = 0.113$; Table S12), possibly due to the difference in adjusted $R^2$ values (this study: 0.74, TimeTree: 0.31; Table S12).

## Discussion

More accessible next-generation sequencing technology is facilitating a discipline-wide shift away from resolving phylogenies using small sets of markers and towards the analysis of thousands of loci from across the genome. While increasing the size of phylogenetic datasets yields more variable sites for tree inference, accurate phylogenetic interpretation of genome-scale data also relies on our ability to model exponentially more substitution rate variation (Yang, 1994), compositional heterogeneity (Duchêne, Duchêne, & Ho, 2017; Foster, 2004), and variable evolutionary constraints (Keightley et al., 2011; Reddy et al., 2017), as well as evolutionary processes like ILS, which can muddle species tree inference (Song et al., 2012). Thus, despite genome-scale analyses, we continue to see conflicting, well-supported phylogenies, including in major groups of interest; however, parsimony-based exploration of phylogenetic information can highlight subsets of genomic data where accurate phylogenetic interpretation is possible even in the absence of complex modeling.

Large phylogenetic datasets afford conservative filtration strategies

For datasets containing only a handful of loci, robust tree inference relies on the use of all available data as well as accurately modeling as much variation as possible in order to generate the necessary phylogenetic signal (Cao et al., 1994). As datasets grow to include millions of variable sites, strategies can afford to shift from signal maximization (which can face computational hurdles) towards site selectivity as we have illustrated here. On the surface, the combination of filtration steps in this study appear to be exceptionally restrictive; our final datasets contain only sites that: (1) were biallelic, (2) fixed within species, (3) with no singletons, (4) no indels, (5) were supported by three or more reads of coverage, (6) uniquely mapped to the reference genome, and (7) had data for all focal taxa. Applying these filters resulted in a massive culling of sites (less than 3% of all assembled sites made it all the way through filtering); yet, the exceptionally large ortholog sets generated by SISRS meant that the final PIBS counts were still over 3Mb for the focal clades and over 300Kb for the combined analysis. Sites that break any of these filtering rules (or even all of them) certainly may contain relevant phylogenetic signal, but (1) filtered PIBS counts in our final datasets surpassed the total site counts (invariant + variable) of many studies, and (2) over two-thirds of the those PIBS provided phylogenetic support for accepted clade relationships that could be interpreted accurately under simple parsimony, including a staggering 90% of the roughly 12 million primate PIBS that we were able to identify here.

Estimating substitution rates (e.g. for purposes of accurate branch length estimation) typically requires information on the relative amounts of invariant sites, hypervariable sites, and

sites evolving at all rates in between (Yang & Nielsen, 2000). While models have been proposed to estimate more realistic substitution rate information from biallelic SNPs (Leaché, Banbury, Felsenstein, de Oca, & Stamatakis, 2015), PIBS filtering generally excludes sites from both ends of the substitution rate spectrum; thus, while PIBS datasets are enriched for informative data from a tree inference perspective, estimating accurate branch lengths on the resulting trees may be more challenging. However, PIBS datasets will typically derive from more traditional phylogenetic datasets (e.g. alignments of whole loci), and substitution rate estimates can be derived from this starting data using traditional methods, as we do in this study when estimating divergence times for our reference topologies.

The substantial overlap between the data generated with SISRS and the loci from a large mammal UCE project (McCormack et al., 2012) suggests that both methods are honing in on similar attributes as potentially useful (i.e. evolving under rates suitable for alignment [UCE] or assembly [SISRS]). Yet, while both ortholog discovery methods provide similar data, the SISRS datasets are substantially larger than many contemporary phylogenomics/UCE-based studies and do not require high-level genome assemblies, alignments, or probe/bait design to generate. For very large datasets (i.e. where WGS data collection for all samples may be impractical), the pipeline described here can also be applied to the analysis (or re-analysis) of reduced-representation datasets such as UCE or RADseq data, albeit with an expected reduction in final dataset sizes.

Site concordance analyses find noncoding loci are a rich source of phylogenetically-reliable data

CDS have been a large focus of phylogenetic research for decades, due in part to the relative ease of processing CDS data along with general interest in protein-coding mutations. However, the phylogenetic reliability of coding sequences relies on accurate modeling (Chen et al., 2017; Doyle et al., 2015; Reddy et al., 2017), such as the incorporation of models that account for nonhomogenous base substitution (Galtier & Gouy, 1998) and codon-usage bias (Galtier et al., 2018). CDS blocks affected by strong linkage may also exacerbate the impact of ILS, which has been shown to impact coding regions even at the within-gene level (Scornavacca & Galtier, 2016). The computational requirements for applying these highly parameterized models will scale with dataset size (Philippe et al., 2011), which suggests that CDS-biased analyses may be a computationally inefficient way to make use of genome-scale data. Furthermore, the comparatively limited research on noncoding loci at the genomic scale leaves much of what we know about CDS largely uncontextualized, but our findings suggest

benefits to intentionally shifting focus towards noncoding loci as a potentially richer and more robust dataset for tree inference (at least in the clades studied here).

PIBS derived from all locus types were sufficient for recovering clade relationships among our selected primate, rodent, and Pecora taxa (provided there were enough variable sites), reinforcing the use of PIBS broadly as a reliable, informative data subset (Leaché & Oaks, 2017); yet, our results also support findings of increased modeling requirements when working with CDS data. In all clades, noncoding-derived PIBS harbored significantly more concordant sites (both proportionally and absolutely) relative to coding loci, and the practical implications of using more model-reliant data subsets can be seen in our interrogation of the genomic sources of topological conflict between trees from our reference studies and those from the TimeTree database. In TimeTree, the placement of *Mastomys* within the murid rodents, and the okapi and giraffe among Pecora, differ from the reference topologies by a single swapped node (Kumar et al., 2017); in both cases, we found that CDS PIBS supported the split from TimeTree at significantly higher rates than other genomic subsets (although not by a majority of sites in either case). This result provides a tangible example of how mis- or undermodeled CDS data may be more likely to result in the inference of an incorrect topology (Wiens, 1998), a problem likely exacerbated when working with small sets of loci (Cao et al., 1994).

