

Title page

Title: Genomic characterization and curation of UCEs improves species tree reconstruction.

Running head: Genomic characterization and curation of UCEs improves species trees

Authors: Matthew H. Van Dam^{1,2*}, James B. Henderson², Lauren Esposito^{1,2},
Michelle Trautwein^{1,2}

¹*Entomology Department, Institute for Biodiversity Science and Sustainability, California Academy of Sciences, 55 Music Concourse Dr., San Francisco, CA, 94118 USA*

²*Center for Comparative Genomics, Institute for Biodiversity Science and Sustainability, California Academy of Sciences, 55 Music Concourse Dr., San Francisco, CA, 94118 USA*

**Correspondence to be sent to: Matthew H. Van Dam, Entomology Department, Institute for Biodiversity Science and Sustainability, California Academy of Sciences, 55 Music Concourse Dr., San Francisco, CA, 94118 USA; matthewhvandam@gmail.com*

ABSTRACT

Ultraconserved genomic elements (UCEs), are generally treated as independent loci in phylogenetic analyses. The identification pipeline for UCE probes is agnostic to genetic identity, only selecting loci that are highly conserved, single copy, without repeats, and of a particular length. Here we characterized UCEs from 12 phylogenomic studies across the animal tree of life, from birds to marine invertebrates. We found that within vertebrate lineages, UCEs are mostly intronic and intergenic, while in invertebrates, the majority are in exons. We then curated 4 different sets of UCE markers by genomic category from 5 different studies including; birds, mammals, fish, Hymenoptera (ants, wasps and bees) and Coleoptera (beetles). Of genes captured by UCEs, we find that many are represented by 2 or more UCEs, corresponding to non-overlapping segments of a single gene. We considered these UCEs to be non-independent, merged all UCEs that belonged to a particular gene, constructed gene and species trees, and then evaluated the subsequent effect of merging co-genic UCEs on gene and species tree reconstruction. Average bootstrap support for merged UCE gene trees were significantly improved across all datasets. Increased loci length appears to drive this increase in bootstrap support. Additionally, we found that gene trees generated from merged UCEs were more accurate than those generated by unmerged and randomly merged UCEs, based on our simulation study. This modest degree of UCE characterization and curation impacts downstream analyses and demonstrates the advantages of incorporating basic genomic characterizations into phylogenomic analyses.

KEYWORDS: Ultraconserved elements, species tree, gene tree, phylogenomic markers, bait capture, ASTRAL, conserved nonexonic elements, exon capture, anchored hybrid enrichment, Hymenoptera, Coleoptera, carangimorph, ants, weevils, mammal, songbird

INTRODUCTION

Phylogenomic methods rely on sampling orthologous loci from the genomes of non-model organisms and then using these loci for the estimation of evolutionary relationships. Commonly used sampling strategies include; ultraconserved genomic elements (UCEs), *sensu* Faircloth et al. 2012, anchored hybrid enrichment (Lemmon et al. 2012), exon capture (Bi et al. 2013), transcriptomes, along with homologous *k*-mer blocks (Sanderson et al. 2017), and conserved nonexonic elements (CNEEs; Edwards et al. 2017). While UCEs and anchored hybrid enrichment markers are generally identified without regard to what genomic class they fall into, transcriptomes, exon capture (Bi et al. 2013) and CNEEs (Edwards et al. 2017) each select for a specific class of marker as their names entail.

UCEs are among the most widely used types of phylogenomic markers, and have been used to resolve both higher level and population level phylogenetic relationships (McCormack et al. 2012, Winker et al. 2018). They are found throughout the tree of life, including Cnidaria, flat worms (Platyhelminthes), arachnids, insects, as well as in birds and mammals (Quattrini et al. 2018, Locke et al. 2018, Faircloth 2017, Van Dam et al. 2018, Moyle et al. 2016, Esselstyn et al. 2017). UCEs are beneficial for their ease of capture from non-model organisms. In addition, UCEs and other probe

based approaches provide access to loci from degraded DNA in museum specimens (Bi et al. 2013, McCormack et al. 2016, Blaimer et al. 2016, Van Dam et al. 2017). The process of identifying UCEs is agnostic to the genetic identity of each UCE and instead simply selects for loci that are highly conserved, single copy, lack repeats (and having a user defined GC content), and are of a particular length (typically 160bp length) (Faircloth et al. 2012; Faircloth 2017).

Independent of phylogenetics, a large volume of research has been conducted on identifying the function of UCEs in the genome and towards understanding why they are highly conserved over many millions of years (Dermitzakis et al. 2003; Bejerano et al. 2004; Sandelin et al. 2004; Woolfe et al. 2004; Vavouri et al. 2007; McCole et al. 2014, 2018; Kushawah and Mishra 2017). The sets of UCEs used in phylogenomics have not been formally documented as to their genomic categorization even at the basic level (intronic, exonic and or intergenic) (but see Jarvis et al. 2014). The genomic identity of phylogenetic markers affects how data is treated in analyses (e.g. models of nucleotide substitution rate based on codon position), as well as their potential phylogenetic informativeness (Gilbert et al. 2018). Generally, UCEs are considered primarily non-coding entities and are treated as such in phylogenetic analyses (but see Jarvis et al. 2014, Bossert et al. 2018, Branstetter et al, 2017b).

Like many different classes of phylogenomic markers, UCEs have long been treated as anonymous, independent loci. A markers' independence is relevant to subsequent analyses— particularly in multi-species coalescent tree estimation. Treating non-independent samples as independent violates the assumption of statistical

independence between samples which could bias results by giving more representation to particular gene tree topologies (see Szöllősi et al. 2015).

Here we examined 12 sets of UCEs that have been used for phylogenomic estimation across the tree of life and identify the genetic class of each UCE as intronic, exonic, or intergenic. We examined how these characterizations compare between organismal classes. We considered 5 previous UCE studies in mammals, birds, fish, and insects (Hymenoptera: wasps, ants and bees, and Coleoptera: beetles) more extensively and examined the phylogenetic utility of intergenic and genic UCEs across taxa. Our data exploration revealed that many genes are actually represented by multiple non-overlapping UCEs (referred to throughout as co-genic UCEs in accordance with Scornavacca and Galtier, 2017). Co-genic UCEs that are then used to generate multiple gene trees from the same gene violate the assumption of independence between loci in subsequent multi-species coalescent analyses. We then concatenated (referred to herein as merged) all co-genic UCEs and performed phylogenetic reconstructions of gene and species trees to examine the effect of merging co-genic UCEs on average bootstrap support values (ABS). Finally, we performed a simulation study to test the effects of merging co-genic UCEs on the accuracy of phylogenetic inference of gene trees and species trees.

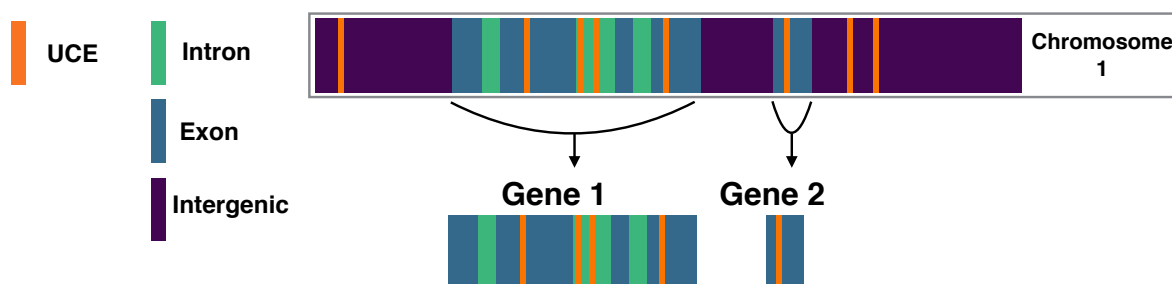
METHODS

Genomic Characterization of UCEs across the Tree of Life

First, we characterized the genomic identity of 12 sets of UCEs representing diverse regions of the tree of life from mammals to marine invertebrates (Figs. 1–2, Table 1). We made our characterizations by blasting (using *blastn* version 2.9.0, Camacho et al. 2008) the base taxon’s probes of each taxon set back to their base genome (the genome to which all other genomes were aligned to in the UCE probe set design). For the tetrapod and acanthomorph fish UCE sets, a base genome was not used in probe design and instead probes were identified by performing an all to all alignment (Faircloth et al. 2012). For these sets, we assigned a new base genome as the taxon for which the most UCE loci were recovered: chicken (*Gallus gallus*) for tetrapods and medaka (*Oryzias latipes*) for the acanthomorph fish probe sets (Faircloth et al. 2012; Faircloth et al. 2013). For all 12 UCE sets, we downloaded each original probe set, identified and extracted the base-genome’s (or assigned base genome’s) specific probes from the total probe fasta file and generated a new fasta containing just the base-genome’s probes. Then we blasted these base genome probes against the base-genome using *blastn* (Camacho et al. 2008). The resulting m8 file from the *blastn* search was then filtered for 100% matches over the 120bp length of the probe.

To identify a UCE’s position within a genome, we identified the UCE’s scaffold and/or chromosome and position from the m8 file. The scaffold and/or chromosome and position information allowed us to search the base-genome’s GFF file to identify overlap between each UCE and particular gene features. The positions of UCEs that fell within introns were inferred from start and stop positions of exons within gene regions. The NCBI records for the specific genomes and GFF files are listed in Supplementary Material S1.

Blast UCE probes against the base genome using blastn

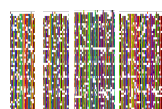


Use the chromosome/scaffold in GFF file to identify which UCEs are found in genes and how many UCEs exist per gene

Merge UCEs found in the same gene



Gene 1: 4 UCE



Gene 2: 1 UCE

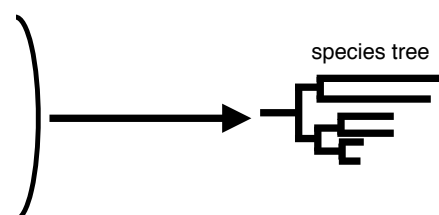
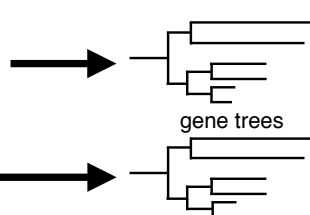


Figure 1. General workflow used to identify co-genic UCEs and merge them for gene tree and species tree reconstruction.

Focal taxa UCE characterization and curation

Data acquisition and alignment

To examine how UCE characterization can affect phylogenetic inference, we more extensively considered UCE data from 5 different phylogenetic studies using 4 different UCE bait sets representing: **weevils** (Coleoptera UCE baits: Faircloth 2017; Van Dam et al. 2017), **ants** (Hymenoptera UCE baits-V2: Branstetter et al. 2017b), **mammals** (Tetrapod 5K-UCE baits: Faircloth et al. 2012; Esselstyn et al. 2017),

songbirds (Tetrapod 5K-UCE baits: Faircloth et al. 2012; Moyle et al. 2016) and carangimorph **fish** (acanthomorph fish UCE baits: Faircloth et al. 2013; Harrington et al. 2016), see Table 2. Using the original data from these studies, we followed their assembly and matrix construction procedures largely using the *PHYLUCE* pipeline (Faircloth et al. 2012; Faircloth 2016).

For three of the datasets (fish, ants and weevils), aligned fasta sequences were already available. For birds and mammals, aligned reads were not available, so we downloaded raw reads and followed the procedures taken by the previous authors to create our aligned matrices. In the case of the mammal dataset (Esselstyn et al. 2017), we downloaded the raw reads and also extracted UCE loci from the same genomes used by the authors, and then combined these data.

For all alignments, we used the *R* package *ips* (Heibl 2008) and removed any ragged ends with the function “*trimEnds*” having a minimum of 4 taxa present in the alignment and filled any gap character “-” with “n” before the first and last non-ambiguous nucleotide.

