Running Head: Noncoding markers for building species trees

Conserved Non-exonic Elements: A Novel Class of Marker for Phylogenomics

Scott V. Edwards[1*], Alison Cloutier[1,2,3], Allan J. Baker[2,3¶]

[1]*Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138 USA*

[2]*Department of Natural History, Royal Ontario Museum, Toronto, Ontario, Canada M5S 2C6*

[3]*Department of Ecology and Evolutionary Biology, University of Toronto, Ontario, Canada M5S 3B2*

[¶]Deceased.

*For correspondence: sedwards@fas.harvard.edu

Abstract

Noncoding markers have a particular appeal as tools for phylogenomic analysis because, at least in vertebrates, they appear less subject to strong variation in GC content among lineages. Thus far, ultraconserved elements (UCEs) and introns have been the most widely used noncoding markers. Here we analyze and study the evolutionary properties of a new type of noncoding marker, conserved non-exonic elements (CNEEs), which consists of noncoding elements that are estimated to evolve slower than the neutral rate across a set of species. Although they often include UCEs, CNEEs are distinct from UCEs because they are not ultraconserved, and, most importantly, the core region alone is analyzed, rather than both the core and its flanking regions. Using a data set of 16 birds plus an alligator outgroup, and ~3600 - ~3800 loci per marker type, we found that although CNEEs were less variable than UCEs or introns and in some cases exhibited a slower approach to branch resolution as determined by phylogenomic subsampling, the quality of CNEE alignments was superior to those of the other markers, with fewer gaps and missing species. Phylogenetic resolution using coalescent approaches was comparable among the three marker types, with most nodes being fully and congruently resolved. Comparison of phylogenetic results across the three marker types indicated that one branch, the sister group to the passerine+falcon clade, was resolved differently and with moderate (> 70%) bootstrap support between CNEEs and UCEs or introns. Overall, CNEEs appear to be promising as phylogenomic markers, yielding phylogenetic resolution as high as for UCEs and introns but with fewer gaps, less ambiguity in alignments and with patterns of nucleotide substitution more consistent with the assumptions of commonly used methods of phylogenetic analysis.

Keywords: intron, conserved element, multispecies coalescent, incomplete lineage sorting, biased-gene conversion

As a result of advances in DNA sequencing and phylogenetic theory, as well as broader and more aggressive taxon sampling and access to museum specimens, phylogenetics is undergoing a renaissance. "Phylogenomics", although a term originally coined to denote the increasing need for a phylogenetic perspective when inferring genome function (Eisen 1998), is now meant also to signify the expanded scale in which phylogenetics typically is executed in the era of high-throughput sequencing (Delsuc et al. 2005; Posada 2016). This scaling up has taken two principle forms: increased taxon sampling as a means of producing greater phylogenetic accuracy, and perhaps even more pointedly, increased amounts of sequence data and numbers of loci generated to test a given phylogenetic hypothesis. Many phylogenies now contain hundreds, if not thousands of taxa, although in many cases highly taxon-rich studies still employ a modest number of loci or base pairs in the phylogenetic analysis. Through a variety of next-generation sequencing technologies, systematists now also have access not only to large numbers of loci for phylogenetic analysis but also a wide diversity of genes and noncoding regions for building phylogenetic trees (Bi et al. 2012; Faircloth et al. 2012; Chen et al. 2015). This access to a diversity of loci for building trees has inevitably increased interest in functional ties between phylogeny and genome history, thereby helping re-capture some of the original intent of the term "phylogenomics". For example, comparison of coding regions generated by transcriptomes across species can reveal key events in the history of adaptation of a clade (Pease et al. 2016), and phylogenetic analyses of conserved noncoding elements and transposable elements in vertebrates has yielded insight into major phases of regulatory evolution (Lowe et al. 2011) and sources of genomic innovation, respectively (Novick et al. 2009).

Despite this progress, in many ways systematists are still constrained by technology in their choice of marker loci for building trees, and this constraint has begun to yield cracks in the

vision for phylogenomics going forward (Edwards et al. 2016). For example, transcriptomes are widely used in plant, invertebrate and vertebrate phylogenomics, and with considerable success, in part due to their ease of access in organisms without available genomes and their relative ease of alignment across broad evolutionary distances. Yet, particularly in vertebrate phylogenetics, the deficiencies of coding regions for phylogenetic analysis have long been noted, even in the PCR-era of phylogenetics (Chojnowski et al. 2008; Jarvis et al. 2014). For example, Chojnowski et al. (2008) suggested that introns were superior to coding regions in the phylogenetic analysis of birds in part because of their higher variability. It is also widely recognized that the third positions of codons can become saturated in vertebrate data sets encompassing deeper divergences, sometimes providing unreliable phylogenetic signal. This trend was previously thought to be confined to fast-evolving mitochondrial genes, but is now generally acknowledged for nuclear genes as well, in many cases necessitating removal of 3rd positions of codons or the use of amino acids rather than nucleotides (Cummins and McInerney 2011; Pisani et al. 2015). A compelling example of the challenges of coding regions for phylogenomic analysis has recently been found for birds, where coding regions showed the highest level of among-lineage variation in base composition, resulting in severe challenges for phylogenetic analysis and ultimately yielding gene and species trees with lower congruence than other types of markers (Jarvis et al. 2014). Some of these deficiencies for phylogenetic analysis can be compensated for by improved models of molecular evolution (Philippe et al. 2011; Pisani et al. 2015), partitioning, use of amino acids instead of nucleotides or dropping sites from analysis, yet at the same time there is a clear need for additional kinds of markers that may yield signals more commensurate with the major assumptions of many tree-building algorithms, such as base compositional stationarity.

Ultraconserved elements (UCEs) have also emerged as a major type of marker for phylogenomics, particularly in vertebrates (Faircloth et al. 2012; McCormack et al. 2012; Lemmon and Lemmon 2013; McCormack et al. 2013). These markers, which consist of and whose signal is dominated by the more variable regions flanking highly-conserved core regions, are found throughout vertebrate and other genomes and have a number of features making them attractive for phylogenetics. They are numerous, allowing the accumulation of thousands of markers for a given study, and most importantly, the flanking regions are characterized by high variability, much more so than the conserved regions that are used to identify them. Although this higher variability yields large numbers of informative sites for phylogenetic analysis, it comes at the cost of decreasing reliability of alignments as one moves away from the core, conserved region (Faircloth et al. 2012; McCormack et al. 2013). Perhaps the most useful aspect of UCEs is their convenience: they can be isolated, through hybrid capture or other methods, without knowing anything about the genome of the species under study. In a similar fashion, anchored hybrid enrichment, while not focusing speficially on UCEs, also yields loci easily comparable among genomically novel taxa (Lemmon et al. 2012). Such loci have been readily isolated from hundreds of taxa that are otherwise genomically unstudied. Although many bioinformatics pipelines specifically exclude UCEs that include coding regions, in some studies, UCEs or 'anchored' conserved loci include exons (e.g., Lemmon et al. 2012; Prum et al. 2015). Additionally, in several studies not explicitly focused on UCEs, the more variable introns flanking exons have also been accessed in genomically unstudied species in a way similar to the flanking regions of UCEs, using approaches such as exon-capture or anchored enrichment (Lemmon et al. 2012; Hamilton et al. 2016). The convenience of UCEs, transcriptomes and exon-capture when studying organisms whose genomes are not yet sequenced is a major driving

force of marker choice in phylogenomics today (Edwards et al. 2016). These markers open up vast areas of biodiversity whose genomes have not yet been sequenced, either due to the unavailability of financial resources, small body size (and hence low DNA yield) of the studied organisms, excessively large or complex genomes, or other factors.