As datasets expand and researchers can afford to be more selective with their data, the ability to contrast the absolute and proportional support for alternative topologies among genome-scale subsets can provide reasonable grounds for down-weighting incompatible phylogenies derived from subsets containing more complex signal. Analyses like those performed here, and related strategies such as the quartet-based calculation of site concordance factors (Minh, Hahn, & Lanfear, 2020), scale easily to accommodate genome-scale data; furthermore, unlike traditional bootstrapping techniques they do not suffer from artificial inflation when applied to large datasets (Kumar et al., 2012; Salichos & Rokas, 2013). However, by leveraging our atypically large datasets which included no missing data, we circumvented the (situationally useful) abstraction of quartet analysis and instead present a novel, site-by-site genome-scale analysis of millions of fixed alleles with data for all sites and taxa, while still maintaining low computational overhead.

SISRS-generated PIBS derived from most locus types provide broad phylogenetic support over evolutionary timescales

For decades, there has been a ‘casual’ understanding regarding the relative utility of locus types over evolutionary timescales based on relatively simplified views of molecular evolution (e.g. CDS evolves slowly and has phylogenetic utility for older splits, while less constrained locus types have increased utility for recent splits due to faster evolution), and some studies bear this out with quantifiable data (Fong & Fujita, 2011; Townsend et al., 2008). Here too, in fact, CDS-derived PIBS provided more support for older splits among rodents and in the combined analysis. The combined analysis is associated with deeper timescales, while rodents have the fastest generation times among the focal clades (i.e. rodents experience more generations, and thus more mutations, per unit time) (Sims, Jun, Wu, & Kim, 2009). Thus, in these two datasets homoplasy is expected to be more common, and our results support the idea that functional constraints in protein-coding sequences may, to some extent, convert CDS into a sort of ‘genomic sanctuary’, providing some protection against repeated mutations through purifying selection (Yang, 1993). Conversely in the same groups, intronic PIBS tended to support more recent splitting events; while it is tempting to explain this trend using the same line of reasoning (i.e. relaxed constraints within intronic regions lead to higher probabilities of overwriting substitutions), if this were the case we should expect to see similar trends among PIBS derived from locus types expected to mutate at the most unconstrained rates (i.e. those within pseudogenic or intergenic regions). Yet, these locus types showed no significant deterioration in signal over evolutionary time in any dataset, suggesting that the trend in introns may involve a more complex interplay of evolutionary forces while also reinforcing findings that suggest the impact of homoplasy within canonically fast-evolving loci may be less dramatic than previously considered (Müller et al., 2006).

However, locus type is often not a simple predictor of time-dependent phylogenetic utility: canonically rapidly-evolving genes like the plastid gene matK have been used to resolve splitting events in plants reaching as far back as 475MYA (Hilu, Black, & Oza, 2014; Lutzoni et al., 2018), while CDS has been used as subspecies population markers for many groups (Biswas et al., 2020; Frenkel et al., 2012). Broadly, assuming the majority of PIBS derive from single-mutation events (which is suggested by the high concordance rates in most locus types), these data should contain a sampling of sites evolving under ideal conditions for species tree inference. Indeed, we found that PIBS derived from seven of the ten locus types queried here provided consistent phylogenetic support to nodes spread over the 50 million years of evolution associated with all study taxa, and PIBS from all locus types provided consistent phylogenetic signal to nodes of all ages among Pecora and primates. Thus, in the absence of explicit substitution rate estimation or modeling, PIBS filtering provides an efficient method to distill

genome-scale datasets down to sites that are informative across evolutionary timescales, while also providing further evidence that phylogenetic information about older or more recent splits is not restricted to any particular locus types, at least at the timescales associated with our clades.

Conclusions

In this study, we provide a genome-scale perspective on the phylogenetic utility of parsimony-informative biallelic sites (PIBS) derived from different locus types as they apply to resolving species relationships among three mammal clades. PIBS derived from noncoding regions provided higher proportions and amounts of phylogenetically concordant sites compared to CDS PIBS in all datasets, underlining the importance of accurate modeling when inferring trees from coding data. These results suggest potential benefits in shifting away from primarily targeting coding regions for phylogenetic studies, particularly in this era of accessible whole-genome sequence data. Across 50MY of mammal evolution, we find that changes in phylogenetic utility of PIBS over time were limited to specific genic subsets, and that these patterns were both subtle and clade-specific. These findings provide motivation to expand locus sets into the more understudied regions of the genome in order to resolve some of the more recalcitrant relationships in evolutionary biology. Additionally, we recognize that our results focus on mammals at a limited timescale; thus, we encourage future work using this approach to examine larger timeframes and a diversity of taxa to provide a greater understanding of the general applicability of our results.

## 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 information in genomic data relies on having sequence data for species with well-supported evolutionary relationships. To that end, we identified three mammalian clades with well-established relationships (Fig. 1) and sufficient whole-genome sequencing [WGS] data: catarrhine primates (dos Reis et al., 2018), murid rodents (Steppan & Schenk, 2017), and members of the infraorder Pecora (Zurano et al., 2019). 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 (Table S1). 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 36 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/).

Generating Ortholog Sequences from WGS Reads

We used the SISRS pipeline to generate *de novo* ortholog data (i.e. a 'composite genome') for each dataset (Primates, Rodents, Pecora, and Combined). SISRS uses WGS reads pooled across all taxa in the dataset to generate a set of genomic loci that are (1) present in the WGS data for most species, and (2) conserved enough among taxa to be assembled together from pooled reads using a typical genome assembly program, and therefore compared among taxa. Briefly, based on a genome size estimate of 3.5Gb per dataset (Kapusta et al., 2017), we first subsampled bases equivalently from each taxon so that the final assembly depth was ~10X genomic coverage (e.g. 35Gb total, equivalently sampled from each taxon). By subsampling reads prior to assembly, regions of relatively high sequence conservation have sufficient depth for assembly while taxon-specific or poorly conserved regions will fail to assemble. We used Ray v.2.3.2-devel (Boisvert, Laviolette, & Corbeil, 2010) to assemble the composite genome using the subsampled reads from all taxa pooled together, default parameters, and a k-value of 31.