Curation of genic UCEs

After determining which UCEs were found within genes from our focal UCE sets, we curated sets of genic and intergenic UCEs for weevils, ants, songbirds, mammals and caragimorph fish, (from here on referred to as **Genic** and **Intergenic**). Next, across our taxa sets we identified genes that contain multiple UCEs (co-genic UCEs) (Table 3). Co-genic UCEs were then merged into a single alignment per gene using scripts we developed (Supplementary Material). We then curated 2 sets of UCEs for each taxon

165 group: one that included all UCEs- called **Unmerged** and another that included all
166 merged co-genic UCEs + all remaining UCEs- called **Merged**.

UCE set	BASE TAXON	EXON	INTRON	EXON/ INTRON	INTER- GENIC	INTER- GENIC/ EXON	TOTAL	CITATION
Tetrapod-5K	<i>Gallus gallus</i>	563	2203	158	2095	4	5023	Faircloth et al. 2012
Acantho- morphs-1K	<i>Oryzias latipes</i>	134	547	28	597	0	1306	Faircloth et. al. 2013
Ostariophy- ns-2.7K	<i>Danio rerio</i>	1125	825	54	92	600	2696	Faircloth et al. 2018
Anthozoan_ V 1-UCE-only	<i>Nematostella vectensis</i>	686	218	0	89	44	1037	Quattrini et al. 2017
Arachnida- 1.1K	<i>Limulus polyphemus</i>	347	235	2	24	0	608	Faircloth 2017
Acari-1.8K	<i>Tetranychus urticae</i>	1342	199	36	3	2	1582	Van Dam et al. 2018
Hymenoptera- v2	<i>Athalia rosae</i>	1170	570	94	306	0	2140	Branstetter et al. 2017b
Coleoptera- 1.1K	<i>Tribolium castaneum</i>	736	228	58	100	0	1122	Faircloth 2017
Lepidoptera- 1.3K	<i>Bombyx mori</i>	937	313	63	25	0	1338	Faircloth 2017
Diptera-2.7K	<i>Aedes aegypti</i>	1766	177	82	17	0	2042	Faircloth 2017
Hemiptera- 2.7K	<i>Diaphorina citri</i>	865	580	104	139	1	1689	Faircloth 2017
Mammals	<i>Homo sapiens</i>	3450	9871	1290	5175	371	20157	Van Dam et al. in prep

167
168 **TABLE 1.** Characterization of 12 UCE probe sets according to the annotated base
169 genome (in the base taxon column) for each set.

170
171
172

TAXA	# of TAXA	Total # of UCEs	Minimum # of TAXA Per Locus /Average # of TAXA
weevils (Van Dam et al. 2017)	48	368	33/38
ants (Branstetter et al. 2017b)	101	2571	10/89
birds (Moyle et al. 2016)	106	4347	79/102
mammals (Esselstyn et al. 2017)	100	3594	70/88
carangimorph fish (Harrington et al. 2016)	55	1014	41/52

Table 2. Taxa column lists the previous studies used here. Second column lists the number of individual taxa used in each study, followed by the total number of UCEs. Last column lists the minimum number of taxa allowed in a UCE locus for it to be included/average number of taxa per UCE locus.

UCE set	Single UCE per gene	Co-genic UCEs per gene
Coleoptera	528	497
Hymenoptera	1024	624
Tetrapods	736	2222
Fish	276	438

Table 3. Count of the number of times single and co-genic UCEs are found in particular datasets. This count includes both exonic and intronic UCEs.

Calculation of distance between UCEs for Gallus gallus

The distances between co-genic UCEs (e.g. Fig. 1, see “Gene 1”) were often found to be thousands to tens of thousands of base pairs long. Merging distantly located regions

of the same genes (which naturally happens in the case of transcriptome data) increases the chance of merging regions that have different recombination histories- which may impact species tree analyses (but see Lanier and Knowles 2012). To understand more about where UCEs occur in a genome in relation to each other, we used the *Gallus gallus* genome (the base genome of the tetrapod-5k-UCE probe set) and custom R script to explore the distance (in base pairs) between co-genic UCEs, as well as the distance to the nearest UCE upstream and downstream from a set of co-genic UCEs (referred to herein as nearest neighbor UCEs). For each set of co-genic UCEs, we used the results of the *blastn* analyses described above to identify where a particular UCE locus was found (position, gene and scaffold/chromosome). We then calculated the distance between co-genic. As the m8 file (results from *blastn*) is already ordered by position along a particular chromosome/scaffold, to find the nearest neighbor distances, we simply found the distance upstream and downstream from the UCEs that bookended a particular set co-genic UCEs.

Species Tree Analyses of curated UCEs

Species Tree Analyses of Genic versus Intergenic UCEs

To evaluate the effects of genomic class on support and topology, we reconstructed species trees based on only **Genic** and only **Intergenic** UCEs for weevils, ants, mammals, songbirds and caragimorph fish. Using *R*/unix scrips modified from Van Dam et al. 2017, we ran a maximum likelihood (ML) analysis in *RAxML* 8.2.11 (Stamatakis 2014) on each individual UCE locus, with the “-f a” options for a rapid bootstrap analysis (100 bootstrap replicates (BS)) and searched for best scoring

tree. A General Time Reversible + gamma (GTRGAMMA) site rate substitution model was used for each locus. We then constructed a **Genic** and **Intergenic** species tree for each taxon set using summary methods implemented in ASTRAL-III (Zhang et al. 2018).

Species Tree Analyses of Merged versus Unmerged UCEs

To evaluate the effects of merging co-genic UCEs, we reconstructed species trees based on **Unmerged** and **Merged** UCEs for weevils, ants, mammals, songbirds and caragimorph fish. For clarity, our 1) **Unmerged** UCE set is based on standard protocols and considers each UCE locus as an independent unit (one UCE locus used to reconstruct a single gene tree) and 2) our **Merged** UCE set includes co-genic UCEs merged together to generate a single gene tree, along with all remaining genic and intergenic UCEs each treated as a single locus.

For each **Unmerged** and **Merged** set of UCEs, we ran a maximum likelihood (ML) analysis in *RAxML* 8.2.11 (Stamatakis 2014) with the “-f a” options for a rapid bootstrap analysis (100 BS replicates) and search for best scoring tree. A General Time Reversible + gamma (GTRGAMMA) site rate substitution model was used for each locus. Next we constructed 2 species trees per taxon group using ASTRAL-III (Zhang et al. 2018) first using the gene trees from all **Unmerged** UCE loci and second, using the gene trees from the **Merged** dataset. We ran ASTRAL with the default settings and performed multi-locus bootstrapping (Seo, 2008).

Effects of characterizing and curating UCEs on bootstrap values and topologies

ABS comparisons

To identify if there was any difference in bootstrap support between the gene trees generated from the **Genic** UCEs versus **Intergenic** UCEs and between the gene trees generated from the **Unmerged genic** UCEs versus the **Merged co-genic** UCEs, we calculated average bootstrap support values for each individual gene tree (see Supplementary Material for R and Python code). Average bootstrap support values (ABS) were calculated using a modified R script from (Borowiec et al. 2015). While non-parametric bootstrapping cannot be taken as a measure of absolute of gene tree estimation error (GTEE), it loosely correlates with it (Efron et al. 1996; Holmes 2003, 2005; Susko 2009; Molloy and Warno 2018).

T-tests and GLM

We performed two-sample t-tests in *R* between the means of the ABS of the **Genic** versus the means of the ABS of the **Intergenic** gene tree sets and between the means of the ABS of for **Merged co-genic** versus the means of the ABS for the **Unmerged genic** gene tree sets for all taxa to see if a UCE's genomic category or if merging co-genic UCEs had a statistically significant effect on ABS values. We then calculated the Cohen's d statistic in *R*. We also investigated the effect that locus length, genomic categorization, and merged or unmerged status had on ABS value using a generalized linear model (GLM).

Comparison of tree topologies and shapes: species trees

To assess the impact of our curation efforts on resulting tree topologies we next calculated the Robinson–Foulds distance (RF-dist) (Robinson and Foulds, 1981) tree distance metrics that rely on the tree topology in the *R* package *Phangorn* (Schliep 2011). In addition, we calculated the KF-distance (KF-dist), (Kuhner and Felsenstein 1994) which measures the sum of squares differences between individual branch lengths.

Comparison of tree topologies and shapes: gene trees

For assessing differences between the gene trees from the **Merged** and **Unmerged** UCE sets, we used two different tree shape metrics from their Laplacian spectrum calculated in the *R* package *RPANDA* (Lewitus and Morlon 2015; Morlon et al. 2016). We selected the skewness (asymmetry) of the spectral density profile and the peakedness (peak height) the largest y-axis value of the spectral density profile (Morlon et al. 2016). Normalized, each of these metrics gives a separate description of tree shape: skewness detects the relative timing between branching events (lower values indicate more branching near the stem of the tree whereas higher values indicate more branching near the tips), peakedness (lower peak height values indicate a more even tree shape, whereas higher values indicate a more ladderized tree shape).

Assessing Species Tree Accuracy for Merged and Unmerged UCEs

To identify if the merging of co-genic UCE loci improves the accuracy of inferred species trees, we conducted a simulation study. In brief, we outline below our procedures based on slight modifications from Mirarab and Warnow, 2015.

Simulation methods for gene trees and species trees

As in Mirarab and Warnow, 2015 and Molloy and Warnow, 2018, we used *SimPhy* (Mallo et al. 2016) to simulate 50 species trees and their constituent 1000 gene trees, each with 50 taxa under high levels of incomplete lineage sorting (see Supplementary Material S1 for specific parameters used in simulations). We chose to simulate under high levels of incomplete lineage sorting because we expect that under these conditions, where topological inference may be at its most ambiguous, we would see performance markedly increase in for the longer “merged” loci we simulated next. Sequences were simulated under these gene trees using Indelible (Fletcher and Yang, 2009). We performed 50 replicates each with 1000 genes, the speciation rate was set to $(1e-07)$ with a tree depth set to 1.0×10^7 generations and a global population size of 2.0×10^5 . Sequence lengths were designated to reflect the lengths of **Merged co-genic** loci constructed from the mammal UCE dataset. Specifically, lengths were drawn from lognormal distributions with the log mean controlled by drawing uniformly between 6.2 – 10.2, which corresponds to lengths between 500 – 26,600 bp, representing the range in length of merged co-genic loci from the mammal dataset.

After creating the 50 sets of loci for each species tree, we then subdivided each of these loci to represent our **Unmerged** dataset. Here we selected loci length from a normal distribution of between 300 – 800bp, with a mean of 500 and standard deviation of 74, also representative of the mammal dataset. This resulted in 546,775 total alignments over our 50 replicates, with on average 10,936 alignments per species tree replicate.

Along with testing the accuracy of species trees based on **Merged** and **Unmerged** loci, we also tested the effects of **Randomly Merged** loci. By building species trees based on **Randomly Merged** loci, we were able to examine whether any accuracy improvements of species trees based on the **Merged** loci set were simply due to longer loci length or shared genealogy. To generate **Randomly Merged** loci, we randomly merged, without replacement, loci from the **Unmerged** dataset. The number of loci that were selected to be merged at any one time was chosen from the distribution of the number of times the **Merged** loci were divided up to generate the **Unmerged** loci set. For a distribution of the loci lengths for each of these three-simulated sets, see Supplementary Material S1.

Gene trees for our simulation study were reconstructed using RAxML under a GTRGAMMA site rate substitution model for each locus. For the species trees based on the 1000 locus datasets, we used ASTRAL-III (Zhang et al. 2018). However, for the larger **Unmerged** (~10,000 locus) datasets we used ASTRAL-MP (Yin et al. 2019) to take advantage of AVX2-CPU and GPU processors so that the runs would finish. We then compared the resulting Robinson–Foulds tree distance between our **Merged**, **Unmerged**, and **Randomly Merged** species trees compared to their associated simulated “true” species trees.