Conserved non-exonic elements in phylogenomics

Here we analyze a new type of marker for phylogenomics that appears a promising addition to the systematists' toolkit. Conserved non-exonic elements (CNEEs) are a class of marker that was originally studied in the context of genome function. CNEEs are noncoding regions of the genome that are designated as 'conserved' because they evolve slower than a putatively neutral class of sites in the focal clade of organisms (Fig. 1). They are called non-exonic to distinguish them from exons, which also usually evolve more slowly than neutral regions of the genome. CNEEs are distinct from what are often called "conserved noncoding elements" (CNEs) in that they can encompass regions of the genome that are sometimes transcribed but are not in exons. It is often unclear whether genomic regions are transcribed and made into proteins, and recent work from the ENCODE and other studies suggests that many regions previously thought to not encode proteins may in fact be transcribed and translated (Ji et al. 2015). Like UCEs, CNEEs have in many cases been found to act as regulatory enhancers, recruiting transcription factors to influence the expression of nearby or distant genes (Kvon et al. 2016; Leal and Cohn 2016). Crucially, however (Table 1), CNEEs differ from UCEs in not being "ultra-conserved": whereas the core regions of UCEs are often identified on the basis of >95% or higher sequence identify between genomes, the core regions of CNEEs are designated as conserved only because they evolve more slowly than a putatively neutral rate. As a result, CNEEs often exhibit moderate levels of variability, especially when compared to the core

131 regions of UCEs (Siepel et al. 2005a). This tendency raises the possibility that CNEEs might
132 contain sufficient variability to be useful in phylogenetics, while at the same time exhibiting
133 alignments of a quality that matches or exceeds those of the flanking regions of UCEs or
134 transcriptomes.

135 CNEEs also differ in the means by which they are identified in genomes. UCEs were
136 initially identified using arbitrary thresholds of conservation and minimum length applied to
137 synteny-aware whole genome alignments of a few exemplar taxa; they are often localized in
138 additional genomes by blast searches using previously identified UCEs (McCormack et al. 2012).
139 By contrast, CNEEs are delimited using statistical approaches, such as hidden Markov models
140 (HMMs), wherein the rate of candidate genomic region is compared to the rate at a class of
141 putatively neutral sites (Siepel et al. 2005b, Table 1). These models are usually applied to the
142 entire set of species under analysis (although here we use a hybrid approach in which vertebrate
143 CNEEs previously identified using a set of aligned vertebrate genomes are identified by blast in
144 additional bird genomes). Fourfold degenerate sites of protein-coding genes are the most
145 commonly used class of site to generate a baseline pattern of substitution. (It is reasonable to
146 question whether 4-fold degenerate sites in coding regions are genuinely neutral and alternatives,
147 such as ancient transposable elements that do not appear to have assumed functions, have been
148 used as the putatively neutral class (Siepel et al. 2005a)) Although CNEEs do overlap with the
149 core regions of UCEs in genomes, and include all noncoding UCEs in principle, the use of only
150 the conserved core region of CNEEs distinguishes this class of marker and ensures that the
151 sequences we use here for phylogenetics do not fully overlap with those of UCEs. Additionally,
152 depending on the thresholds used to identify UCEs and tuning parameters of the HMM used to
153 identify CNEEs, some UCEs will not be found in the set of CNEEs identified; this frequently

154 happens when the core regions of UCEs are short, resulting in non-significant log-odds scores,
155 which depend on length, when comparing the likelihood of CNEE sequences on trees with
156 'neutral' or 'conserved' branch lengths, the two states often used in the HMM. Finally, whereas
157 UCEs are by definition found throughout the clade of interest (e.g., vertebrates) -- widespread
158 presence across a clade is one criterion for inclusion in this marker class -- CNEEs can appear at
159 variable nodes in the Tree of Life; the appearance and disappearance of CNEEs has been studied
160 to gain insight into the origin of phenotypic traits (e.g., Lowe et al. 2011; Lowe et al. 2014).

161 In this study, we compare the phylogenetic performance and evolutionary dynamics of
162 three classes of noncoding genomic markers: CNEEs, UCEs, and introns. We focus on
163 noncoding regions because they appear to be promising for vertebrate phylogenetics, and we
164 agree with suggestions that transcriptomes may have undesirable phylogenetic properties,
165 especially at high taxonomic levels. Our questions in analyzing CNEEs in a phylogenetic
166 context include: do CNEEs resolve phylogenetic relationships as well as UCEs or introns? How
167 do the substitution dynamics of CNEEs compare with those of UCEs and introns? Do CNEEs
168 exhibit alignment and evolutionary properties that are desirable for phylogenomic analysis? And
169 finally, how easily can CNEEs be accessed in non-model species, and what sorts of protocols are
170 recommended for their large-scale deployment in phylogenomics?

171 ## METHODS

172 ### *Compiling CNEEs in Genomic Data*

173 We previously explored the use of CNEEs as markers of regulatory evolution in vertebrates
174 (Lowe et al. 2014). As part of an effort to understand the genomic basis of feather evolution
175 using CNEEs, we first aligned 19 vertebrate genomes using BlastZ, MultiZ and chaining of local

alignments (Kent et al. 2003; Schwartz et al. 2003; Blanchette et al. 2004). We then used a hidden a phylogenetic hidden Markov model (HMM, Siepel and Haussler 2004) to determine regions of the genome (both coding and noncoding) that evolved slower than a benchmark set of 4-fold degenerate sites. The phylogeny of the 19 vertebrates used by Lowe et al. (2014) is well known and was assumed as fixed for all genes and genomic regions prior to analysis. This assumption is standard in pipelines for identifying CNEEs; although it ignores the possibility of incomplete lineage sorting (ILS), discordance due to ILS between the local genomic region and the vertebrate species tree we assumed, which has relatively long branches, is likely to be rare. Still, the potential biases incurred by assuming a fixed tree while identifying CNEEs should be explored, since ILS is known to influence parameter estimates of other macroevolutionary phenomena, such as molecular clocks, substitution rates and reconstruction of ancestral sequences (Burbrink and Pyron 2011; Groussin et al. 2015; Mendes and Hahn 2016). Branch lengths of the tree for the neutral class of sites were determined using maximum likelihood to find optimal branch lengths for the set of 4-fold degenerate sites. Conserved sites were defined as those exhibiting a better fit to a tree with branches no greater than 0.3 (30%) of the length of the 4-fold degnereate tree. The HMM had two states, “conserved” and “neutral” and the tuning parameters for the transition rate between states in the HMM were set with an expectation that CNEEs would on average have a length of 45 bp. This protocol yielded a total of 957,409 conserved elements in total, of which 605,756 fulfilled the criteria for a CNEE. Whereas Lowe et al. (2014) used 602,539 CNEEs in their study, we retained 3207 CNEEs that were discarded in Lowe et al. (2014) because they were not assigned to chromosomes in the chicken assembly used (galGal3), making for a starting total of 605,756 CNEEs. For a detailed account of the bioinformatics pipeline by which we initially determined a working set of CNEEs, see

199 Lowe et al. (2014). Candidate CNEEs were filtered from this vertebrate-wide set of 605,746
200 elements referenced on chicken to retain loci ≥ 400 bp in length (n = 6182). We focused on
201 CNEEs ≥ 400 bp long so as to use a set of loci expected to contain at least a moderate number of
202 variable and parsimony-informative sites.