In order to generate species-specific ortholog sets from this composite assembly, SISRS maps all 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. SISRS uses the mapping information from each species to replace bases in the composite genome with species-specific bases when two key conditions are met: (1) sites must have been covered by at least three reads, and (2) must not have variation within the taxon. Any sites with insufficient read coverage or within-taxon variation were denoted as 'N', resulting in orthologs containing only information about alleles that are fixed within species.

In order to contextualize our results in light of alternative methods for identifying orthologs, we identified the overlap between our SISRS orthologs and markers generated as part of a large and well-cited mammal UCE phylogenomics study (McCormack et al., 2012). Briefly, this UCE study generated multiple sets of loci through whole-genome alignment and used each dataset to better understand the relationships among different mammalian clades.

Two of the four locus sets from this study (hereafter referred to as UCE-183 and UCE-917 based on the total number of loci) contained each of the reference species used in our study (*Homo sapiens, Mus musculus,* and *Bos taurus*). For UCE-183 and UCE-917, we mapped the UCE loci onto the appropriate reference genome and derived genome mapping coordinates in the same way we processed our SISRS contigs. We calculated the percent overlap between our loci and the UCE loci using the *intersect* function from BEDTools v.2.26 (Quinlan, 2014) on the corresponding coordinate files.

Composite genome annotation

We obtained chromosomal and mitochondrial scaffolds along with associated annotation data for *Homo sapiens, Mus musculus,* and *Bos taurus* from the Ensembl Build 98 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 & 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 locus types: (1) coding sequences (CDS, including all annotated transcript variants), (2) 3' untranslated regions (3'-UTR), (3) 5'-UTR, (4) intronic regions, (5) 'other' genic regions (sites within genes that were not annotated as CDS, UTR, or intronic), (6) long-noncoding RNAs (lncRNAs), (7) noncoding genes (genes without annotated CDS; none annotated in Pecora), (8) pseudogenes, or (9) small RNAs (smRNA including miRNAs + ncRNAs + rRNAs + scRNAs + smRNAs + snoRNAs + snRNAs + tRNAs + vaultRNAs). Any reference genome position that was not annotated as one of these locus types was denoted as (10) intergenic, although these could also be called '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 Ensembl v98 annotation files, the output from the Bowtie2 reference genome mapping, and the intersect function in BEDTools v.2.26.

In this study, we perform multiple percentage comparisons among locus types; due to the small number of categories (9 locus types in Pecora, 10 in primates and rodents), we assessed statistical significance between locus types using a two-tailed modified Z-score analysis, which is robust at detecting deviations within small sample sizes (e.g. n=9 or n=10) (Leys, Ley, Klein, Bernard, & Licata, 2013). We used this modified Z-score analysis to assess

locus-type differences in the proportion of sites from each reference genome that were assembled into the composite genome. Based on the number of annotation subsets present in each dataset (10 in Primates, Rodents, and Combined; 9 in Pecora) critical Z-score values indicative of significant assembly biases were identified at a Bonferroni-corrected α = 0.05/10 ($Z_{Critical}$ = 2.81) or α = 0.05/9 ($Z_{Critical}$ = 2.77).

Isolation of parsimony-informative biallelic sites (PIBS)

We used SISRS to scan each site along the mapped composite contigs, identifying and flagging parsimony-informative sites with different patterns of sequence variation. Filtering phylogenetic data down to parsimony-informative sites involves removing sites with no interspecific variation (i.e. invariant sites) as well as any site where a single taxon had its own unique allele (i.e. singletons). Furthermore, we only included sites where there was fixed allele data for all taxa (i.e. no 'N's) and did not include indel sites (i.e. sites where the variation consists of a gap and an otherwise invariant nucleotide). While the remaining parsimony-informative sites included bi-, tri-, and quadallelic sites, the binary nature of PIBS allows for the most straightforward statistical assessment; thus, sites with biallelic variation were selected for full phylogenetic site concordance profiling. In order to assess whether certain locus types carried a higher or lower proportion of PIBS, we used the modified Z-score test as described above.

PIBS phylogeny-building and concordance analysis

We built phylogenies using concatenated PIBS data from each locus type and dataset. We inferred all trees using a maximum-likelihood approach as implemented in IQ-TREE v1.7-beta16 (Nguyen, Schmidt, von Haeseler, & Minh, 2015), using the best-fit model as determined by IQ-TREE and 5000 ultrafast bootstrap replicates. PIBS partition a dataset into a pair of taxonomic groups, with each defined by one of two possible alleles. To assess phylogenetic site concordance, we used custom scripts in Biopython (Cock et al., 2009) and R v.3.6.3 (R Core Team, 2020) (scripts available in the GitHub repo) to scan each site in the alignments and report back the two sets of clustered taxa. We then scored each site as concordant or discordant with respect to the reference trees from the literature. We identified locus types that carried higher percentages of concordant (or discordant) signal using the modified Z-score analysis as previously described.

Detecting changes in phylogenetic utility over time

To assess whether PIBS derived from certain locus types informed broadly about splits in the tree (or conversely, contained more information about older or younger splits), we broke down the PIBS support by locus type for each node in the reference trees (e.g. 5% of the support for 'Split A' came from CDS, 30% from intergenic, etc.). Using this annotation breakdown along with the estimated age of each node (see below), we then applied linear models using R to detect time-dependent trends in PIBS support. Statistical significance of the regressions was interpreted at Bonferroni-corrected α values based on the number of locus types per dataset. Two sets of divergence times were used to test the phylogenetic utility of PIBS over time: (1) We generated divergence time estimates from our whole-ortholog alignments, and (2) we used divergence time downloaded from the TimeTree database (Kumar et al., 2017).

To estimate the node ages based off our SISRS orthologs, we first concatenated the alignments of all composite contigs that could be uniquely mapped back to the reference genome. We then used these alignments to estimate branch lengths on the reference tree using the best-fitting evolutionary model in IQ-TREE. With these branch lengths, we applied penalized likelihood (Sanderson, 2002) to estimate node ages on each reference tree in R using the *chronos* function as implemented in the package *ape* v.5.3 (Paradis & 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. The focal group trees (Pecora, primates, and rodents) were calibrated at the root node using the TimeTree divergence time confidence intervals as the minimum and maximum bound estimates. In the same way, the combined topology was calibrated at the base of the tree, but also at the calibration nodes from the focal group analyses. 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.