RESULTS

Genomic Characterization of UCEs across the Tree of Life

For the 12 UCE sets examined, roughly 51%, ± 16.8 95%CI, of UCEs were found within exonic regions of the genome (Fig. 2; Table 1). The percentage of UCEs located within

exons varied greatly between organismal Classes; for the Insecta UCEs, 86.5% (Diptera) and 51.2% (Hemiptera) were found in exons, whereas in Vertebrata, the exon percentage varied between 41.7% (Ostariophysan fish) and 10.3% (Acanthomorph fish). The vertebrate sets tended to be composed mostly of non-coding regions compared to the invertebrates. The percentages of UCEs found in intergenic regions of the genome varied between 0.83% (Diptera) to 45.7% (Acanthomorph fish). There were some UCEs that could not be placed in an intron, exon or an intergenic region alone and were found to be spanning any two of these regions (Fig. 2). For some UCE sets, these loci were a relatively major component (see Ostariophysan fish Fig. 2), but for most UCE sets they were a small fraction or absent all together (Fig. 2; Table 1). Also, within the tetrapod UCE probe set, we found several probes that could not be recovered in more recent genome assemblies. Though this affected very few probes, it demonstrates that probe recovery may change as genomic assemblies are improved or updated, and highlights the fact that these UCE annotations are only as good as the annotations of their base genomes.

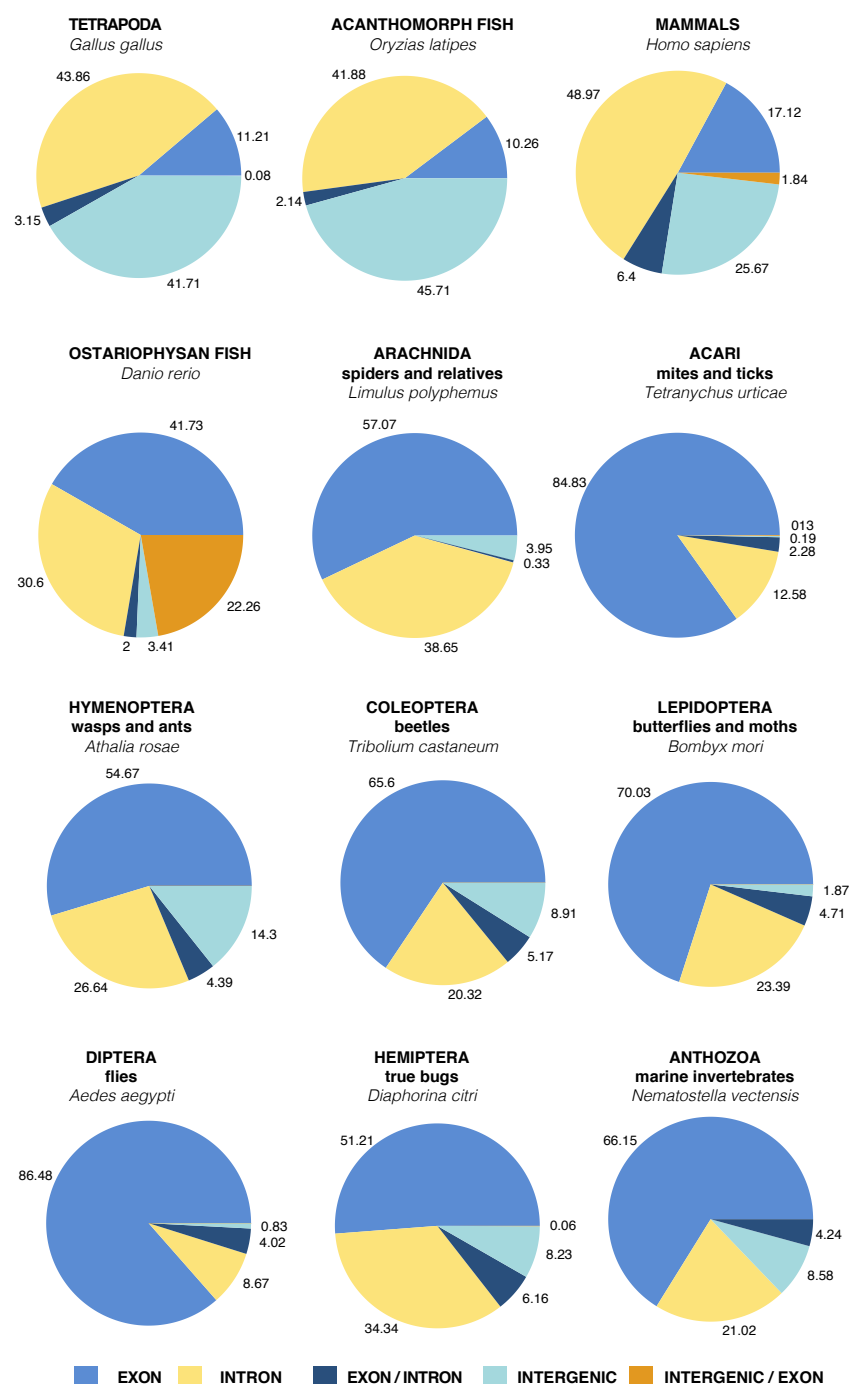


Figure 2. Characterization of 12 UCE probe sets according to the annotated base genome for each set. UCEs are put into 5 different categories: intergenic (not in a gene), exon, intron, exon/intron, and intergenic/exon. The categories exon/intron and

intergenic/exon represents those UCEs that span an intron and an exon or an exon and an intergenic region.

Focal taxa UCE characterization and curation

For weevils, ants, mammals, songbirds and caragimorph fish, we filtered the UCEs located within genic regions of the genome and identified the genes that each UCE represented. We found that within Insecta, between ~52% (weevils) and ~62% (ants) of the UCEs are found as single UCEs in a gene (singleton UCEs) (Fig. 3). Almost as many UCEs belong to genes that are represented by more than one UCE (co-genic) (Table 1, Fig. 3). For the tetrapod UCE set, there are roughly three times as many co-genic UCEs as there are singleton UCEs. Within the caragimorph fish UCE set, ~63% of UCEs are singletons.

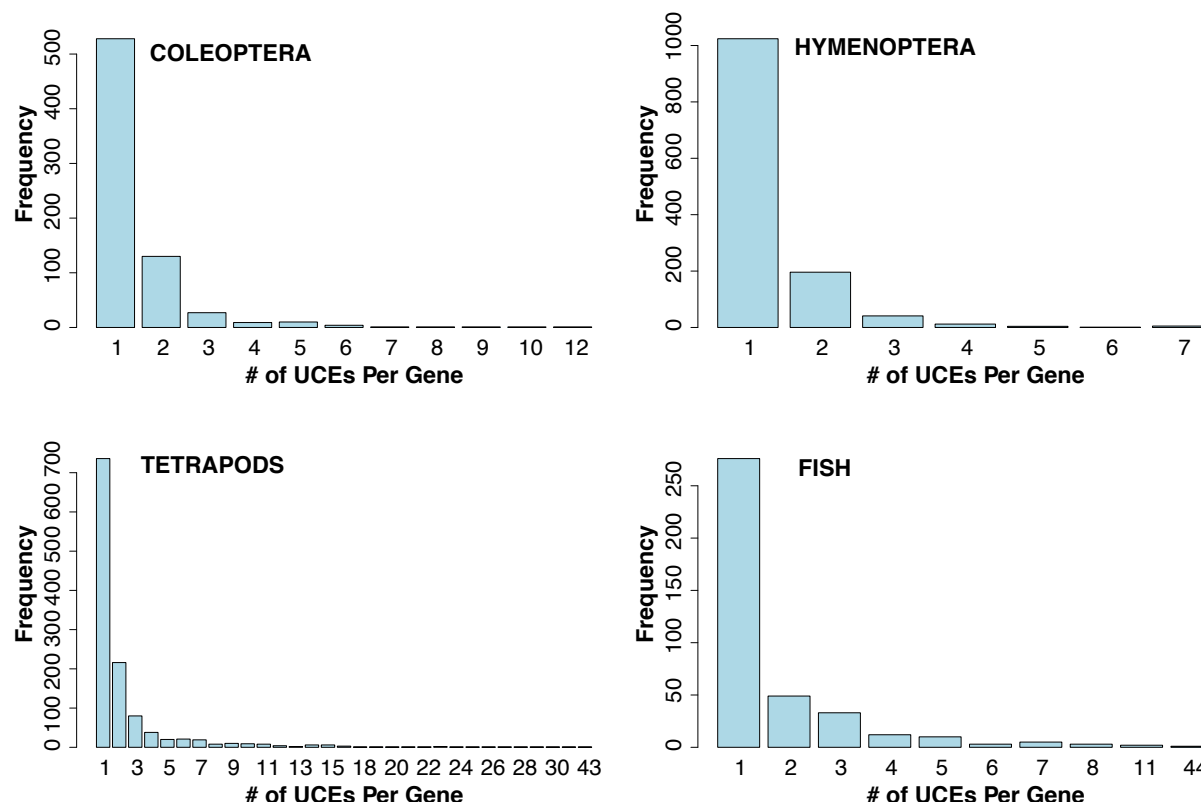


Figure 3. Bar-plot of frequency of the number of UCEs that can be found within a gene, e.g. ~50 separate genes each containing 2 UCEs as seen in the lower right panel. The X-axis labels indicate the number of UCEs per gene.

Calculation of distance between UCEs for Gallus gallus

We found that the majority of UCEs within a gene were clumped and not widely dispersed in comparison to their nearest neighbor distances. The majority of co-genic UCEs were within 20kb to each other, with the average distance of 27kb. While some of the nearest neighbor UCEs were within 10kbp (see Fig. 4), the majority were much farther apart >100kb with the average distance of 458kb.

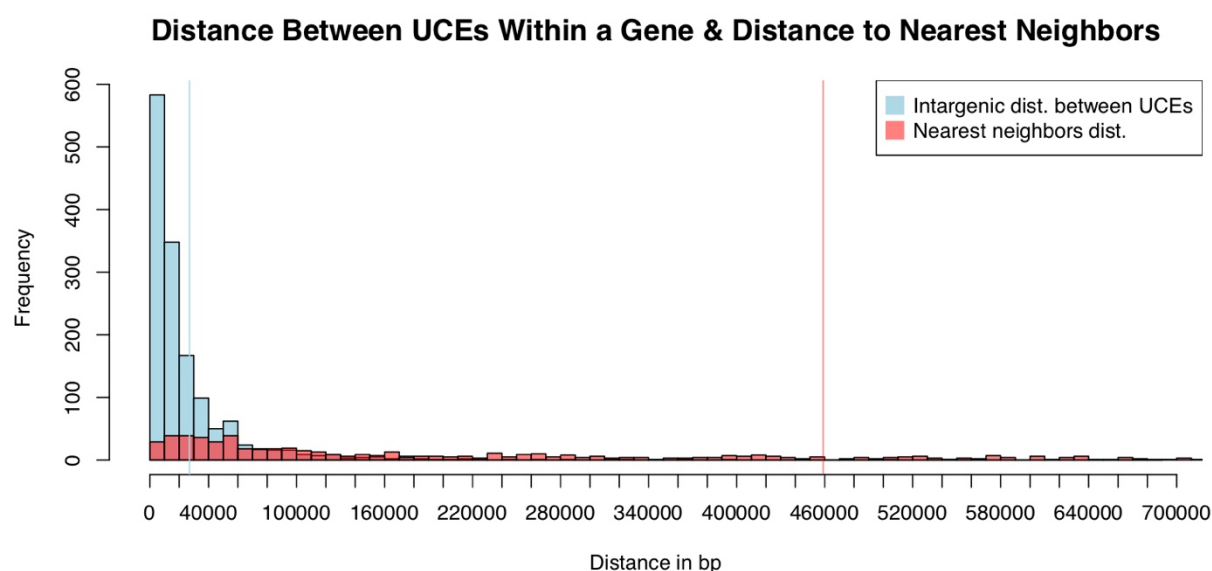


Figure 4. Histogram of distances between UCEs in the 5k-Tetrapod UCE set. UCEs were mapped back to the chicken genome (blastn 100% matches of probe to base over 100% of the length of UCE probe). The distances between co-genic UCEs were then measured (light blue bars), followed by measuring the distance up and downstream from a genic set of UCEs to its nearest neighboring UCE (light red bars). Vertical lines

indicate the mean distance between co-genic UCEs (blue) or mean distance to nearest neighboring UCE (red) for a co-genic set of UCEs. Both estimates are local to a UCE's scaffold/chromosome.