203 We then chose 14 exemplar species from the Avian Phylogenomics Project (Jarvis et al.
204 2014), including chicken, as a test case for phylogenomic analysis (see Supplementary File S1).
205 These species were chosen so as to capture major branches of the avian tree as it is now known,
206 and in some cases pairs of species were chosen to determine if our analyses could recapitulate
207 known or expected relationships (e.g., flamingo and grebe, penguin and loon). This group of 14
208 species also contains clades that are still unresolved or contentious, such as the precise order of
209 the multiple outgroups to passerine birds (Hackett et al. 2008; Jarvis et al. 2014; Prum et al.
210 2015). We also included data from draft genomes of an emu (*Dromaius novaehollandiae*) and
211 Chilean Tinamou (*Nothoprocta perdicaria*) from Baker et al. (2014) so as to explore the
212 hypothesis of ratite paraphyly (Harshman et al. 2008; Phillips et al. 2010; Smith et al. 2013), to
213 make a total of 16 ingroup species. Using an American Alligator (Alligator mississippiensis,
214 Green et al. 2014) genome as an outgroup sequence brought the total taxa used to 17
215 (SupplementaryFile S1). Blastn searches with chicken query CNEE sequences were used against
216 each of the 16 non-chicken target genomes at an e-value cutoff $1e^{-10}$. CNEEs with no missing
217 species were retained (n = 3822), and *de novo* aligned with default global alignment parameters
218 in MAFFT v. 7.245 (Katoh and Standley 2013).

219 Intron alignments were assembled from the Avian Phylogenomics Project data of Jarvis
220 et al. (2014); however, individual introns were used rather than alignments concatenating introns
221 within each protein-coding gene. The SATé-MAFFT alignments provided by Jarvis et al. (2014)

were reduced to the taxon subset of interest and gap-only columns removed. Loci greater than 400 bp in aligned sequence length, including the alligator outgroup sequence and with no more than 3 missing species were retained (n = 3733). It is noteworthy that it was straightforward to compile ~3700 fully populated CNEE alignments of 400 bp or greater, whereas there were only 998 (26.7%) fully-populated orthologous introns from birds available; we will return to this point in the discussion. Orthologous sequences from Emu and Chilean Tinamou were identified with blastn searches against draft genome assemblies for these species with chicken, ostrich, and white-throated tinamou queries, and were profile-aligned to the existing Jarvis et al. (2014) alignment with MAFFT. Because SATé-MAFFT yields relatively gappy alignments that are nonetheless "better" than MAFFT-only alignments by some optimality criteria (B. Faircloth, pers. comm.), comparing alignment statistics using SATé-MAFFT and MAFFT may bias the results. We therefore applied both SATé-MAFFT and MAFFT to all three marker types to enable side-by-side comparisons. For the CNEE alignments, we recapitulated the precise SATé-MAFFT alignment protocol of Jarvis et al. {, 2014 #4295), including post-alignment trimming with their custom python script 'filter_alignment_fasta_v1.3B.pl', except that we used SATe v. 2.2.7 with MAFFT v. 6.717 (Jarvis et al. used SATé 2.1.0 and MAFFT 6.860b). Ultraconserved elements (UCEs: n = 3679, representing the full set from Jarvis et al. 2014) were compiled as described for introns. There was a higher number of fully-populated alignments for UCEs of 400 bp or greater (n = 3669; 99.7%) than for introns.

As expected, there was overlap between our sets of CNEEs, UCEs and introns. For example, 1497 (39.2%) CNEEs overlapped at least one UCE. The degree of overlap between introns and the other two data sets was much lower: there were 6 introns overlapping CNEEs and 3 introns overlapping UCEs (both < 0.2%). Because UCE loci typically include the conserved

core region in addition to the flanking regions, this overlap could lead to non-independence of our analyses. Therefore, in addition to analyzing the full set of UCEs and CNEEs, we also analyzed non-overlapping data sets of CNEEs and UCEs; in general, we found that our results held for overlapping and non-overlapping subsets of data, and we suggest that even if our CNEE and UCE data sets overlapped completely, analyzing just the core or flanking regions alone would help clarify the difference in dynamics and performance between these genomic regions.

Measures of alignment quality and substitution dynamics

Alignment lengths, proportions of variable and parsimony informative sites, GC content, and the amount of missing data per alignment matrix (here defined as the number of gaps and uncalled bases per total cells in the nucleotide matrix) were calculated with AMAS (Borowiec 2016). Average pairwise nucleotide identity between species within each locus, and the proportion of gaps per base pair of aligned sequence were calculated with custom Perl scripts. Unlike the AMAS calculation of missing data, gaps per bp aligned considers only genuine gap characters (ignoring uncalled bases) and excludes leading and trailing gaps as well as gaps adjacent to uncalled bases; it is equivalent to internal gaps in the alignment per total called bases. TrimAl v. 1.2rev59 (Capella-Gutiérrez et al. 2009) was used for column-based alignment filtering, with the 'automated1' option to choose trimming parameters heuristically based on input alignment characteristics. We recognize that TrimAl and other alignment trimmers may not necessarily improve phylogenetic analysis in some cases (Tan et al. 2015), but we use them here strictly as a standard metric for comparing alignment "quality", without subsequent phylogenetic analysis on the trimmed alignments. Additionally, we note that in many of the analyses in Tan et al. (2015), alignment trimmers performed marginally worse only under unsustainably high levels of trimming. Model-averaged transition/transversion rate ratios (Ti/Tv), the proportion of invariant

sites when appropriate and the gamma shape parameter (α) were estimated for each alignment with jModelTest v. 2.1.7 (Darriba et al. 2012). jModelTest runs included six substitution models (JC, F81, K80, HKY, SYM, and GTR), with invariant sites and unequal base frequencies allowed and rate variation modeled with 4 gamma categories.

Phylogenetic analyses and measures

RAxML v. 8.1.4 (Stamatakis 2014) was used to construct 200 bootstrap replicate gene trees from each unpartitioned alignment for each locus with a GTR + Γ substitution model; these were rooted with the Alligator outgroup in DendroPy v. 3.12.0 (Sukumaran and Holder 2010; Sukumaran and Holder 2015). MP-EST v. 1.5 (Liu et al. 2010) was used to infer species trees for each marker type from the input set of rooted RAxML bootstrap trees. Each analysis used three full MP-EST runs starting with a different random number seed and 10 independent tree searches within each run. Highest scoring trees from each search were used to build a majority-rule extended (MRE) consensus tree for each MP-EST run using RAxML. Per-site consistency indices (CI) were calculated with PAUP v. 4a149 (Swofford 2002) using the MRE consensus gene tree of the 200 RAxML bootstrap replicates for each locus. We did not compute consistency indices on species trees because gene tree heterogeneity can distort statistics like CI when all gene trees are forced onto a single topology (Mendes and Hahn 2016). Average bootstrap supports are also reported for MRE consensus gene trees.

Phylogenomic subsampling

Phylogenomic subsampling (Edwards 2016) was used to assess the stability of specific clades for different subsets of each of the CNEE, intron and UCE data sets. Data sets of increasing numbers of loci (n = 50, 100, 200, 300, 400, 500, 1000, 1500, 2000, 2500, 3000, and 3500 loci)

were built by sampling loci with replacement from within each marker type, and repeating the process to generate 10 independent replicates of a given number of loci within each marker type. MP-EST was then run on each of the 10 replicates as described above, except that only a single MP-EST run (but with 10 independent tree searches) was performed for each replicate. Summary measures are reported by counting the frequency of splits from among the set of MP-EST output trees for each replicate rather than from a consensus tree.