Concatenating all loci and modeling them under one substitution model is an overly simplistic method to estimate branch lengths; however, due to the size of our datasets, some commonly-used node dating strategies (e.g. Bayesian inference, partition modeling) were too computationally costly to implement. Therefore, to provide robustness to discrepancies in estimated divergence times we also assessed time-dependent trends using node ages pulled directly from the TimeTree database, which compiles divergence dates from multiple published studies. For each locus type and dataset, we determined whether slopes varied between dating methods by using R to fit linear models to the data, both with and without an interaction term for the dating method. We ran an ANOVA on the two models to determine whether removing the

dating method interaction term significantly affected the model fit (i.e. whether dating method affects slope estimation).

For both the Rodents and Pecora datasets, the TimeTree topologies differed from the reference topologies at one node each (red outlined nodes in Fig. 1). These topological discrepancies provided a direct opportunity to test whether PIBS could be used to identify potential sources of such conflict. For each annotation subset, except for smRNA, which contained too few sites to query, we calculated the proportion of PIBS that supported the reference topology and the TimeTree topology and detected annotation biases in PIBS split support using the modified Z-score test described above.

**Acknowledgements**

This work was supported by the National Science Foundation (1812201). We thank the anonymous reviewers whose comments helped improve and clarify this manuscript.

**Figure Legends**

Figure 1: Evolutionary relationships among study taxa. These relationships, supported by three independent phylogenomic studies, were also fully resolved in 36/39 trees inferred in this study. For each split in the tree, the size of filled node icons is proportional to the number of parsimony-informative biallelic sites [PIBS] that support that split under parsimony (i.e. clustering taxa by alleles and assessing monophyly). Split support ranges for each focal group were as follows: Pecora (green squares): 173K - 1.04M; Primates (orange triangles): 148K - 1.86M; 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 - 33.9K sites). Tip labels for reference annotation species are red and bolded. Relative to splits seen in the reference topologies, nodes outlined in red are swapped in the TimeTree database.

Figure 2: Concordance rates of parsimony-informative biallelic site [PIBS] derived from different locus types. Modified Z-score analysis of genome-wide PIBS concordance (i.e. the proportion of sites where biallelic variation reflects a true split event) reveals that PIBS derived from different locus types varied significantly the proportion of sites supporting **(a)** the entire reference tree, and **(b)** two conflicting nodes from the TimeTree database for rodents and Pecora. Filled shapes indicate locus types with concordance percentages that are either significantly higher or lower than the median concordance among locus types. **(a)** Across datasets, PIBS derived from

CDS displayed the lowest concordance relative to all locus types (all p <= $2.13E^{-7}$). **(b)** When comparing support for the correct relationships and the incompatible phylogenies from TimeTree, CDS PIBS were most likely to support the incorrect topology in both cases (both p <= $1.08E^{-6}$). Conversely, 5'-UTR PIBS provided proportionally more support for the reference relationships (both p <= $2.12E^{-11}$).

Figure 3: Changes in phylogenetic utility over time among locus types. Based on divergence times estimated from SISRS orthologs (displayed here) as well as dates from the TimeTree database, we ran linear regression analyses to determine whether the proportion of parsimony-informative biallelic sites [PIBS] from different locus types changed in their phylogenetic utility over time. Filled shaped indicate locus types where PIBS inform disproportionately on older or more recent splits. Among rodents and in the combined analysis, CDS-derived PIBS (upper left) provided proportionally more support for older splits (both p <= $1.08E^{-6}$), while conversely and for the same groups, intron-derived PIBS (upper right) informed disproportionately about younger splits (both p <= $2.34E^{-3}$). Sites from genes that were not annotated as CDS, UTR, or intron ('Genic (Other)'; lower left) show a weaker trend towards increased utility at younger nodes in rodents (p = $3.77E^{-3}$), but the relationship is not significant when using dates from TimeTree (p=.113). No other locus type, including intergenic/unannotated sites (lower right), displayed any time-dependent shifts in phylogenetic support.


Aguileta, G., Marthey, S., Chiapello, H., Lebrun, M. H., Rodolphe, F., Fournier, E., . . . Giraud, T. (2008). Assessing the performance of single-copy genes for recovering robust phylogenies. *Syst. Biol., 57*(4), 613-627. doi:10.1080/10635150802306527

Bejerano, G. (2004). Ultraconserved Elements in the Human Genome. *Science, 304*(5675), 1321-1325. doi:10.1126/science.1098119

Biswas, M. K., Bagchi, M., Nath, U. K., Biswas, D., Natarajan, S., Jesse, D. M. I., . . . Nou, I.-S. (2020). Transcriptome wide SSR discovery cross-taxa transferability and development of marker database for studying genetic diversity population structure of Lilium species. *Scientific reports, 10*(1), 1-13.

Bleidorn, C. (2017). Sources of Error and Incongruence in Phylogenomic Analyses. *Phylogenomics*, 173-193. doi:10.1007/978-3-319-54064-1_9

Boisvert, S., Laviolette, F., & Corbeil, J. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J. Comput. Biol., 17*(11), 1519-1533. doi:10.1089/cmb.2009.0238

Brown, J. M. (2014). Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Systematic Biology, 63*(3), 334-348.

Brown, J. M., & Thomson, R. C. (2016). Bayes Factors Unmask Highly Variable Information Content, Bias, and Extreme Influence in Phylogenomic Analyses. *Systematic Biology, 66*(4), 517-530. doi:10.1093/sysbio/syw101

Cao, Y., Adachi, J., Janke, A., Pääbo, S., & Hasegawa, M. (1994). Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *J. Mol. Evol., 39*(5), 519-527. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7807540

Chen, M.-Y., Liang, D., & Zhang, P. (2017). Phylogenomic Resolution of the Phylogeny of Laurasiatherian Mammals: Exploring Phylogenetic Signals within Coding and Noncoding Sequences. *Genome Biol. Evol., 9*(8), 1998-2012. doi:10.1093/gbe/evx147

Chifman, J., & Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics, 30*(23), 3317-3324.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., . . . Wilczynski, B. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics, 25*(11), 1422-1423.