Results of phylogenetic analyses of curated UCEs

Gene Tree Analyses of Genic versus Intergenic UCEs

Our comparison of gene trees based on unmerged **Genic** UCE loci versus those found in the **Intergenic** regions show largely no difference in average bootstrap support (ABS) (Fig. 5; Table 4). In the weevils and fish datasets, there was no significant difference between ABS of intergenic and unmerged genic gene trees, according to a t-test P-value and Cohen's d. In the mammal set, there was significantly more support for the intergenic gene trees according to the t-test P-value but not according to Cohen's d. Ants and birds showed significantly more support for genic gene trees (Table 4), but the Cohen's d shows that there is a very weak difference between the ant datasets and no difference in the bird datasets.

UCE set	Mean ABS INTERGENIC	Mean ABS GENIC	t-test p- value	Cohen's d
weevils	47.96106	48.45831	0.7336	0.1251066
ants	43.17255	47.35368	< 2.2e-16	0.2306506
mammals	56.19731	54.98488	0.0009687	0.06939678
birds	22.98138	24.49267	8.02E-07	0.03047168
fish	29.90741	30.18136	0.6196	0.04413178

Table 4. Means of average bootstrap support values (ABS) per-gene tree for **Intergenic** and **Genic** UCEs across taxa. T-tests show significant differences in ants and birds, yet Cohen's D shows no difference in ants and a weak difference in birds.

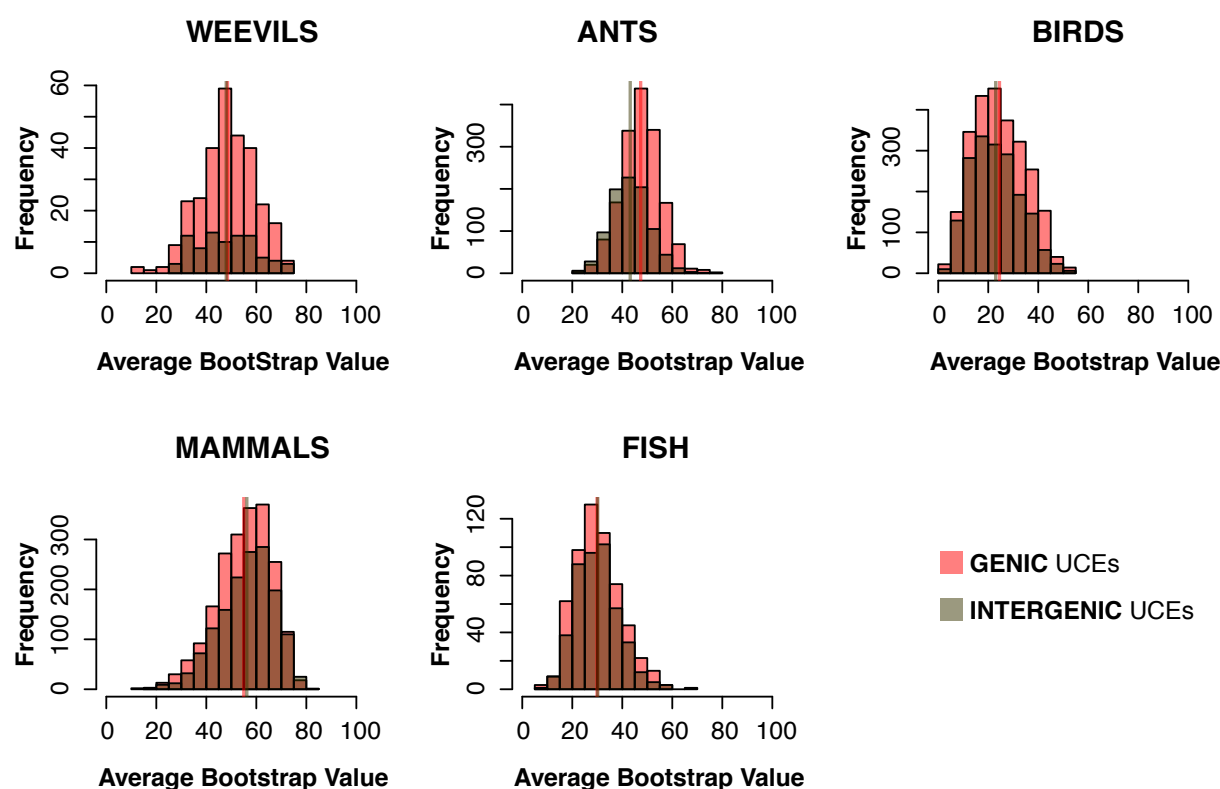


Figure 5. Histograms of average bootstrap support (ABS) values for gene trees based on **Intergenic** UCEs and **Genic** UCEs. Vertical lines represent the mean of ABS values for each set of gene trees (overlapping or adjacent in all but ants). Generally, the distribution of ABS values for gene trees based on **Intergenic** UCEs and **Genic** UCEs are overlapping and largely similar.

Gene Tree and Species Tree Analyses of Merged UCEs versus Unmerged loci

Regarding gene tree analyses across all taxa for **Merged** versus their corresponding **Unmerged** UCEs, we found that the average bootstrap support per-gene tree was significantly higher for **Merged** UCEs than for their corresponding **Unmerged** UCEs (Table 5, Fig. 6) according to both the t-test and Cohen's D. (Note here we are not comparing the full **Merged** and **Unmerged** datasets, but simply the merged co-genic UCEs compared to the same corresponding UCEs, yet unmerged).

UCE set	Mean ABS MERGED	Mean ABS UNMERGED	t-test p- value	Cohen's d
weevils	54.54468	47.32786	0.0007544	0.6321575
ants	57.67213	47.2517	< 2.2e-16	1.27434
mammals	74.11935	55.07754	< 2.2e-16	1.748992
birds	46.13401	23.69862	< 2.2e-16	2.061103
fish	46.30876	29.7862	< 2.2e-16	1.661292

Table 5. Means of average bootstrap support (ABS) values per-gene tree and t-tests between gene trees that were merged when multiple UCEs were found within a single gene, and the same set of UCEs but unmerged and treated as single individual genes. Results show highly significant difference between **Merged** and **Unmerged** gene tree ABS values.

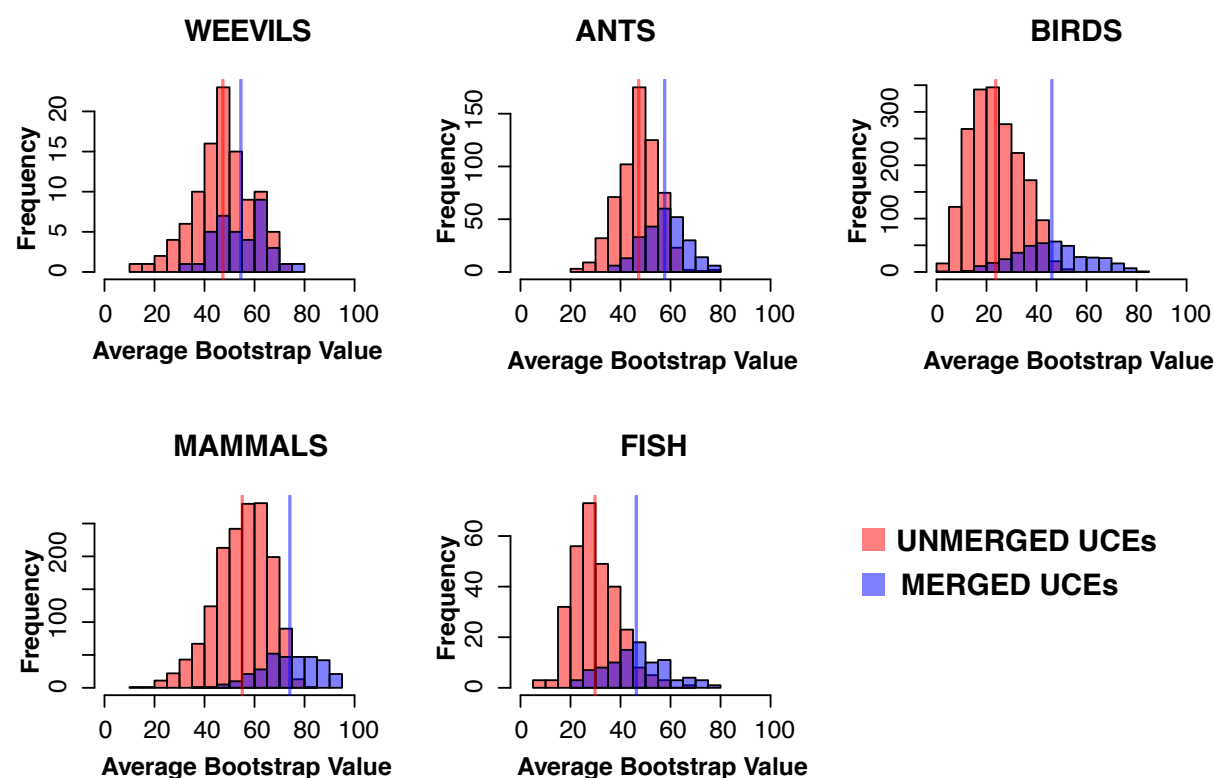


Figure 6. Histograms of average bootstrap support values (ABS) of gene trees generated for **Merged** UCEs and their corresponding **Unmerged** UCEs (at the exclusion of all intergenic and singleton UCEs). Red bars represent ABS values of gene trees for **Unmerged** UCEs, where each individual UCE provides a single gene tree estimate. The purple bars represent the ABS of gene trees generated by **Merged** UCEs, where all UCEs representing a particular gene were merged together to estimate a single gene tree. Vertical lines represent the mean ABS of each UCE set. The distributions are significantly different between **Unmerged** and **Merged** treatments, with the distribution of **Merged** UCE gene trees ABS values shifted towards higher support.

The species tree results from the analyses based on **Merged** UCEs (merged co-genic UCE gene trees + all remaining UCE gene trees) and from the analyses of the

standard UCE treatment (**Unmerged**, one UCE locus - one gene tree) produced similar, but not identical, results in terms of ABS and topologies. Average bootstrap support for the species tree analyses across taxa (Table 6) shows an increase in support for the trees based on **Merged** UCEs (0.29–5.13 ABS support improvement), in all but the fish species tree, which decreased by 0.38 ABS support. Average Astral quartet support values, that measure the conflict between gene trees by node, improve in **Merged** species trees across all taxa. Regions where support was weak in the analyses based on standard UCE treatment remained similar with slight improvement in the **Merged** UCEs analyses; however, these two data treatments often resulted in different topologies. Here we present an annotated topological comparison between the **Unmerged** and the **Merged** species trees for birds (Fig. 7). Comparisons of the topological differences of the remaining **Unmerged** versus **Merged** species trees can be seen in Supplementary Material S1.