RESULTS

Alignment and variability metrics for non-coding markers in birds

Alignment lengths and variability: Fig. 2 shows the distribution of alignment lengths among the three marker types and the percentage of variable sites within each alignment. With the constraint that each alignment must equal or exceed 400 bp, introns had longer alignments (up to 22,138 bp) than CNEEs (longest alignment, 1829 bp; Fig. 2a-c). UCE alignments based on those of Jarvis et al. (2014) varied from 2,126 – 4279 bp. CNEE alignments exhibit a higher fraction of populated bases per alignment than do introns and UCEs, with 1210 out of 3822 CNEE alignments (31.7%) possessing >99 % of populated bases (Fig. 3a and b). No intron alignments and only a single UCE alignment possessed this high a nucleotide matrix occupancy, whether considering any undetermined base or gaps between called sequence alone (Fig. 3c). CNEEs also exhibited a much lower percent of each alignment that was deemed low quality by trimAl than did introns or UCEs (Fig. 3d, Supplementary File S2. Whereas 1003 out of 3822 CNEE alignments (26.2%) retained >99 % of bases after trimming, only 1 of the UCE alignments and none of the intron alignments retained this much after trimming (Fig. 3d). As expected, both introns and UCEs were more variable than CNEEs (Fig. 2d-f;Supplementary Fig. S1a). The

312 number of parsimony informative sites per alignment varied among markers in a similar way,
313 with CNEEs having the fewest and introns having the most (Supplementary Fig. S1b). The
314 number of variable sites scaled more linearly with alignment length for introns ($r = 0.992$, $P <$
315 0.00001) than for UCEs ($r = 0.666$, $P < 0.0001$) or CNEEs ($r = 0.228$, $P < 0.00001$; Fig. 2d-f).
316 Although the alignment and variability statistics for UCEs changed significantly when analyzed
317 using the MAFFT-only pipeline we used for CNEEs, the magnitude of the differences were
318 small and trends among markers did not change (Supplementary File S3). Similarly, when we
319 re-aligned all three marker types with the SATé-MAFFT used by Jarvis et al. (2014), overall
320 trends and differences between markers were unchanged (Supplementary File S4).

321 *GC-content and substitution dynamics of noncoding markers:* CNEEs exhibited systematically
322 lower GC-contents than did introns or UCEs (Fig. 4a and b). There was a correlation between
323 the GC-contents of different noncoding markers across species, presumably indicating a genome-
324 wide effect on base composition that influences all three marker types (Supplementary File S5).
325 A notable outlier in GC-content across all three marker types is the Downy Woodpecker
326 (*Picoides pubescens*), with average GC contents of 37.52% (CNEEs), 42.44% (introns) and
327 40.48% (UCEs), values that deviate from the grand mean for each marker type often 10 times
328 more than for other species (Fig. 4a and b). High variance in GC-content can complicate
329 phylogenomic analyses, since most phylogenetic models assume that all species in the analysis
330 share a similar equilibrium base composition (Lockhart et al. 1994; Foster and Hickey 1999;
331 Mooers and Holmes 2000). We found that the variance in GC-content among species was lowest
332 for CNEE markers (average variance = 0.82), and higher for intron and UCE markers (average
333 variance = 5.76 and 3.91, respectively; Fig. 4c). These substitution dynamics held in non-
334 overlapping sets of CNEEs and UCEs (Supplementary File S6).

Using jModelTest, we evaluated the substitution dynamics and optimal substitution model for each alignment. On average, CNEEs exhibited higher transition/transversion rate ratios (average 2.44) than did introns (1.90) or UCEs (1.79; Supplementary Fig. S1d). CNEEs also exhibited intermediate estimates of the gamma shape parameter (average 1.46) compared to introns (7.81) or UCEs (0.92; Supplementary Fig. S1). Overall, although all three markers displayed a similar range of nucleotide substitution models, the most complex models (GTR+ G+ I and GTR+ G) were least prevalent as the best-fitting model for CNEEs (7.2 and 29.2% of loci, respectively) than for introns (13. 7 and 73.2%) or UCEs (74.7 and 23.6%; Supplementary File S7). CNEEs displayed significantly higher consistency indices (mean = 0.92 for full and non-overlapping set) than UCEs (mean = 0.82; $p < 0.00001$) or introns (mean = 0.82; $p < 0.00001$); Supplementary Fig. S1, Supplementary File S6).

Phylogenomic signal and consistency of noncoding sequences

As expected from the rank order of variability of each of the three marker types, gene trees made from CNEE alignments exhibited the lowest average bootstrap support, with introns and UCEs having progressively higher support (Supplementary Fig. S1c). However, the estimates of overall phylogenetic relationships and clade support as judged by species tree analyses were generally concordant among marker types and with previous analyses using larger data sets (Jarvis et al. 2014). All markers recovered ratite paraphyly, with the emu clustering with the two tinamous to the exclusion of the ostrich at 100% bootstrap support (Fig. 5a-c). In all three trees, the Neognathae are monophyletic and the three taxa representing Galloanseriformes (Chicken, Turkey and Peking Duck) were monophyletic at 100%, appearing as expected as sister to all the remaning taxa (Neoaves). All branches in the MP-EST species trees in this study achieved ≥ 95% for all marker types, except for two branches in the total

CNEE tree, two branches in the total intron tree, and one branch in the total UCE tree. The branches in question invariably involved relationships among the outgroups to passerine birds and falcons, a clade termed Australaves (Jarvis et al. 2014; Prum et al. 2015, Fig. 5d-f). Whereas the total CNEE tree suggests that the Bald Eagle is closer to this clade than the Downy Woodpecker (albeit with only 72% and 56% bootstrap support, respectively, for these two branchings), both the total intron and UCE trees support the reverse branching order, with first Downy Woodpecker (at 87% and 70% bootstrap support for introns and UCEs, respectively), then Bald Eagle (with 100% support in both cases) forming successive sister groups to the Australaves. Depending on how one likes to draw bootstrap support cutoffs in phylogenomics analyses, there is no case among the total marker trees of strongly supported conflict in overall species tree estimates among the three marker types for any cutoff greater than 87%. This trend largely held for phylogenetic analysis of the non-overlapping subsets of CNEEs and UCEs (Supplementary Fig. S2): support values increase for CNEEs (72% to 89% for eagle+falcon/passerines and 56% to 85% for woodpecker+other 'land birds'), and decrease for UCEs (70% to 62% for woodpecker+falcon/passerines). When we confine phylogenetic analysis to the 1000 loci with the highest variability or most highly supported gene trees, the results are largely similar (Supplementary Fig. S3).

The relationships obtained for the three marker types are also similar to most of the analyses produced by the Avian Phylogenomics Project (Jarvis et al. 2014, Fig. 5d-f) and a recent, more taxon-rich tree for birds produced with 259 loci, most of which were derived from coding sequences (Prum et al. 2015). A source of disagreement for the taxa that we have sampled involved the sister group to Australaves (Prum et al. 2015). Although both Prum et al. (2015) and Jarvis et al (2014) generally produced trees placing Woodpeckers closer to

Australaves than eagles, neither paper produced this result unambiguously; whereas Jarvis et al. (2014) achieved 100% support for a sister clade to Australaves that included both woodpeckers and eagles in their total evidence concatenation tree (TENT) tree using ExaML, other analyses from Jarvis et al. (2014), as well as the results of Prum et al., (2015) , placed woodpeckers as sister to Australaves, with eagles falling outside this clade, albeit with highly varying levels of support. The relationships among waterbirds (penguin, loon, flamingo and grebe), although consistent across analyses and markers in this study, constitute another region of disagreement with studies employing more taxa. Whereas this study and Prum et al. (2015) suggest monophyly of the four water birds sampled here (Aequorlitornithes), many of the Jarvis analyses, including their TENT analysis, suggested paraphyly of this clade. Because our taxon sampling is so low we naturally defer to these larger studies for the provisionally 'correct' results for these clades, although we note that these larger studies were not able to robustly resolve all relationships, including the two clades discussed here.