Dornburg, A., Su, Z., & Townsend, J. P. (2019). Optimal Rates for Phylogenetic Inference and Experimental Design in the Era of Genome-Scale Data Sets. *Syst. Biol., 68*(1), 145-156. doi:10.1093/sysbio/syy047

dos Reis, M., Gunnell, G. F., Barba-Montoya, J., Wilkins, A., Yang, Z., & Yoder, A. D. (2018). Using Phylogenomic Data to Explore the Effects of Relaxed Clocks and Calibration Strategies on Divergence Time Estimation: Primates as a Test Case. *Syst. Biol., 67*(4), 594-615. doi:10.1093/sysbio/syy001

Doyle, V. P., Young, R. E., Naylor, G. J. P., & Brown, J. M. (2015). Can We Identify Genes with Increased Phylogenetic Reliability? *Syst. Biol., 64*(5), 824-837. doi:10.1093/sysbio/syv041

Duchêne, D. A., Duchêne, S., & Ho, S. Y. (2017). New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. *Molecular Biology and Evolution, 34*(6), 1529-1534.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol., 61*(5), 717-726. doi:10.1093/sysbio/sys004

Fong, J. J., & Fujita, M. K. (2011). Evaluating phylogenetic informativeness and data-type usage for new protein-coding genes across Vertebrata. *Molecular Phylogenetics and Evolution, 61*(2), 300-307. doi:10.1016/j.ympev.2011.06.016

Foster, P. G. (2004). Modeling compositional heterogeneity. *Systematic Biology, 53*(3), 485-495.

Frenkel, O., Portillo, I., Brewer, M., Peros, J.-P., Cadle-Davidson, L., & Milgroom, M. (2012). Development of microsatellite markers from the transcriptome of Erysiphe necator for analysing population structure in North America and Europe. *Plant Pathology, 61*(1), 106-119.

Galtier, N., & Gouy, M. (1998). Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol., 15*(7), 871-879. doi:10.1093/oxfordjournals.molbev.a025991

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glemin, S., . . . Duret, L. (2018). Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Mol Biol Evol, 35*(5), 1092-1103. doi:10.1093/molbev/msy015

Gardner, S. N., & Slezak, T. (2010). Scalable SNP analyses of 100+ bacterial or viral genomes. *J Forensic Res, 1*(03), 1-5.

Gee, H. (2003). Evolution: ending incongruence. *Nature, 425*(6960), 782. doi:10.1038/425782a

Gilbert, P. S., Chang, J., Pan, C., Sobel, E. M., Sinsheimer, J. S., Faircloth, B. C., & Alfaro, M. E. (2015). Genome-wide ultraconserved elements exhibit higher phylogenetic

informativeness than traditional gene markers in percomorph fishes. *Mol. Phylogenet. Evol., 92*, 140-146. doi:10.1016/j.ympev.2015.05.027
Girault, G., Blouin, Y., Vergnaud, G., & Derzelle, S. (2014). High-throughput sequencing of Bacillus anthracis in France: investigating genome diversity and population structure using whole-genome SNP discovery. *BMC genomics, 15*(1), 1-10.
Graybeal, A. (1994). Evaluating the Phylogenetic Utility of Genes: A Search for Genes Informative About Deep Divergences Among Vertebrates. *Systematic Biology, 43*(2), 174. doi:10.2307/2413460
Heath, T. A., Zwickl, D. J., Kim, J., & Hillis, D. M. (2008). Taxon Sampling Affects Inferences of Macroevolutionary Processes from Phylogenetic Trees. *Systematic Biology, 57*(1), 160-166. doi:10.1080/10635150701884640
Hilu, K. W., Black, C. M., & Oza, D. (2014). Impact of gene molecular evolution on phylogenetic reconstruction: A case study in the rosids (superorder Rosanae, angiosperms). *PLoS One, 9*(6), e99725.
Ishiwata, K., Sasaki, G., Ogawa, J., Miyata, T., & Su, Z.-H. (2011). Phylogenetic relationships among insect orders based on three nuclear protein-coding gene sequences. *Mol. Phylogenet. Evol., 58*(2), 169-180. doi:10.1016/j.ympev.2010.11.001
Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., . . . Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science, 346*(6215), 1320-1331. doi:10.1126/science.1253451
Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. U. S. A., 114*(8), E1460-E1469. doi:10.1073/pnas.1616702114
Keightley, P. D., Eory, L., Halligan, D. L., & Kirkpatrick, M. (2011). Inference of mutation parameters and selective constraint in mammalian coding sequences by approximate Bayesian computation. *Genetics, 187*(4), 1153-1161. doi:10.1534/genetics.110.124073
Klopfstein, S., Massingham, T., & Goldman, N. (2017). More on the Best Evolutionary Rate for Phylogenetic Analysis. *Syst. Biol., 66*(5), 769-785. doi:10.1093/sysbio/syx051
Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L., & Tamura, K. (2012). Statistics and truth in phylogenomics. *Mol. Biol. Evol., 29*(2), 457-472. doi:10.1093/molbev/msr202
Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol., 34*(7), 1812-1819. doi:10.1093/molbev/msx116
Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923
Leaché, A. D., Banbury, B. L., Felsenstein, J., de Oca, A. n.-M., & Stamatakis, A. (2015). Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. *Systematic Biology, 64*(6), 1032-1047. doi:10.1093/sysbio/syv053
Leaché, A. D., & Oaks, J. R. (2017). The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics, 48*(1), 69-84. doi:10.1146/annurev-ecolsys-110316-022645
Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., . . . Cochrane, G. (2011). The European Nucleotide Archive. *Nucleic Acids Res., 39*(Database issue), D28-31. doi:10.1093/nar/gkq967
Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764-766. doi:10.1016/j.jesp.2013.03.013
Lutzoni, F., Nowak, M. D., Alfaro, M. E., Reeb, V., Miadlikowska, J., Krug, M., . . . Magallón, S. (2018). Contemporaneous radiations of fungi and plants linked to symbiosis. *Nature Communications, 9*(1), 5451. doi:10.1038/s41467-018-07849-9