UCE set	ABS UNMERGED species tree	ABS MERGED species tree	Quartet Support UNMERGED species tree	Quartet Support MERGED species tree
weevils	90.38	91.11	58.83	59.84
ants	94.1	94.39	61.39	61.98
mammals	93.65	96.23	69.29	71.05
birds	86.1	91.23	50.56	52.65
fish	89.73	89.35	55.40	56.62

Table 6. Comparison between average bootstrap support and Astral quartet support between species trees constructed in ASTRAL of Unmerged and Merged UCE sets.

GLM

We found a general trend that ABS values were highly correlated with locus length (Table 7; Fig. 8). Another influential factor contributing to model fit of the GLM was the merging of co-genic UCEs in all but the weevil dataset. In contrast, including unmerged UCEs as a category did not significantly improve model fit in all but the ant dataset, which showed slight improvement.

GLM variable	Weevils	Ants	Birds	Mammals	Fish
(Intercept)	9.99E-09	< 2E-16	1.50E-07	< 2E-16	< 2E-16
loci_length	0.0118	< 2E-16	< 2E-16	9.23E-13	< 2E-16
Merged_UCE	0.1913	< 2E-16	< 2E-16	0.00912	1.19E-13
Unmerged_genic_UCE	0.8021	0.000114	0.555	0.78102	0.796
loci_length: Merged_UCE	0.0799	< 2E-16	< 2E-16	1.61E-14	1.11E-07
loci_length:					
Unmerged_genic_UCE	0.9534	0.001967	0.280	0.70277	0.185

Table 7. P-Values from GLM analyses. Loci length was consistently found to be a significant influence on average bootstrap support across datasets. Merging of UCEs also strongly influenced loci length and thus had an effect on average bootstrap support. Significant P-values (less than 0.05) are highlighted in bold font.

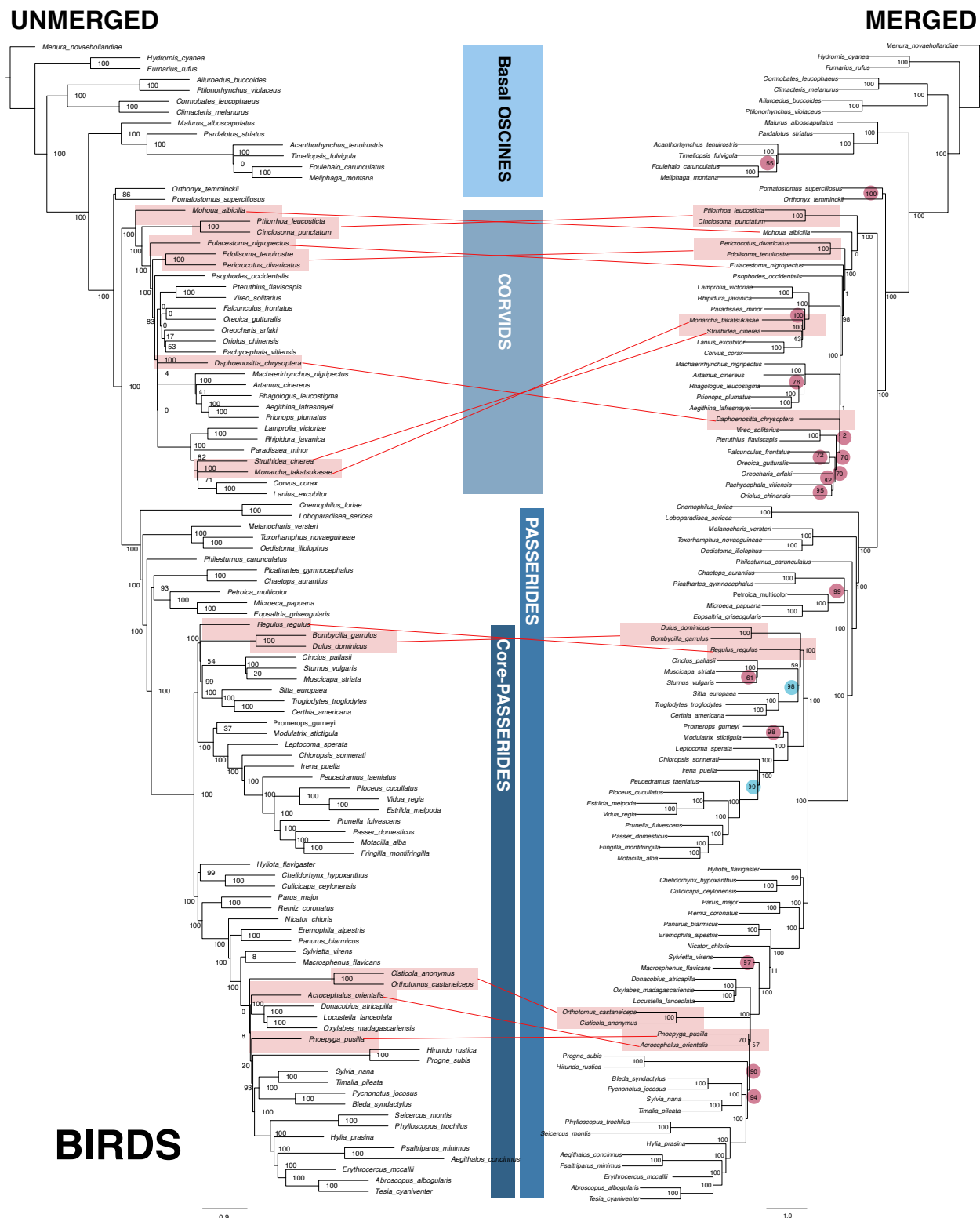


Figure 7. Comparison between the **Unmerged** and **Merged** species tree for birds.

Differences in bootstrap values are indicated on the **Merged** tree (red circles denote

increases, blue—decreases). Bootstrap values are generally higher in the **Merged** species tree and topological shifts exist within corvids and core passerides (shown with red lines).

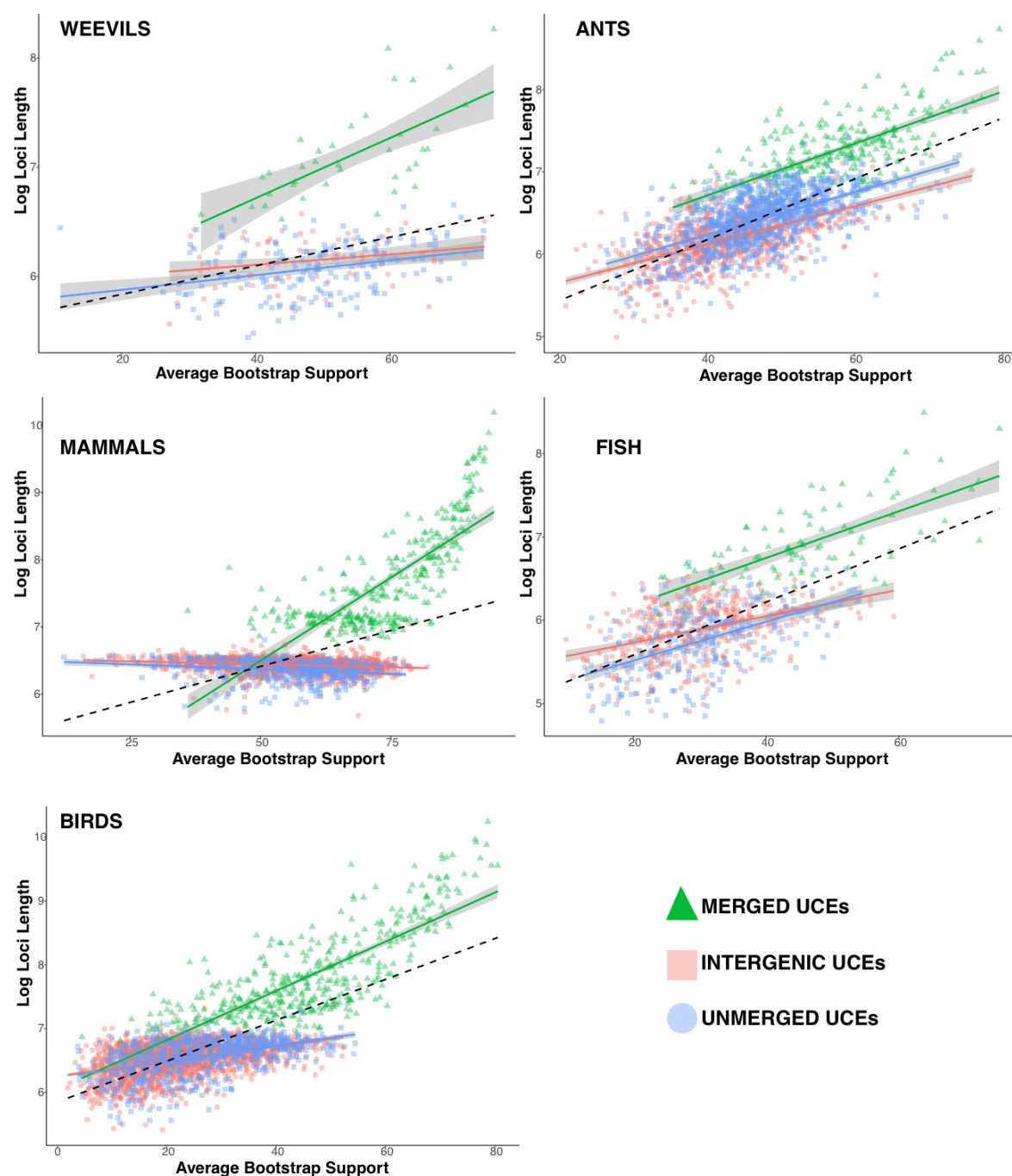


Figure 8. Linear regression of log of UCE length verses average bootstrap support (ABS) values of its corresponding gene tree. Categories include **Intergenic** (all UCES that fell in intergenic regions), **Merged** (all merged co-genic UCES), and unmerged (all remaining genic UCES that were single representatives of single genes). Dashed line represents the correlation of ABS versus log of loci length for all gene trees. Gray regions around regression lines represent 95% confidence interval. ABS values generally increase with increasing loci length.

Comparison of tree metrics

The RF-dist distances and KF-dist distances were computed for the **Genic** versus **Intergenic** species trees and the **Merged** versus **Unmerged** species trees for all taxa (Supplementary Material S1). For **Genic** versus **Intergenic**, RF-dist ranged from 6 (mammals) to 22 (ants), and KF-dist distances ranged from 1.3 (mammals) to 4.4 (ants). For **Merged** versus **Unmerged** species trees, RF distances ranged from 0 (ants) to 20 (birds) and KF-dist ranged from 0.52 (fish) to 1.63 (birds).

The results from the spectral density profiles, including skewness and peakedness (peak height) measures of gene tree shape, both show that on average the distribution **Merged** trees shape is less variable and more similar to one another, whereas their component **Unmerged** gene trees are more widely dispersed (see Fig. 9). In addition, **Merged** versus their component **Unmerged** species trees occupy subtly different regions of trees shape—with most significantly shifted into another region (Supplementary Material S1). These metrics largely indicate that the **Merged** gene trees are converging on a narrower and slightly different region of tree shape. The spectral density profiles for the component **Unmerged** and **Merged** gene trees for mammals are

shown here (Fig. 9), while the remaining taxa can be found in the supplement (Supplementary Material S1).