We conducted phylogenomic subsampling to study the accumulation of signal as the number of loci increases for two expected clades that ultimately achieve high certainty for all data sets as well as for the two uncertain clades described above. The two high-confidence relationships we examined were the paraphyly of ratites and the sister group to passerines (i.e. falcons; Fig. 6a and b). We found that all three marker types established high confidence in the paraphyly of ratites by 200 genes, with introns accumulating signal somewhat faster than CNEEs and UCEs (Fig. 6a). By contrast, the falcon+passerine clade achieved consistent 100% support at 1000 loci for introns and UCEs, whereas CNEEs did not achieve an average of 100% support for the number of loci analyzed here, peaking at 98% support at 3500 loci and 99% with the full data set (Fig. 6b). For the monophyly of the waterbird clade (Fig. 6c), we found that the

accumulation of signal was more rapid for CNEEs and UCEs, and less rapid for introns. Introns achieved an average bootstrap support of only ~70% for subsamples of 3500 loci (only 61% for the full data set), whereas average support of similarly sized subsamples of CNEEs and UCEs approached 100% (98 and 99%, respectively, for the full data set). For this clade, no marker type exhibited monotonically increasing average support with larger subsamples of loci, although the lack of monotonic increase was much more pronounced for introns than for the other two markers (Fig. 6c). The subsampling results for the sister to Australaves are more interesting, in so far as they begin to suggest genuine conflicts between the marker types. Whereas both introns and UCEs accumulate stronger signal favoring a woodpecker+Australaves clade (87 and 70%, respectively; Fig. 5b and c; Fig. 6d), the CNEEs instead accumulate stronger signal favoring an eagle+Australaves clade, approaching 72% (Fig. 5a, 6d). Whereas CNEEs exhibit a threshold of sorts for the accumulation of signal for the waterbird clade, increasing in average support and number of replicates achieving > 70% support at 500 loci (in part an artifact of the particular intervals chosen for subsampling), introns suggest a threshold at 1500 loci for the woodpecker/Australaves clade (Supplementary Fig. S4).

DISCUSSION

In this study we explored the evolution of CNEEs, a class of noncoding marker that has not received attention in terms of its utility for phylogenomics, and compared them to the performance of two other classes of noncoding markers, introns and UCEs. Overall, the full data set of CNEEs performed well compared to introns and UCEs, with a similar number (1-2) of branches in our 17-taxon tree achieving less than 95% support. The utility of CNEEs for phylogenomic analysis will depend somewhat on the values held by different researchers. If a researcher values high support for branches achieved quickly as numbers of loci are increased, at

427 the expense of more uncertain and gappy alignments with missing species, then introns and
428 UCEs clearly outperform CNEEs. However, if a researcher favors higher certainty and quality
429 of alignment, and a better fit of alignments to the equilibrium assumptions of most phylogenetic
430 models of nucleotide substitution, then CNEEs may offer advantages. The major advantages of
431 CNEEs are the ease of obtaining large numbers of high-quality alignments without missing
432 species, their low homoplasy and their low variance in GC across species. Despite their low
433 variability, and the correspondingly weak support in gene trees, CNEEs produced a species tree
434 that rivaled those produced by a similar number of intron and UCE alignments (Fig. 6). Indeed,
435 given that the CNEE alignments were the shortest of the three marker types, one could argue for
436 the overall efficiency of CNEEs in terms of phylogenetic resolution per base pair sequenced as
437 compared to the other two markers.

438 A critical factor is the number of markers of reasonable length available to
439 phylogeneticists, and the ease of producing fully-populated alignments, since these factors could
440 place a limit on phylogenomic resolution of a particular marker type. Introns are numerous in
441 vertebrate genomes, on the order of several times the number of genes, which usually number
442 about 15,000-20,000. However, orthologous introns often vary substantially in length among
443 taxa (Vinogradov 2002; Waltari and Edwards 2002; Pozzoli et al. 2007; Zhang and Edwards
444 2012); due to their high variability and length differences, gaps will be frequent, with many
445 alignments > 400 bp having large numbers of unfilled (missing) bases. Conserved elements are
446 very numerous in vertebrate genomes, with as many as 3.6 million elements detected in
447 mammals, over 80% of which are noncoding (Lindblad-Toh et al. 2011). However, the average
448 length of these elements is often < 50 bp. Faircloth et al. (2012) were able to assemble 5599
449 unique UCEs, which need to achieve a certain minimum length of the core region to be

450 detectable by hybrid capture methods. Bejerano et al. (2004) found only 481 fully conserved
451 UCEs longer than 200 bp in vertebrate genomes, and the total number of UCEs > 100 bp in
452 vertebrates is estimated to be ~14,000 (Stephen et al. 2008). Today, markers designated as
453 "UCEs" often contain loci that are not strictly UCEs, but rather CNEEs, whose core is often
454 more variable than the original definition of UCEs (Bejerano et al. 2004). In this respect, the
455 number of "UCE" loci has increased in recent studies and can overlap even more with loci
456 designated here as CNEE.

457 It was straightforward to compile a data set of several thousand CNEE markers > 400 bp
458 which contained all species, and most of which contained < 2 alignment indels. By contrast,
459 because of their intrinsic variability or reliance on variable flanking regions, both introns and
460 UCEs had between 20-40% missing data (gaps) per species per alignment, a consequence of
461 their high indel rate, and it was challenging to find intron alignments that contained all 17
462 species in our study. These trends were evident despite the fact that all three marker types were
463 harvested from whole genomes, as opposed to being generated using molecular methods such as
464 hybrid capture. The reasons for the lower incidence of fully-populated alignments for introns
465 does not seem to lie in the lower coverage of some of the genomes used since the CNEEs were
466 harvested from the same source data. Rather, it seems to lie in the greater challenges of
467 detecting introns via blast, or by challenges with genome annotations, or the great length of
468 many introns, which undercuts search algorithms.

469 The three marker types exhibited differing patterns of nucleotide substitution, which
470 could influence their phylogenetic performance, and which appear to be driven in part by overall
471 levels of variation. For example CNEEs had the lowest level of among-lineage variation in GC
472 content, a trait that conforms well with the equilibrium assumptions of most models of

473 nucleotide substitution. To our knowledge we are the first to report the anomalous GC content
474 of the Downy Woodpeker genome. It is likely that the higher fraction of transposable elements
475 in this genome (~22%) compared to other birds investigated thus far, as reported by (Zhang et al.
476 2014), is linked to the outlier status of the Downy Woodpecker in terms of GC content, although
477 we have not verified the prediction that TEs in this genome are higher in GC than other genomic
478 regions. As expected due to their inclusion of the slowly evolving core as well as more rapidly
479 evolving flanking sequences, UCEs exhibited high levels of among-site rate variation (low α)
480 compared to introns and CNEEs. Although not necessarily detrimental to phylogenetic analysis,
481 it is widely acknowledged that high levels of among site rate variation are more difficult to
482 model than low levels (Vogler et al. 2005; Marshall et al. 2006; Holland et al. 2013). On the
483 other hand, CNEEs exhibited the highest transition/transversion ratio among the markers;
484 although high ts/tv ratios, like among-site rate variation, often lead to homoplasy, the higher
485 consistency index among CNEEs appears here to be driven more by their low substitution rate
486 than ts/tv ratio. CNEEs were markedly more AT-rich than the other two classes of markers,
487 which, as a group, tend to be more AT-rich than coding regions . Although AT- versus GC- rich
488 markers do not present any obvious advantages, Romiguier et al. (2013) recently suggested that,
489 in mammals, GC-rich markers result in higher gene tree heterogeneity than AT-rich markers,
490 possibly due to biased gene conversion, making phylogenetic analysis more challenging.