Massatti, R., Reznicek, A. A., & Knowles, L. L. (2016). Utilizing RADseq data for phylogenetic analysis of challenging taxonomic groups: A case study in Carex sect. Racemosae. *American Journal of Botany, 103*(2), 337-347.
McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res., 22*(4), 746-754. doi:10.1101/gr.125864.111
McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., . . . Hill, E. W. (2012). A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet, 8*(1), e1002451.
Minh, B. Q., Hahn, M. W., & Lanfear, R. (2020). New methods to calculate concordance factors for phylogenomic datasets. *Molecular Biology and Evolution, 37*(9), 2727-2733.
Moeller, A. H., & Townsend, J. P. (2011). Phylogenetic informativeness profiling of 12 genes for 28 vertebrate taxa without divergence dates. *Molecular Phylogenetics and Evolution, 60*(2), 271-272. doi:10.1016/j.ympev.2011.04.023
Müller, K. F., Borsch, T., & Hilu, K. W. (2006). Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting matK, trnT-F, and rbcL in basal angiosperms. *Molecular Phylogenetics and Evolution, 41*(1), 99-117.
Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol., 32*(1), 268-274. doi:10.1093/molbev/msu300
Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., . . . Wörheide, G. (2013). Deep metazoan phylogeny: when different genes tell different stories. *Mol. Phylogenet. Evol., 67*(1), 223-233. doi:10.1016/j.ympev.2013.01.010
Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics, 35*(3), 526-528. doi:10.1093/bioinformatics/bty633
Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., & Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol., 9*(3), e1000602. doi:10.1371/journal.pbio.1000602
Phillips, M. J., Delsuc, F., & Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol, 21*(7), 1455-1458. doi:10.1093/molbev/msh137
Quinlan, A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics, 47*, 11.12.11-34. doi:10.1002/0471250953.bi1112s47
R Core Team. (2020). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org
Rannala, B., & Yang, Z. (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Systematic Biology, 66*(5), 823-842.
Reddy, S., Kimball, R. T., Pandey, A., Hosner, P. A., Braun, M. J., Hackett, S. J., . . . Braun, E. L. (2017). Why Do Phylogenomic Data Sets Yield Conflicting Trees? Data Type Influences the Avian Tree of Life more than Taxon Sampling. *Syst. Biol., 66*(5), 857-879. doi:10.1093/sysbio/syx041
Regier, J. C., Shultz, J. W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., . . . Cunningham, C. W. (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature, 463*(7284), 1079-1083. doi:10.1038/nature08742
Rodriguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B. F., & Philippe, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol, 56*(3), 389-399. doi:10.1080/10635150701397643
Rokas, A., & Carroll, S. B. (2008). Frequent and widespread parallel evolution of protein sequences. *Mol. Biol. Evol., 25*(9), 1943-1953. doi:10.1093/molbev/msn143

Rokas, A., King, N., Finnerty, J., & Carroll, S. B. (2003). Conflicting phylogenetic signals at the base of the metazoan tree. *Evol. Dev., 5*(4), 346-359. doi:10.1046/j.1525-142x.2003.03042.x

Russo, C. A., Takezaki, N., & Nei, M. (1996). Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol., 13*(3), 525-536. doi:10.1093/oxfordjournals.molbev.a025613

Salichos, L., & Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature, 497*(7449), 327-331. doi:10.1038/nature12130

Sanderson, M. J. (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol., 19*(1), 101-109. doi:10.1093/oxfordjournals.molbev.a003974

Schwartz, R. S., Harkins, K. M., Stone, A. C., & Cartwright, R. A. (2015). A composite genome approach to identify phylogenetically informative data from next-generation sequencing. *BMC Bioinformatics, 16*, 193. doi:10.1186/s12859-015-0632-y

Scornavacca, C., & Galtier, N. (2016). Incomplete Lineage Sorting in Mammalian Phylogenomics. *Systematic Biology*, syw082. doi:10.1093/sysbio/syw082

Sharma, P. P., Kaluziak, S. T., Pérez-Porro, A. R., González, V. L., Hormiga, G., Wheeler, W. C., & Giribet, G. (2014). Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. *Mol. Biol. Evol., 31*(11), 2963-2984. doi:10.1093/molbev/msu235

Sims, G. E., Jun, S.-R., Wu, G. A., & Kim, S.-H. (2009). Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proc. Natl. Acad. Sci. U. S. A., 106*(40), 17077-17082. doi:10.1073/pnas.0909377106

Song, S., Liu, L., Edwards, S. V., & Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. U. S. A., 109*(37), 14942-14947. doi:10.1073/pnas.1211733109

Steel, M., & Leuenberger, C. (2017). The optimal rate for resolving a near-polytomy in a phylogeny. *Journal of Theoretical Biology, 420*, 174-179. doi:10.1016/j.jtbi.2017.02.037

Steppan, S. J., & Schenk, J. J. (2017). Muroid rodent phylogenetics: 900-species tree reveals increasing diversification rates. *PLoS One, 12*(8), e0183070. doi:10.1371/journal.pone.0183070

Su, Z., & Townsend, J. P. (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 Evol. Biol., 15*, 86. doi:10.1186/s12862-015-0364-7

Townsend, J. P. (2007). Profiling phylogenetic informativeness. *Syst. Biol., 56*(2), 222-231. doi:10.1080/10635150701311362

Townsend, J. P., & Leuenberger, C. (2011). Taxon sampling and the optimal rates of evolution for phylogenetic inference. *Syst. Biol., 60*(3), 358-365. doi:10.1093/sysbio/syq097

Townsend, J. P., López-Giráldez, F., & Friedman, R. (2008). The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree. *J. Mol. Evol., 67*(5), 437-447. doi:10.1007/s00239-008-9142-0

Wiens, J. J. (1998). The accuracy of methods for coding and sampling higher-level taxa for phylogenetic analysis: a simulation study. *Systematic Biology, 47*(3), 397-413.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution, 10*(6), 1396-1401. doi:10.1093/oxfordjournals.molbev.a040082

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution, 39*(3), 306-314. doi:10.1007/BF00160154

Yang, Z., & Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution, 17*(1), 32-43.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., . . . Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Res., 46*(D1), D754-D761. doi:10.1093/nar/gkx1098