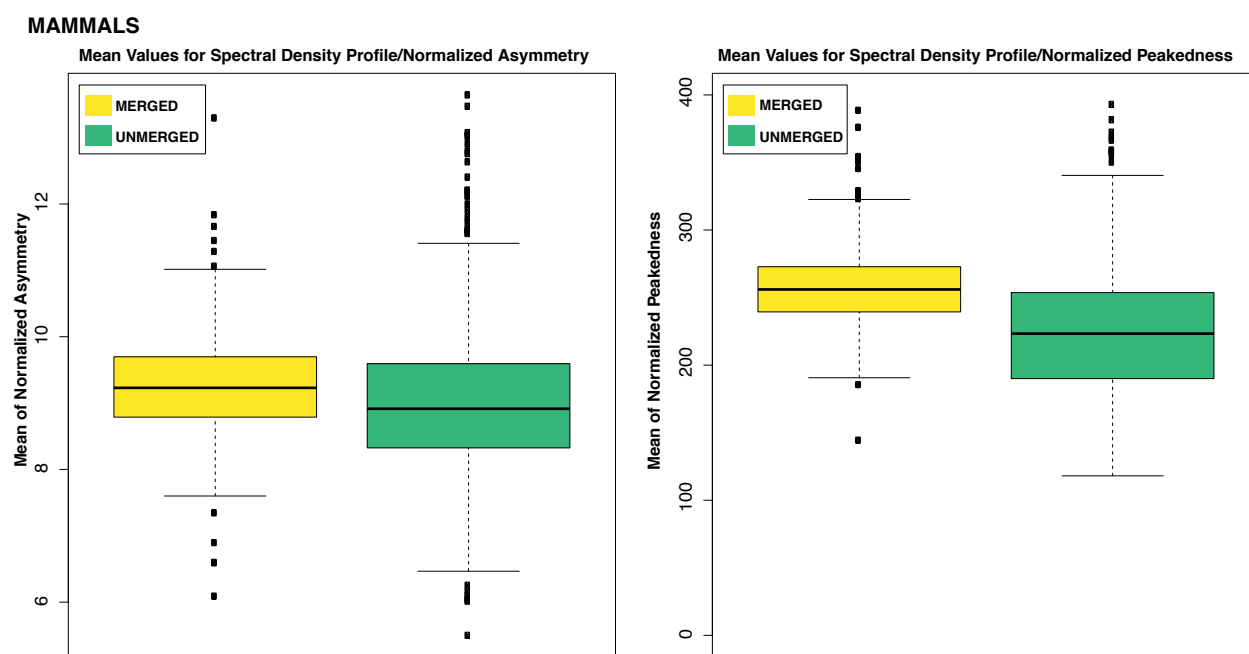


Figure 9. Box plots of normalized Laplacian spectral density profiles, skewness (left panel) and peakedness (right panel) for both **Merged** (yellow) and their component **Unmerged** (green) gene trees from the mammal data set. Both tree shape metrics show that the **Merged** gene trees have a narrower range of values. In addition, their means are slightly shifted into a different region of tree space compared to their component **Unmerged** gene tree spectral density profile metrics.

Assessing Species Tree Accuracy for Merged, Unmerged and Randomly Merged loci

The results of the simulations clearly favored the **Merged** loci (which are essential just long loci) over loci that were **Unmerged** (the long loci broken up). The **Merged** gene trees reconstructed the original (true) simulated species tree in 84% of the simulations,

the **Randomly Merged** 74% of the time, whereas the **Unmerged** gene trees found the original species tree only 2% of the time (Fig. 10). The average RF-dist between the true species tree and the **Merged** species tree set was the smallest (mean RF-dist: 0.4, SD 0.9), slightly higher between the true species tree and the **Randomly Merged** species tree set (mean RF-dist: 0.8, SD 1.8), and greatest between the true and **Unmerged** species trees (mean RF-dist: 88, SD 18).

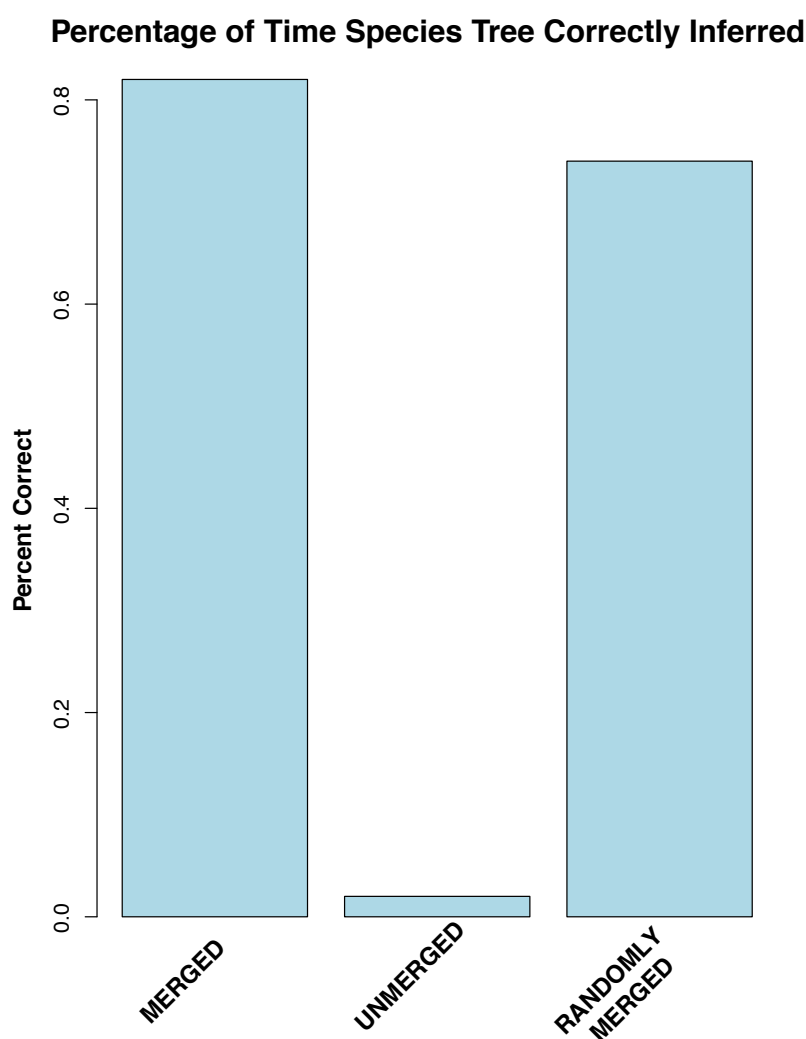


Figure 10. Robinson and Foulds tree distance between the true species tree and the results of the **Merged** (long simulated loci), **Unmerged** (the long loci broken up) and **Randomly Merged** loci (the broken-up parts of the original long loci randomly merged

together). The average RF-distance for the inferred tree to true tree are: 2.2 **Merged**, 90.6 **Unmerged**, 3.1 **Randomly Merged**.

DISCUSSION

Here we characterized UCE sets from a diversity of organisms and found that in each UCE set there are loci from exonic, intronic and intergenic regions. The identification of coding regions in UCEs has been mentioned in previous studies, but not thoroughly explored (Branstetter et al., 2017b, Jarvis et al. 2014). Interestingly, the genomic identity of UCEs appears to vary between invertebrates and vertebrates, with invertebrate UCEs being primarily coding and vertebrate UCEs being mostly non-coding (Fig. 2). It is important to note that our characterizations were based on recent annotations of the base genomes (or assigned base genome) for each UCE set (see Supplementary Material S1). Our categorization of UCEs as exonic relied on gene annotations based on transcriptomes and, in many cases, algorithmic predictions. Thus, we expect that over time some of these UCE characterizations will change in accordance with updated annotations of the base genomes. Also, the annotations for the base genomes for these UCE sets are in varying stages of completeness, yet in most cases more complete than the genomes for other taxa included in these UCE studies. Because we assume orthology of UCEs between taxa within a single study, we expect that the genomic categorization of these UCEs for the base taxon also extends to the UCEs from other taxa (though equally complete and annotated genomes would be required to test this further).

UCEs were first described from the mouse and human genomes as non-coding

regions (Dermitzakis et al. 2003), and so this characterization appears to have then been carried over to all organisms, though it is perhaps only partially true for vertebrates (Fig. 2). An explanation for the contrast between the genomic characterization of vertebrate versus invertebrate UCEs is unclear. The taxa that the invertebrate and vertebrate UCE sets were designed across share common ancestry at a similar age of ~300 million years ago (Bethoux 2009; Smith and Marcot 2015), so the difference is not necessarily driven by the evolutionary age of the lineage. It is possible that the contrasting breakdown of invertebrate and vertebrate UCEs instead relates to the quality of the assembled genomes used in probe design, variation in the probe design pipeline used for vertebrate and invertebrate UCE sets (Faircloth et al. 2012, Faircloth 2017), or more interestingly, genome size and evolution.

Our results on the categorization of vertebrate UCEs are similar in composition to those reported from UCEs identified between mouse and human genomes (McCole et al. 2018). It has been suggested that UCEs play a role in genome stability, as they are enriched in contact domains (McCole et al. 2018), and have been shown to also exhibit elevated synteny (Dimitrieva and Bucher, 2012). In mice and human genomes, boundary regions flanking contact domains, as well as loop anchors, are relatively depleted of UCEs; however, the UCEs that do occur in these regions are disproportionately exonic and play a suggested role in splicing (McCole et al. 2018). It is possible that invertebrate UCEs are more often pulled from boundary regions flanking contact domains and loop anchors to explain their high exonic content.

Genomic categorization is increasingly relevant in phylogenetic analyses. Studies based on coding regions often yield varying results based on how genes are

analyzed, either as amino acids or by nucleotides or by only certain nucleotide positions, sometimes depending on the age of the radiation being assessed. UCEs have been treated as non-coding units even when combined with exons, but see (Jarvis et al. 2014; Bossert et al. 2018). Yet, once coding regions from UCEs are characterized, subsequent analyses can take into consideration results based on amino acids versus nucleotides and specific nucleotide positions.

Additionally, the phylogenetic utility of varying genomic regions has been considered by authors, particularly in respect to bird phylogeny (Jarvis et al 2014; Edwards et al. 2017). Non-coding regions' potential for species tree reconstruction has been highlighted over coding regions because non-coding regions (particularly CNEEs, Conserved Non-Exonic Elements) have less possibility of GC bias and differences in evolutionary rate, as well as an ease of alignment. When we consider the distribution of ABS values from gene trees reconstructed from **Intergenic** versus **Genic** UCEs, we found no general pattern differentiating them (Fig. 5; Table 4). The genic and intergenic gene trees result in a similar, overlapping distributions of bootstrap values across taxa.

When we considered genic UCEs, we not only found that many were exonic, but we also found that genic UCEs often occur in multiples within a single gene or are co-genic. Across taxa, co-genic UCEs were more numerous than UCEs that were single representative of particular genes, except in the Hymenoptera UCE set. Although many UCEs are co-genic, they have been historically treated as independent loci in species tree analyses, thus over-representing particular gene trees in the summary species tree analysis. By merging together all co-genic UCEs, we ameliorate the issue of non-independence (at least in these obvious cases).

After analyzing the gene trees of these merged loci, we found that the ABS of **Merged** gene trees are significantly higher than these same loci **Unmerged** (treated as singletons) (Fig. 6; Table 5). **Merged** and **Unmerged** gene tree topologies varied as well, with the distribution of **Merged** gene tree topologies generally showing less variability and covering a narrower region of tree space according to spectral analyses that consider branch length and tree shape (evenness vs ladderized) (Fig 9, Supplementary Material S1). This suggests that the longer merged loci provide more decisive phylogenetic signal. Also, the ABS of the **Merged** species trees were improved across taxa (excluding the fish dataset, Table 6) and topologies varied from those based on standard protocols (Fig. 7, Supplementary Material S1).