491 *Information Content of CNEEs for Phylogenetic Analysis*

492 Overall, we found that CNEEs delivered an estimate of phylogenetic relationships that
493 was as strong as that for UCEs and introns. For some expected phylogenetic results, such as the
494 paraphyly of the ratites, the approach to phylogentic "certainty" (100% bootstrap support) was as
495 fast as that for the other two markers. However, for other questions that appear to be gaining

consistent support among phylogenomic data sets, such as the falconid sister group of the passerine birds, the approach to phylogenomic resolution was markedly slower than for UCEs or introns. And yet for other clades, such as the sister relationship between penguin/loon and flamingo/grebe, it was introns that failed to achieve high resolution compared to CNEEs and UCEs. Finally, CNEEs suggested a different sister group to falconids and passerines, namely eagles, at fairly high (~80%) and increasing support as more loci were accumulated, as compared to introns and UCEs, which favored woodpeckers as the sister group, again with high support. This result was the only case of moderately strong conflict among markers in our data set, and in our view, either result is plausible, given that this node was not resolved with certainty among larger data sets (Jarvis et al. 2014). For example, the fact that among the taxa we studied the woodpecker is a base compositional outlier more strongly for introns and UCEs than for CNEEs could be driving this difference in result. We were able to achieve high and consistent confidence for nearly all branches in our analysis without binning (Mirarab et al. 2014), suggesting that large numbers of loci, rather than concatenation of loci, remains a plausible way forward for phylogenomics (Liu and Edwards 2015). Although our results point to possible differences in performance and useful trends among these markers, because we only sampled 16 ingroup bird species as exemplars, the generality of these trends requires further investigation.

Finally, we do not expect CNEEs to provide resolution at low taxonomic levels or in phylogeography, where UCEs appear useful (Lemmon and Lemmon 2012; McCormack et al. 2013; Smith et al. 2014; Hamilton et al. 2016; Manthey et al. 2016), in part due to their low variability but also because, like UCEs, they are likely to be under strong background selection (Katzman et al. 2007).

In summary, CNEEs appear to be a promising tool for phylogenomic research. Their low variability compared to introns and UCEs is offset by the larger numbers of moderately long and high quality alignments that can be gathered from whole-genome data sets. In the future, as whole genomes become more readily available, phylogenomic data sets will increasingly be generated via statistical tools or extraction of large sets of alignments from aligned or unaligned genomes (Costa et al. 2016), rather than directly by wet lab bench work. Until that time arrives, wet-lab approaches to gathering loci, such as hybrid capture, will continue to be used. In either scenario, CNEEs should fare well, because they are readily identified by statistical means from whole genomes, and yet they would also be amenable to hybrid capture approaches. We expect that mixtures of noncoding phylogenomic markers, including CNEEs, will be helpful in understanding the dynamics of currently popular markers such as UCEs and introns and will contribute to resolving the Tree of Life.

SUPPLEMENTARY INFORMATION

Supplementary File S1. Sources of noncoding markers used in this paper, including 14 exemplar species from the Avian Phylogenomics Project.

Supplementary File S2. Alignment summary measures for all loci.

Supplementary File S3. Comparison of UCE alignment and variability metrics for SATé-MAFFT alignments with emu and Chilean tinamou profile aligned, and for de novo alignment of all sequences with MAFFT.

539 **Supplementary File S4.** Comparison of alignment and variability metrics for de novo
540 alignment with MAFFT (CNEEs) and SATé-MAFFT alignment with emu and Chilean tinamou
541 profile aligned (introns and UCEs) as used throughout the manuscript, versus alignment
542 following the pipeline described in Jarvis et al. (2015).

543 **Supplementary File S5.** Pairwise correlations in per-taxon GC content between marker types.

544 **Supplementary File S6.** Substitution dynamics of non-overlapping sets of CNEEs and UCEs.

545 **Supplementary File S7.** Summary of jModelTest output for all loci.

546 **Supplementary Fig. S1**. Distribution of variable and parsimony informative sites, gene tree
547 resolution, and substitution dynamics across marker sets.

548 **Supplementary Fig. S2.** Phylogenetic analysis of non-overlapping sets of CNEEs (n = 2318
549 loci), introns (n = 3685 loci) and UCEs (n = 2232 loci).

550 **Supplementary Fig. S3.** Phylogenetic analysis of the 1000 most strongly supported gene trees
551 for each marker (top row) and the 1000 most variable markers (bottom row).

552 **Supplementary Fig. S4.** Threshold analysis for subsampling across markers.

553

554

555

556

557

FUNDING

This research was supported by NSF grant DEB 1355343 (EAR 1355292) to SVE and Julia Clarke.

ACKNOWLEDGEMENTS

We thank Brant Faircloth for help in assembling the UCE data set, and Brant Faircloth and Craig Lowe for helpful discussion and comments on the manuscript. This research was generously supported by the Harvard University FAS Research Computing and the Odyssey Cluster.

565 LITERATURE CITED

566 Baker, A. J., Haddrath O., Mcpherson J. D., and Cloutier A. 2014. Genomic Support for a Moa-
567 Tinamou Clade and Adaptive Morphological Convergence in Flightless Ratites.
568 Molecular Biology and Evolution 31:1686-1696.

569 Bejerano, G., Pheasant M., Makunin I., Stephen S., Kent W. J., Mattick J. S., and Haussler D.
570 2004. Ultraconserved elements in the human genome. Science 304:1321-1325.

571 Bi, K., Vanderpool D., Singhal S., Linderoth T., Moritz C., and Good J. M. 2012.
572 Transcriptome-based exon capture enables highly cost-effective comparative genomic
573 data collection at moderate evolutionary scales. BMC Genomics 13:403.

574 Blanchette, M., Kent W. J., Riemer C., Elnitski L., Smit A. F. A., Roskin K. M., Baertsch R.,
575 Rosenbloom K., Clawson H., Green E. D., Haussler D., and Miller W. 2004. Aligning
576 multiple genomic sequences with the threaded blockset aligner. Genome Research
577 14:708-715.

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

580 Burbrink, F. T., and Pyron R. A. 2011. The Impact of Gene-Tree/Species-Tree Discordance on
581 Diversification-Rate Estimation. Evolution 65:1851-1861.

582 Capella-Gutiérrez, S., Silla-Martínez J. M., and Gabaldón T. 2009. trimAl: a tool for automated
583 alignment trimming in large-scale phylogenetic analyses. Bioinformatics (Oxford,
584 England) 25:1972-1973.

585 Chen, M. Y., Liang D., and Zhang P. 2015. Selecting Question-Specific Genes to Reduce
586 Incongruence in Phylogenomics: A Case Study of Jawed Vertebrate Backbone Phylogeny.
587 Systematic Biology 64:1104-1120.

588 Chojnowski, J. L., Kimball R. T., and Braun E. L. 2008. Introns outperform exons in analyses of
589 basal avian phylogeny using clathrin heavy chain genes. Gene 410:89-96.
590 Costa, I. R., Prosdocimi F., and Jennings W. B. 2016. In silico phylogenomics using complete
591 genomes: a case study on the evolution of hominoids. Genome Res 26:1257-1267.
592 Cummins, C. A., and Mcinerney J. O. 2011. A Method for Inferring the Rate of Evolution of
593 Homologous Characters that Can Potentially Improve Phylogenetic Inference, Resolve
594 Deep Divergence and Correct Systematic Biases. Systematic Biology 60:833-844.
595 Darriba, D., Taboada G. L., Doallo R., and Posada D. 2012. jModelTest 2: more models, new
596 heuristics and parallel computing. Nature Methods 9:772-772.
597 Delsuc, F., Brinkmann H., and Philippe H. 2005. Phylogenomics and the reconstruction of the
598 tree of life. Nature Reviews Genetics 6:361-375.
599 Edwards, S. V. 2016. Phylogenomic subsampling: a brief review. Zoologica Scripta 45:63-74.
600 Edwards, S. V., Potter S., Schmitt C. J., Bragg J. G., and Moritz C. 2016. Reticulation,
601 divergence, and the phylogeography–phylogenetics continuum. Proceedings of the
602 National Academy of Sciences 113:8025-8032.
603 Eisen, J. A. 1998. Phylogenomics: Improving functional predictions for uncharacterized genes
604 by evolutionary analysis. Genome Research 8:163-167.
605 Faircloth, B. C., Mccormack J. E., Crawford N. G., Harvey M. G., Brumfield R. T., and Glenn T.
606 C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple
607 evolutionary timescales Systematic Biology 61:717-726.
608 Foster, P. G., and Hickey D. A. 1999. Compositional bias may affect both DNA-based and
609 protein-based phylogenetic reconstructions. Journal of Molecular Evolution 48:284-290.