Zurano, J. P., Magalhães, F. M., Asato, A. E., Silva, G., Bidau, C. J., Mesquita, D. O., & Costa, G. C. (2019). Cetartiodactyla: Updating a time-calibrated molecular phylogeny. *Mol. Phylogenet. Evol., 133*, 256-262. doi:10.1016/j.ympev.2018.12.015

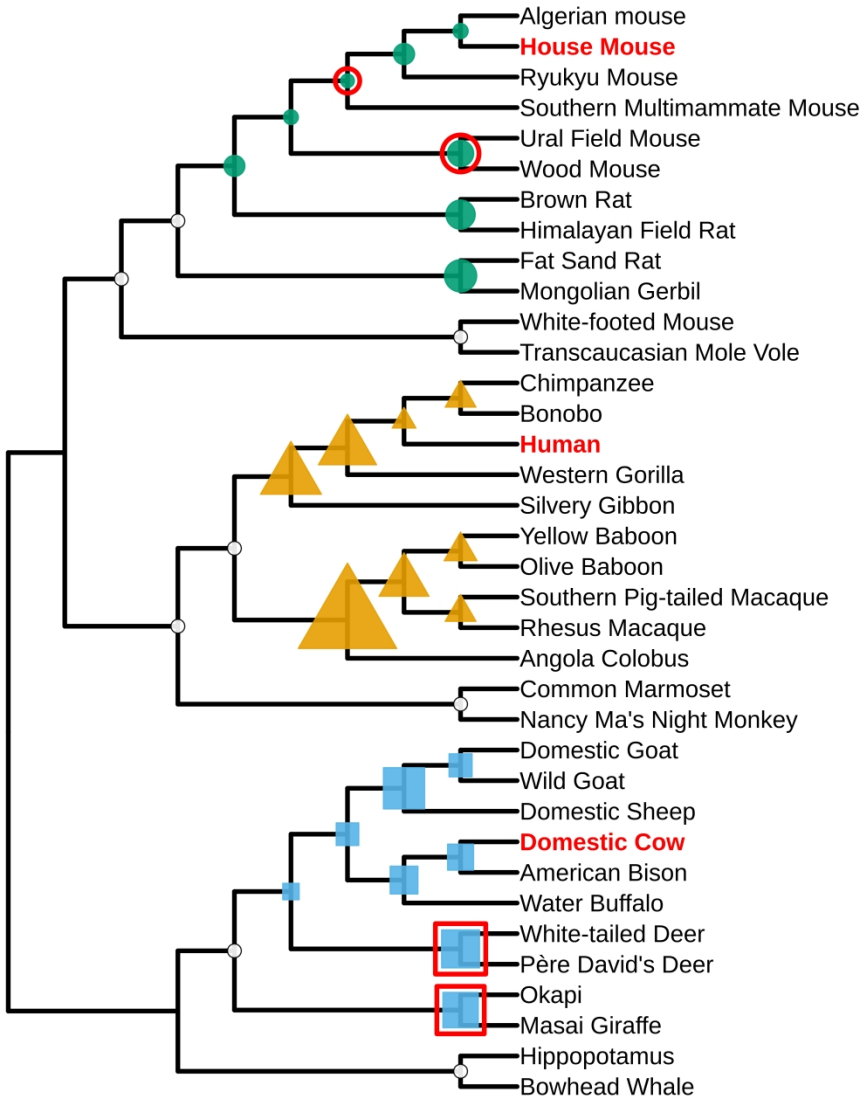


Figure 1: Evolutionary relationships among study taxa. These relationships, supported by three independent phylogenomic studies, were also fully resolved in 36/39 trees inferred in this study. For each split in the tree, the size of filled node icons is proportional to the number of parsimony-informative biallelic sites [PIBS] that support that split under parsimony (i.e. clustering taxa by alleles and assessing monophyly). Split support ranges for each focal group were as follows: Pecora (green squares): 173K - 1.04M; Primates (orange triangles): 148K - 1.86M; 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 - 33.9K sites). Tip labels for reference annotation species are red and bolded. Relative to splits seen in the reference topologies, nodes outlined in red are swapped in the TimeTree database

(Full color requested)