In general, we find that loci length was the predominate driver for the increased ABS based on merging UCEs by gene (Fig. 8; Table 7). Loci length has been found to be highly correlated with phylogenetic information content (Faircloth et al. 2012; Portik et al. 2016; Branstetter et al. 2017a; Edwards et al. 2017; Van Dam et al. 2017). Given the relationship demonstrated between ABS and accuracy of estimated gene trees (Liu et al. 2015; Molloy and Warnow 2018; Zhang et al. 2018), we expected that, broadly speaking, the merging of UCEs from the same genes into longer loci would result in more highly supported gene trees that would improve the accuracy of the resulting species tree. Our simulation study supported this hypothesis. Our investigation of the accuracy between simulated sets of **Merged** (complete, long loci), **Unmerged** (subdivided loci, the length of standard UCEs) and **Randomly Merged** UCEs, decisively favored the merged loci in terms of their ability to recover the correct species tree under high levels of incomplete lineage sorting. This finding indicates that it is not

only loci length that contributes to enhancing bootstrap support values and improving the accuracy of tree inference, it is also the shared phylogenetic signal in the **Merged** loci. **Randomly Merged** loci, however, did outperform **Unmerged** loci, again pointing to the importance of loci length even at the expense of shared phylogenetic history across a locus (Adams and Castoe 2019; Bayzid and Warnow, 2013). Our simulations suggest that using fewer longer loci is preferable to many shorter, less informative loci.

Our results are also in accord with other studies (Edwards et al. 2017; Adams and Castoe 2019) that identify that longer loci with more information are preferable to shorter less informative ones, despite the fact that longer loci increase the probability of spanning recombination blocks. The effect of recombination on summary species tree methods has received recent attention (Lanier and Knowles 2012; Gatesy and Springer 2014; Edwards et al. 2016; Jennings 2017). A series of papers by Gatesy and Springer suggest that recombination misleads species tree methods (Gatesy and Springer 2014, Gatesy and Springer, 2018) and thus the authors advocate for concatenation methods. Yet a simulation study by Lanier and Knowles, 2012, found recombination to not have an overtly negative influence on coalescent-based phylogenetic analyses under high levels of incomplete lineage sorting (though only relatively short loci were considered- and see Gatesy and Springer, 2018).

Another advocated approach to address the issue of recombination in species tree analyses has been to select loci that are separated by an intrachromosomal distance threshold (Jennings 2017) to satisfy the evolutionary independence assumption of coalescent-based phylogenetic methods (Arbogast et al. 2002). In some UCE studies, UCEs within 10k base pairs of each other were discarded to avoid

physical linkage (Faircloth et al. 2013; Alfaro et al. 2018)- a physical recombination distance estimated for fish. However, the accurate estimation of recombination blocks across diverse non-model organisms is currently an unrealistic approach for phylogenomics, considering that recombination rates even within species show substantial heterogeneity across the genomes of individuals (Comeron et al 2012).

Here we did find, as an example, that within the *Gallus gallus* genome, co-genic UCEs were generally clustered together within a distance of 20kb, while the distance between co-genic UCEs and their next neighbor outside of the gene were much longer (>400 kb on average). This clustering suggests that merging multiple sections of the same gene may not be as problematic in regard to chromosomal distance as suggested by Springer and Gatesy (2018). Though 20k base pairs may be longer than some suggested distances (*Drosophila* 12.5 kb: Jennings 2017), it is also shorter than some (*Tiger salamander* 17 kb-1.7 Mb: Jennings 2017). Again, this highlights the ambiguity surrounding the determination of appropriate recombination distances and where they stop and start over potentially millions of years of evolution.

In addition, the “merging” of distinct and perhaps distant regions of a single gene naturally occurs in the production of a transcriptome, and coalescent-based phylogenetic analyses based on transcriptome data are widely used (Lin et al. 2014; Fernández et al. 2018; Esstelystyn et al., 2017; Wipfler et al., 2018). The combined effect of linkage, recombination, and selection on species tree accuracy still remains unclear. In general, the interplay between recombination and selection (e.g. selective sweeps and recombination hotspots) has not been adequately addressed in the systematics debate between recombination and its effect on the accuracy of

multispecies coalescent methods. Yet, as more fully syntenic genomes become available (Dudchenko et al. 2017) this more nuanced investigation can begin.

When multiple UCEs fall within a single gene, treating them as independent units in species tree analyses potentially over-represents a single gene in species tree analyses (but see Scornavacca and Galtier, 2017). The perimeter of this evolutionary unit is dependent on how genes are defined. Here we used the GFF files from well-annotated base genomes that had predefined units termed genes. This process of identifying genes in a genome by default necessitates the importance of high-quality genome annotation and the criteria/methods used in the annotation (e.g. identifying genes from robust transcriptomes and or algorithmically). The chicken genome, for example, was annotated through masking repetitive regions and then using transcripts (cDNA and ESTs) and RNA-Seq data to identify the potential genes, which were then filtered further using standard gene (codon) models (Warren et al. 2016). This high-quality annotation lessens human arbitration in gene definition, as defining genes is primarily based on biological evidence.

UCEs are increasingly important and frequently used in phylogenomics due to their accessibility in specimens of varying quality, relatively low cost, as well as due to the existence of a user-friendly bioinformatics pipeline (Faircloth, et al. 2102). Amongst genomic subsampling methods, though, they also return the shortest loci on average (Karin et al. 2019). The capture of short loci is a benefit when dealing with degraded DNA from old specimens, but short loci are a disadvantage in multispecies coalescent-based phylogenetic analyses. Short loci can also contribute to gene tree estimation error, which in turn hampers species tree inference. When using species tree methods

based on UCE data, merging co-genic UCEs may help reduce the negative impacts of uninformative loci and result in a more highly supported and potentially more accurate phylogenetic estimate.

Acknowledgments: The authors would like to thank Brian Simison and Joe Russack (California Academy of Sciences, CAS) for access to AWS servers. We would also like to thank the CAS Center for Comparative Genomics Reading Group for discussion of this paper. Thanks to the Doolin Foundation for Biodiversity and NSF 1754376, 1856402 for funding this research. We would also like to thank Adam Leaché, Brant Faircloth and the other 3 reviewers for their useful comments that improved the quality of this manuscript.

Supplementary Material

UCE Identification Pipeline: *R* and *Python* code:

<https://github.com/matthewhvandam/integrating-functional-genomics-into-phylogenomics>

Supplementary Material S1: tree comparison metrics and reference tables: link pending submission:

References

Adams R.H., Castoe T.A. 2019. Statistical binning leads to profound model violation due to gene tree error incurred by trying to avoid gene tree error. *Molecular Phylogenetics and Evolution*. 134:164–171.

701 Alfaro M.E., Faircloth B.C., Harrington R.C., Sorenson L., Friedman M., Thacker C.E.,
702 Oliveros C.H., Černý D., Near T.J. 2018. Explosive diversification of marine fishes at the
703 Cretaceous-Paleogene boundary. *Nature Ecology & Evolution* 2: 688–696.

704 Arbogast B., Edwards S.V., Wakeley J., Beerli P., Slowinski J.B. 2002. Estimating
705 divergence times from molecular data on phylogenetic and population genetic time
706 scales. *Annu. Rev. Ecol. Syst.*, 33:707–740.

707 Bayzid M.S. and Warnow, T. 2013. Naive binning improves phylogenomic
708 analyses. *Bioinformatics*, 29:2277–2284.

709 Bejerano G., Pheasant M., Makunin I., Stephen S., Kent W.J., Mattick J.S., Haussler D.
710 2004. Ultraconserved Elements in the Human Genome. *Science* (80). 304:1321–1325.

711 Bethoux O. 2009. The earliest beetle identified. *J. Paleontol.* 83:931–937.

712 Bi K., Linderroth T., Vanderpool D., Good J.M., Nielsen R., Moritz C. 2013. Unlocking the
713 vault: next-generation museum population genomics. *Mol. Ecol.* 22:6018–6032.

714 Blaimer B.B., Lloyd M.W., Guillory W.X., Brady S.G. 2016. Sequence Capture and
715 Phylogenetic Utility of Genomic Ultraconserved Elements Obtained from Pinned Insect
716 Specimens. *PLoS One.* 11: e0161531.

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

720 Bossert S., Murray E.A., Almeida E.A.B., Brady S.G., Blaimer B.B., Danforth B.N. 2018.

721 Combining transcriptomes and ultraconserved elements to illuminate the phylogeny of
722 Apidae. *Molecular Phylogenetics and Evolution*. 2018

723 Branstetter M.G., Danforth B.N., Pitts J.P., Faircloth B.C., Ward P.S., Buffington M.L.,
724 Gates M.W., Kula R.R., Brady S.G. 2017a. Phylogenomic Analysis of Ants, Bees and
725 Stinging Wasps: Improved Taxon Sampling Enhances Understanding of Hymenopteran
726 Evolution. *Curr Biol*. 27:1019–1025.

727 Branstetter M.G., Longino J.T., Ward P.S., Faircloth B.C. 2017b. Enriching the ant tree
728 of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other
729 Hymenoptera. *Methods Ecol. Evol.* 8:768–776.

730 Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden
731 T.L. 2008. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

732 Comeron, J.M., Ratnappan, R. and Bailin, S., 2012. The many landscapes of
733 recombination in *Drosophila melanogaster*. *PLoS genetics*, 8(10), p.e1002905.),

734 Dermitzakis E.T., Reymond A., Scamuffa N., Ucla C., Kirkness E., Rossier C.,
735 Antonarakis S.E. 2003. Evolutionary discrimination of mammalian conserved non-genic
736 sequences (CNGs). *Science*. 302(5647):1033–1035.

737 Dimitrieva S. and Bucher P. 2012. Genomic context analysis reveals dense interaction
738 network between vertebrate ultraconserved non-coding elements. *Bioinformatics*.
739 28:395–401.

740 Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C.,
741 Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P. and Aiden, E.L., 2017. De novo
742 assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length
743 scaffolds. *Science*. 356:92–95.

744

745 Edwards S.V., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B.,
746 Wu S., Lemmon E.M., Lemmon A.R. 2016. Implementing and testing the multispecies
747 coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics*
748 and *Evolution*. 94:447–462.

749 Edwards S.V., Cloutier A., Baker A.J. 2017. Conserved Nonexonic Elements: A Novel
750 Class of Marker for Phylogenomics. *Syst. Biol.* 66:1028–1044.

751 Efron B., Halloran E., Holmes S. 1996. Bootstrap confidence levels for phylogenetic
752 trees. *Proc. Natl. Acad. Sci. USA*. 93:13429–13429.

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

756 Faircloth B.C. 2016. PHYLUCE is a software package for the analysis of conserved
757 genomic loci. *Bioinformatics*. 32:786–788.

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

760 Faircloth B.C., Branstetter M.G., White N.D., Brady S.G. 2015. Target enrichment of
761 ultraconserved elements from arthropods provides a genomic perspective on
762 relationships among Hymenoptera. *Mol. Ecol. Resour.* 15:489–501.

763 Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn
764 T.C. 2012. Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning
765 Multiple Evolutionary Timescales. *Syst. Biol.* 61:717–726.

766 Faircloth B.C., Sorenson L., Santini F., Alfaro M.E. 2013. A Phylogenomic Perspective
767 on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of
768 Ultraconserved Elements (UCEs). *PLoS One.* 8:e65923.

769 Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the
770 bootstrap. *Evolution.* 39:783–791.

771 Fernández R., Kallal R.J., Dimitrov D., Ballesteros J.A., Arnedo M.A., Giribet G.,
772 Hormiga G. 2018. Phylogenomics, Diversification Dynamics, and Comparative
773 Transcriptomics across the Spider Tree of Life. *Current Biology.* 28:1489-1497.