610 Green, R. E., Braun E. L., Armstrong J., Earl D., Nguyen N., Hickey G., Vandewege M. W., St
611 John J. A., Capella-Gutierrez S., Castoe T. A., Kern C., Fujita M. K., Opazo J. C., Jurka
612 J., Kojima K. K., Caballero J., Hubley R. M., Smit A. F., Platt R. N., Lavoie C. A.,
613 Ramakodi M. P., Finger J. W., Suh A., Isberg S. R., Miles L., Chong A. Y., Jaratlerdsiri
614 W., Gongora J., Moran C., Iriarte A., Mccormack J., Burgess S. C., Edwards S. V., Lyons
615 E., Williams C., Breen M., Howard J. T., Gresham C. R., Peterson D. G., Schmitz J.,
616 Pollock D. D., Haussler D., Triplett E. W., Zhang G., Irie N., Jarvis E. D., Brochu C. A.,
617 Schmidt C. J., Mccarthy F. M., Faircloth B. C., Hoffmann F. G., Glenn T. C., Gabaldoon
618 T., Paten B., and Ray D. A. 2014. Three crocodilian genomes reveal ancestral patterns of
619 evolution among archosaurs. Science 346:1335-+.
620 Groussin, M., Hobbs J. K., Szöllősi G. J., Gribaldo S., Arcus V. L., and Gouy M. 2015. Toward
621 More Accurate Ancestral Protein Genotype–Phenotype Reconstructions with the Use of
622 Species Tree-Aware Gene Trees. Molecular Biology and Evolution 32:13-22.
623 Hackett, S. J., Kimball R. T., Reddy S., Bowie R. C. K., Braun E. L., Braun M. J., Chojnowski J.
624 L., Cox W. A., Han K. L., Harshman J., Huddleston C. J., Marks B. D., Miglia K. J.,
625 Moore W. S., Sheldon F. H., Steadman D. W., Witt C. C., and Yuri T. 2008. A
626 phylogenomic study of birds reveals their evolutionary history. Science 320:1763-1768.
627 Hamilton, C. A., Lemmon A. R., Lemmon E. M., and Bond J. E. 2016. Expanding anchored
628 hybrid enrichment to resolve both deep and shallow relationships within the spider tree of
629 life. BMC Evol Biol 16:212.
630 Harshman, J., Braun E. L., Braun M. J., Huddleston C. J., Bowie R. C., Chojnowski J. L.,
631 Hackett S. J., Han K. L., Kimball R. T., Marks B. D., Miglia K. J., Moore W. S., Reddy
632 S., Sheldon F. H., Steadman D. W., Steppan S. J., Witt C. C., and Yuri T. 2008.

633 Phylogenomic evidence for multiple losses of flight in ratite birds. Proc Natl Acad Sci U
634 S A 105:13462-13467.
635 Holland, B. R., Jarvis P. D., and Sumner J. G. 2013. Low-Parameter Phylogenetic Inference
636 Under the General Markov Model. Systematic Biology 62:78-92.
637 Jarvis, E. D., Mirarab S., Aberer A. J., Li B., Houde P., Li C., Ho S. Y. W., Faircloth B. C.,
638 Nabholz B., Howard J. T., Suh A., Weber C. C., Da Fonseca R. R., Li J. W., Zhang F., Li
639 H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M. S., Zavidovych V.,
640 Subramanian S., Gabaldon T., Capella-Gutierrez S., Huerta-Cepas J., Rekepalli B.,
641 Munch K., Schierup M., Lindow B., Warren W. C., Ray D., Green R. E., Bruford M. W.,
642 Zhan X. J., Dixon A., Li S. B., Li N., Huang Y. H., Derryberry E. P., Bertelsen M. F.,
643 Sheldon F. H., Brumfield R. T., Mello C. V., Lovell P. V., Wirthlin M., Schneider M. P.
644 C., Prosdocimi F., Samaniego J. A., Velazquez A. M. V., Alfaro-Nunez A., Campos P. F.,
645 Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D. M.,
646 Zhou Q., Perelman P., Driskell A. C., Shapiro B., Xiong Z. J., Zeng Y. L., Liu S. P., Li Z.
647 Y., Liu B. H., Wu K., Xiao J., Yinqi X., Zheng Q. M., Zhang Y., Yang H. M., Wang J.,
648 Smeds L., Rheindt F. E., Braun M., Fjeldsa J., Orlando L., Barker F. K., Jonsson K. A.,
649 Johnson W., Koepfli K. P., O'brien S., Haussler D., Ryder O. A., Rahbek C., Willerslev
650 E., Graves G. R., Glenn T. C., Mccormack J., Burt D., Ellegren H., Alstrom P., Edwards
651 S. V., Stamatakis A., Mindell D. P., Cracraft J., Braun E. L., Warnow T., Jun W., Gilbert
652 M. T. P., and Zhang G. J. 2014. Whole-genome analyses resolve early branches in the
653 tree of life of modern birds. Science 346:1320-1331.
654 Ji, Z., Song R., Regev A., and Struhl K. 2015. Many lncRNAs, 5'UTRs, and pseudogenes are
655 translated and some are likely to express functional proteins. Elife 4:e08890.

656 Katoh, K., and Standley D. M. 2013. MAFFT multiple sequence alignment software version 7:
657 improvements in performance and usability. Mol Biol Evol 30:772-780.
658 Katzman, S., Kern A. D., Bejerano G., Fewell G., Fulton L., Wilson R. K., Salama S. R., and
659 Haussler D. 2007. Human genome ultraconserved elements are ultraselected. Science
660 317:915.
661 Kent, W. J., Baertsch R., Hinrichs A., Miller W., and Haussler D. 2003. Evolution's cauldron:
662 duplication, deletion, and rearrangement in the mouse and human genomes. Proceedings
663 of the National Academy of Sciences of the United States of America 100:11484-11489.
664 Kvon, E. Z., Kamneva O. K., Melo S., Dickel D. E., Melo S., Barozzi I., Osterwalder M.,
665 Mannion B. J., Kvon E. Z., Kamneva O. K., Pickle C. S., Plajzer-Frick I., Lee E. A., Kato
666 M., Garvin T. H., Akiyama J. A., Afzal V., Lopez-Rios J., Rubin E. M., Dickel D. E., and
667 Pennacchio L. A. 2016. Progressive Loss of Function in a Limb Enhancer during Snake
668 Evolution Article Progressive Loss of Function in a Limb Enhancer during Snake
669 Evolution. Cell 167:633-642.
670 Leal, F., and Cohn M. J. 2016. Loss and Re-emergence of Legs in Snakes by Modular Evolution
671 of Sonic hedgehog and HOXD Report Loss and Re-emergence of Legs in Snakes by
672 Modular Evolution of Sonic hedgehog and HOXD Enhancers. Current Biology:1-8.
673 Lemmon, A. R., Emme S. A., and Lemmon E. M. 2012. Anchored hybrid enrichment for
674 massively high-throughput phylogenomics. Systematic Biology 61:727-744.
675 Lemmon, A. R., and Lemmon E. M. 2012. High-Throughput Identification of Informative
676 Nuclear Loci for Shallow-Scale Phylogenetics and Phylogeography. Systematic Biology
677 61:745-761.