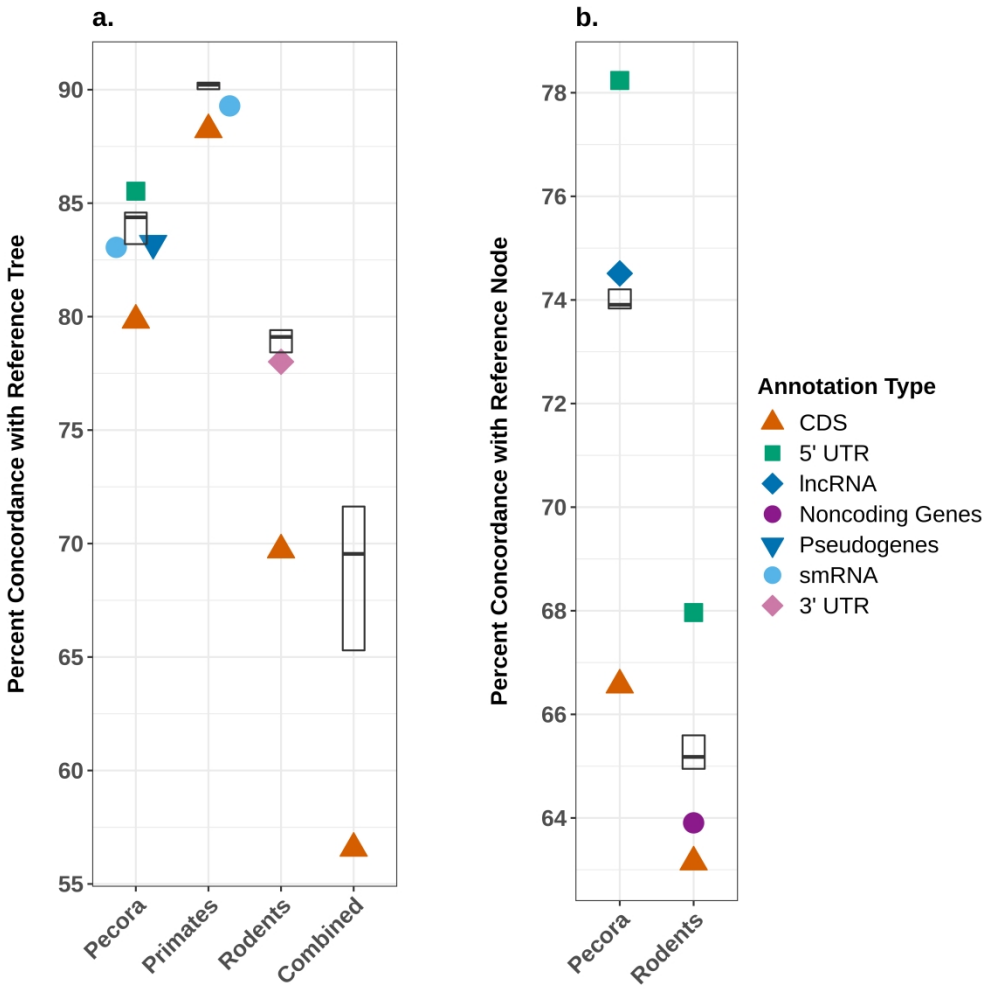


Figure 2: Concordance rates of parsimony-informative biallelic site [PIBS] derived from different locus types. Modified Z-score analysis of genome-wide PIBS concordance (i.e. the proportion of sites where biallelic variation reflects a true split event) reveals that PIBS derived from different locus types varied significantly the proportion of sites supporting (a) the entire reference tree, and (b) two conflicting nodes from the TimeTree database for rodents and Pecora. Filled shapes indicate locus types with concordance percentages that are either significantly higher or lower than the median concordance among locus types. (a) Across datasets, PIBS derived from CDS displayed the lowest concordance relative to all locus types (all p <= 2.13E-7). (b) When comparing support for the correct relationships and the incompatible phylogenies from TimeTree, CDS PIBS were most likely to support the incorrect topology in both cases (both p <= 1.08E-6). Conversely, 5'-UTR PIBS provided proportionally more support for the reference relationships (both p <= 2.12E-11).

(Full color requested)

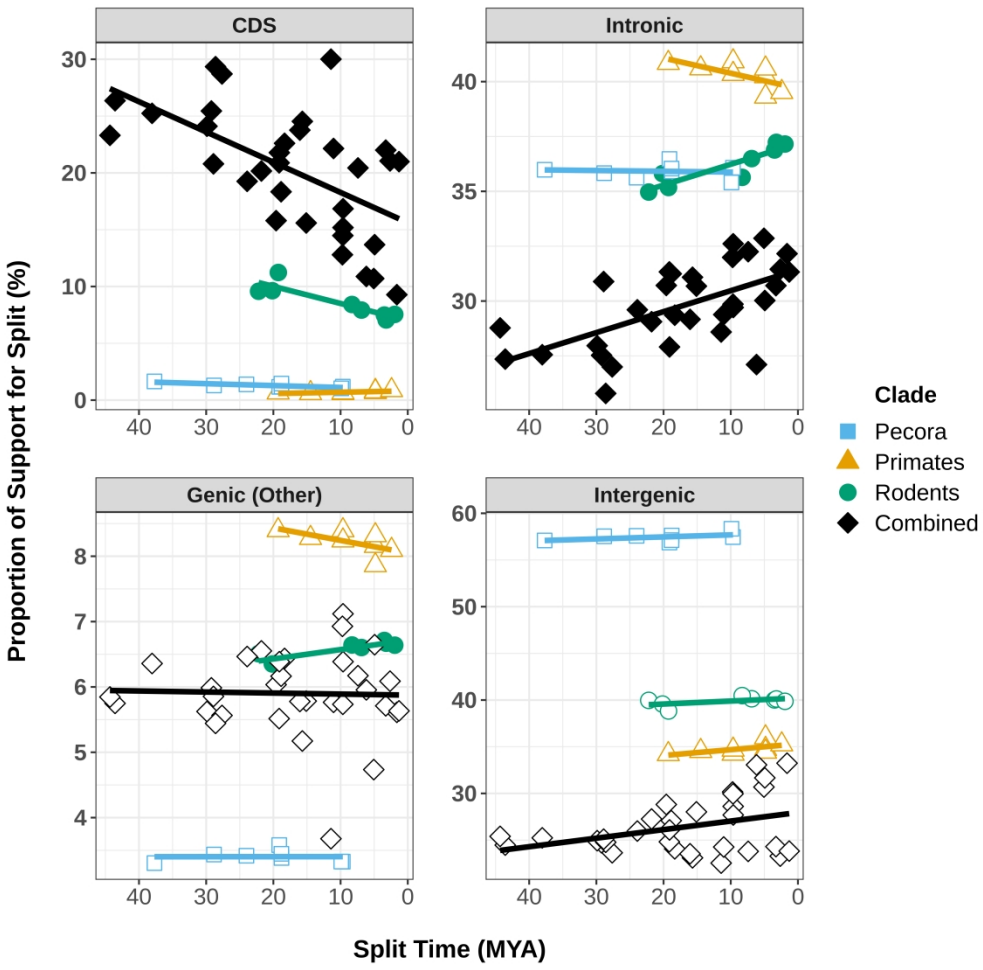


Figure 3: Changes in phylogenetic utility over time among locus types. Based on divergence times estimated from SISRS orthologs (displayed here) as well as dates from the TimeTree database, we ran linear regression analyses to determine whether the proportion of parsimony-informative biallelic sites [PIBS] from different locus types changed in their phylogenetic utility over time. Filled shaped indicate locus types where PIBS inform disproportionately on older or more recent splits. Among rodents and in the combined analysis, CDS-derived PIBS (upper left) provided proportionally more support for older splits (both p <= 1.08E-6), while conversely and for the same groups, intron-derived PIBS (upper right) informed disproportionately about younger splits (both p <= 2.34E-3). Sites from genes that were not annotated as CDS, UTR, or intron ('Genic (Other)'; lower left) show a weaker trend towards increased utility at younger nodes in rodents (p = 3.77E-3), but the relationship is not significant when using dates from TimeTree (p=.113). No other locus type, including intergenic/unannotated sites (lower right), displayed any time-dependent shifts in phylogenetic support.

(Full color requested)