774 Gatesy J., Springer M.S. 2014. Phylogenetic analysis at deep timescales: Unreliable
775 gene trees, bypassed hidden support, and the coalescence/concatalescence
776 conundrum. *Molecular Phylogenetics and Evolution.* 80:231–266.

777 Gilbert P.S., Wu J., Simon M.W., Sinsheimer J.S., Alfaro M.E. 2018. Filtering nucleotide
778 sites by phylogenetic signal to noise ratio increases confidence in the Neoaves
779 phylogeny generated from ultraconserved elements. *Molecular Phylogenetics and*
780 *Evolution.* 126:116–128.

781 Harrington R.C., Faircloth B.C., Eytan R.I., Smith W.L., Near T.J., Alfaro M.E., Friedman
782 M. 2016. Phylogenomic analysis of carangimorph fishes reveals flatfish asymmetry
783 arose in a blink of the evolutionary eye. *BMC Evol. Biol.* 16:224.

784 Heibl C. PHYLOCH: R language tree plotting tools and interfaces to diverse
785 phylogenetic software packages. Available from
786 <http://www.christopheibl.de/Rpackages.html>.

787 Holmes S. 2003. Bootstrapping phylogenetic trees: theory and methods. *Stat. Sci.*
788 18:241–255.

789 Holmes S. 2005. Statistical approach to tests involving phylogenies. *Mathematics of*
790 *Evolution and Phylogeny* (Gascuel O. ed.). Oxford, UK: Oxford University Press, pp.
791 91–120.

792 Jennings W.B. 2017. On the independent gene trees assumption in phylogenomic
793 studies. *Molecular Ecology*. 26:4862–4871.

794 Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C.,
795 Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H.,
796 Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V.,
797 Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B.,
798 Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W.,
799 Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H.,
800 Brumfield R.T., Mello C. V., Lovell P. V., Wirthlin M., Schneider M.P.C., Prosdocimi F.,
801 Samaniego J.A., Velazquez A.M.V., Alfaro-Núñez A., Campos P.F., Petersen B.,

802 Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q.,
803 Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K.,
804 Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun
805 M., Fjeldsa J., Orlando L., Barker F.K., Jönsson K.A., Johnson W., Koepfli K.-P.,
806 O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C.,
807 McCormack J., Burt D., Ellegren H., Alström P., Edwards S. V., Stamatakis A., Mindell
808 D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014.
809 Whole-genome analyses resolve early branches in the tree of life of modern birds.
810 Science (80). 346.

811 Karin B.R., Gamble T., Jackman T.R. 2019. Optimizing Phylogenomics with Rapidly
812 Evolving Long Exons: Comparison with Anchored Hybrid Enrichment and
813 Ultraconserved Elements. bioRxiv 672238; doi: <https://doi.org/10.1101/672238>

814 Kuhner M.K., Felsenstein J. 1994. A simulation comparison of phylogeny algorithms
815 under equal and unequal evolutionary rates. Molecular Biology and Evolution. 11:459–
816 468.

817 Kushawah G., Mishra R.K. 2017. Ultraconserved Sequences Associated with HoxD
818 Cluster Have Strong Repression Activity. Genome Biol. Evol. 9:2134–2139.

819 Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. PartitionFinder: Combined
820 Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses.
821 Mol. Biol. Evol. 29:1695–1701.

822 Lanfear R., Frandsen P.B., Wright A.M., Senfeld T., Calcott B. 2016. PartitionFinder 2:
823 new methods for selecting partitioned models of evolution for molecular and
824 morphological phylogenetic analyses. *Molecular biology and evolution*. DOI:
825 [dx.doi.org/10.1093/molbev/msw260](https://doi.org/10.1093/molbev/msw260)

826 Lanier H.C., Knowles L.L. 2012. Is recombination a problem for species-tree analyses?
827 *Syst. Biol.* 61:691–701.

828 Lewitus E., Morlon H., Characterizing and Comparing Phylogenies from their Laplacian
829 Spectrum. *Syst. Biol.* 65:495–507.

830 Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored Hybrid Enrichment for
831 Massively High-Throughput Phylogenomics. *Syst. Biol.* 61:727–744.

832 Lin G.H., Wang K., Deng X.G., Nevo E., Zhao F., Su J.P., Guo S.C., Zhang T.Z., Zhao
833 H. 2014. Transcriptome sequencing and phylogenomic resolution within Spalacidae
834 (Rodentia). *BMC Genomics* 15:32.

835 Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015. Estimating phylogenetic trees from
836 genome-scale data. *Ann. N. Y. Acad. Sci.* 1360:36–53.

837 Locke S.A., Van Dam A.R., Caffara M., Pinto H.A., Lopez-Hernandez D., Blonar C.
838 2018. Nuclear and mitochondrial phylogenomics of the Diplostomoidea and
839 Diplostomida (Digenea, Platyhelminthes) bioRxiv 333518; doi:
840 <https://doi.org/10.1101/333518>

841 Mallo D., De Oliveira Martins L., Posada D. 2016. *SimPhy*: Phylogenomic Simulation of
842 Gene, Locus, and Species Trees. *Syst. Biol.* 65:334–344.

843 McCole R.B., Erceg J., Saylor W., Wu C. 2018. Ultraconserved Elements Occupy
844 Specific Arenas of Three-Dimensional Mammalian Genome Organization. *Cell Rep.*
845 24:479–488.

846 McCole R.B., Fonseka C.Y., Koren A., Wu C. -tin. 2014. Abnormal Dosage of
847 Ultraconserved Elements Is Highly Disfavored in Healthy Cells but Not Cancer Cells.
848 *PLoS Genet.* 10:e1004646.

849 McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield R.T., Glenn
850 T.C. 2012. Ultraconserved Elements Are Novel Phylogenomic Markers that Resolve
851 Placental Mammal Phylogeny when Combined with Species Tree Analysis. *Genome*
852 *Res.* 22:746–754.

853 Molloy E.K., Warnow T. 2018. To Include or Not to Include: The Impact of Gene
854 Filtering on Species Tree Estimation Methods. *Syst. Biol.* 67:285–303.

855 Morlon H, Lewitus E, Condamine FL, Manceau M., Clavel J., Drury J., 2016. RPANDA:
856 an R package for macroevolutionary analyses on phylogenetic trees. *Methods in*
857 *Ecology and Evolution*, 7:589–597.

858 Moyle R.G., Oliveros C.H., Andersen M.J., Hosner P.A., Benz B.W., Manthey J.D.,
859 Travers S.L., Brown R.M., Faircloth B.C. 2016. Tectonic collision and uplift of Wallacea
860 triggered the global songbird radiation. *Nat. Commun.* 7:12709.

861 Portik D.M., Smith L.L., Bi K. 2016. An evaluation of transcriptome-based exon capture
862 for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order:
863 Anura). *Mol. Ecol. Resour.* 16:1069–1083.

864 Quattrini A.M., Faircloth B.C., Dueñas L.F., Bridge T.C.L., Brugler M.R., Calixto-Botía
865 I.F., DeLeo D.M., Forêt S., Herrera S., Lee S.M.Y., Miller D.J., Prada C., Rádis-Baptista
866 G., Ramírez-Portilla C., Sánchez J.A., Rodríguez E., McFadden C.S. 2018. Universal
867 target-enrichment baits for anthozoan (Cnidaria) phylogenomics: New approaches to
868 long-standing problems. *Molecular Ecology Resources.* 18:281–295.

869 Robinson D., Foulds L. 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53:131–
870 147.

871 Sandelin A., Bailey P., Bruce S., Engström P.G., Klos J.M., Wasserman W.W., Ericson
872 J., Lenhard B. 2004. Arrays of ultraconserved non-coding regions span the loci of key
873 developmental genes in vertebrate genomes. *BMC Genomics.* 5:99.

874 Sanderson M.J., Marius N., McMahon M.M. 2017. Homology-Aware Phylogenomics at
875 Gigabase Scales. *Systematic Biology.* 66: 590–603.

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

877 Scornavacca, C. and Galtier, N., 2017. Incomplete lineage sorting in mammalian
878 phylogenomics. *Systematic Biology.* 66:112–120.

879
880 Seo T. 2008. Calculating Bootstrap Probabilities of Phylogeny Using Multilocus
881 Sequence Data. *Molecular Biology and Evolution.* 25:960–971.

882 Smith D.M., Marcot J.D. 2015. The fossil record and macroevolutionary history of the
883 beetles. *Proc. R. Soc. London B Biol. Sci.* 282:Stamatakis A. 2014. RAxML version 8: a
884 tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*.
885 30:1312–1313.

886 Susko E. 2009. Bootstrap support is not first-order correct. *Systematic Biology*. 58:211–
887 223.

888 Szöllősi G.J., Tannier E., Daubin V., Boussau B. 2015. The Inference of Gene Trees
889 with Species Trees. *Systematic Biology*. 64:42–62.

890 Van Dam M.H., Lam A.W., Sagata K., Gewa B., Laufa R., Balke M., Faircloth B.C.,
891 Riedel A. 2017. Ultraconserved elements (UCEs) resolve the phylogeny of Australasian
892 smurf-weevils. *PLoS One*. 12:e0188044.

893 Van Dam M.H., Trautwein M., Spicer G.S., Esposito L. 2018. Advancing mite
894 phylogenomics: Designing ultraconserved elements for Acari phylogeny. *Molecular*
895 *Ecology Resources*. doi:10.1111/1755-0998.12962

896 Vavouri T., Walter K., Gilks W.R., Lehner B., Elgar G. 2007. Parallel evolution of
897 conserved non-coding elements that target a common set of developmental regulatory
898 genes from worms to humans. *Genome Biol.* 8:R15.

899 Warren W.C., Hillier L.W., Tomlinson C., Minx P., Kremitzki M., Graves T., Markovic C.,
900 Bouk N., Pruitt K.D., Thibaud-Nissen F., Schneider V., Mansour T.A., Brown C.T., Zimin
901 A., Hawken R., Abrahamsen M., Pyrkosz A.B., Morisson M., Fillon V., Vignal A., Chow
902 W., Howe K., Fulton J.E., Miller M.M., Lovell P., Mello C.V., Wirthlin M., Mason A.S.,

903 Kuo R., Burt D.W., Dodgson J.B., Cheng H.H. (2016). A New Chicken Genome
 904 Assembly Provides Insight into Avian Genome Structure. *G3 (Bethesda, Md.)*. 7:109–
 905 117. doi:10.1534/g3.116.035923

906 Winker K., Glenn T.C., Faircloth B.C. 2018. Ultraconserved elements (UCEs) illuminate
 907 the population genomics of a recent, high-latitude avian speciation
 908 event. *PeerJ* 6:e5735 <https://doi.org/10.7717/peerj.5735>

909 Wipfler, B., Letsch, H., Frandsen, P.B., Kapli, P., Mayer, C., Bartel, D., Buckley, T.R.,
 910 Donath, A., Edgerly-Rooks, J.S., Fujita, M. and Liu, S., 2019. Evolutionary history of
 911 Polyneoptera and its implications for our understanding of early winged
 912 insects. *Proceedings of the National Academy of Sciences*. 116:3024–3029.

913 Woolfe A., Goodson M., Goode D.K., Snell P., McEwen G.K., Vavouri T., Smith S.F.,
 914 North P., Callaway H., Kelly K., Walter K., Abnizova I., Gilks W., Edwards Y.J.K., Cooke
 915 J.E., Elgar G. 2004. Highly Conserved Non-Coding Sequences Are Associated with
 916 Vertebrate Development. *PLoS Biol.* 3:e7.

917 Yin J., Zhang C., Mirarab S. 2019. *ASTRAL-MP*: scaling *ASTRAL* to very large datasets
 918 using randomization and parallelization. *Bioinformatics*.
 919 <https://doi.org/10.1093/bioinformatics/btz211>

920 Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. *ASTRAL-III*: polynomial time species
 921 tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*. 19:153.

922