678 Lemmon, E. M., and Lemmon A. R. 2013. High-throughput genomic data in systematics and
679 phylogenetics. Annual Review of Ecology, Evolution, and Systematics 44:99-121.
680 Lindblad-Toh, K., Garber M., Zuk O., Lin M. F., Parker B. J., Washietl S., Kheradpour P., Ernst
681 J., Jordan G., Mauceli E., Ward L. D., Lowe C. B., Holloway A. K., Clamp M., Gnerre S.,
682 Alfoldi J., Beal K., Chang J., Clawson H., Cuff J., Di Palma F., Fitzgerald S., Flicek P.,
683 Guttman M., Hubisz M. J., Jaffe D. B., Jungreis I., Kent W. J., Kostka D., Lara M.,
684 Martins A. L., Massingham T., Moltke I., Raney B. J., Rasmussen M. D., Robinson J.,
685 Stark A., Vilella A. J., Wen J. Y., Xie X. H., Zody M. C., Worley K. C., Kovar C. L.,
686 Muzny D. M., Gibbs R. A., Warren W. C., Mardis E. R., Weinstock G. M., Wilson R. K.,
687 Birney E., Margulies E. H., Herrero J., Green E. D., Haussler D., Siepel A., Goldman N.,
688 Pollard K. S., Pedersen J. S., Lander E. S., Kellis M., Inst B., Med B. C., and Univ W.
689 2011. A high-resolution map of human evolutionary constraint using 29 mammals.
690 Nature 478:476-482.
691 Liu, L., and Edwards S. V. 2015. Comment on “Statistical binning enables an accurate
692 coalescent-based estimation of the avian tree”. Science 350:171.
693 Liu, L., Yu L., and Edwards S. 2010. A maximum pseudo-likelihood approach for estimating
694 species trees under the coalescent model. BMC Evolutionary Biology 10:302.
695 Lockhart, P. J., Steel M. A., Hendy M. D., and Penny D. 1994. Recovering evolutionary trees
696 under a more realistic model of sequence evolution. Molecular Biology and Evolution
697 11:605-612.
698 Lowe, C. B., Clarke J. A., Baker A. J., Haussler D., and Edwards S. V. 2014. Feather
699 development genes and associated regulatory innovation predate the origin of Dinosauria.
700 Molecular Biology and Evolution 32:23-28.

701 Lowe, C. B., Kellis M., Siepel A., Raney B. J., Clamp, Michele, Salama S. R., Kingsley D. M.,
702 Lindblad-Toh K., and Haussler D. 2011. Three periods of regulatory innovation during
703 vertebrate evolution. Science 333:1019-1024.
704 Manthey, J. D., Campillo L. C., Burns K. J., and Moyle R. G. 2016. Comparison of Target-
705 Capture and Restriction-Site Associated DNA Sequencing for Phylogenomics: A Test in
706 Cardinalid Tanagers (Aves, Genus: Piranga). Systematic biology 65:640-650.
707 Marshall, D. C., Simon C., and Buckley T. R. 2006. Accurate branch length estimation in
708 partitioned Bayesian analyses requires accommodation of among-partition rate variation
709 and attention to branch length priors. Systematic Biology 55:993-1003.
710 Mccormack, J. E., Faircloth B. C., Crawford N. G., Gowaty P. A., Brumfield R. T., and Glenn T.
711 C. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental
712 mammal phylogeny when combined with species tree analysis. Genome Research
713 22:746-754.
714 Mccormack, J. E., Hird S. M., Zellmer A. J., Carstens B. C., and Brumfield R. T. 2013.
715 Applications of next-generation sequencing to phylogeography and phylogenetics.
716 Molecular Phylogenetics and Evolution 66:526-538.
717 Mendes, F. K., and Hahn M. W. 2016. Gene Tree Discordance Causes Apparent Substitution
718 Rate Variation. Systematic Biology 65:711-721.
719 Mirarab, S., Bayzid M. S., Boussau B., and Warnow T. 2014. Statistical binning enables an
720 accurate coalescent-based estimation of the avian tree, Pages 1250463, Science.
721 Mooers, A. O., and Holmes E. C. 2000. The evolution of base composition and phylogenetic
722 inference. Trends in Ecology & Evolution 15:365-369.

723 Novick, P. A., Basta H., Floumanhaft M., Mcclure M. A., and Boissinot S. 2009. The
724 Evolutionary dynamics of autonomous non-LTR retrotransposons in the lizard *Anolis*
725 *carolinensis* shows more similarity to fish than mammals. Molecular Biology and
726 Evolution 26:1811-1822.
727 Pease, J. B., Haak D. C., Hahn M. W., and Moyle L. C. 2016. Phylogenomics Reveals Three
728 Sources of Adaptive Variation during a Rapid Radiation. Plos Biology 14.
729 Philippe, H., Brinkmann H., Lavrov D. V., Littlewood D. T. J., Manuel M., Worheide G., and
730 Baurain D. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are
731 Not Enough. Plos Biology 9.
732 Phillips, M. J., Gibb G. C., Crimp E. A., and Penny D. 2010. Tinamous and moa flock together:
733 mitochondrial genome sequence analysis reveals independent losses of flight among
734 ratites. Syst Biol 59:90-107.
735 Pisani, D., Pett W., Dohrmann M., Feuda R., Rota-Stabelli O., Philippe H., Lartillot N., and
736 Worheide G. 2015. Genomic data do not support comb jellies as the sister group to all
737 other animals. Proceedings of the National Academy of Sciences of the United States of
738 America 112:15402-15407.
739 Posada, D. 2016. Phylogenomics for Systematic Biology. Systematic Biology 65:353-356.
740 Pozzoli, U., Menozzi G., Comi G. P., Cagliani R., Bresolin N., and Sironi M. 2007. Intron size in
741 mammals: complexity comes to terms with economy. Trends in Genetics 23:20-24.
742 Prum, R. O., Berv J. S., Dornburg A., Field D. J., Townsend J. P., Lemmon E. M., and Lemmon
743 A. R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation
744 DNA sequencing. Nature 526:569-U247.

745 Romiguier, J., Ranwez V., Delsuc F., Galtier N., and Douzery E. J. P. 2013. Less is more in
746 mammalian phylogenomics: AT-Rich genes minimize tree conflicts and unravel the root
747 of placental mammals. Molecular Biology and Evolution 30:2134-2144.
748 Schwartz, S., Kent W. J., Smit A., Zhang Z., Baertsch R., Hardison R. C., Haussler D., and
749 Miller W. 2003. Human-mouse alignments with BLASTZ. Genome Research 13:103-107.
750 Siepel, A., Bejerano G., Pedersen J. S., Hinrichs A. S., Hou M., Rosenbloom K., Clawson H.,
751 Spieth J., Hillier L. W., Richards S., Weinstock G. M., Wilson R. K., Gibbs R. A., Kent
752 W. J., Miller W., and Haussler D. 2005a. Evolutionarily conserved elements in vertebrate,
753 insect, worm, and yeast genomes. Genome Research 15:1034-1050.
754 Siepel, A., Bejerano G., Pedersen J. S., Hinrichs A. S., Hou M. M., Rosenbloom K., Clawson H.,
755 Spieth J., Hillier L. W., Richards S., Weinstock G. M., Wilson R. K., Gibbs R. A., Kent
756 W. J., Miller W., and Haussler D. 2005b. Evolutionarily conserved elements in vertebrate,
757 insect, worm, and yeast genomes. Genome Research 15:1034-1050.
758 Siepel, A., and Haussler D. 2004. Phylogenetic estimation of context-dependent substitution
759 rates by maximum likelihood. Mol Biol Evol 21:468-488.
760 Smith, B. T., Harvey M. G., Faircloth B. C., Glenn T. C., and Brumfield R. T. 2014. Target
761 capture and massively parallel sequencing of ultraconserved elements for comparative
762 studies at shallow evolutionary time scales. Syst Biol 63:83-95.
763 Smith, J. V., Braun E. L., and Kimball R. T. 2013. Ratite nonmonophyly: independent evidence
764 from 40 novel Loci. Syst Biol 62:35-49.
765 Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
766 large phylogenies. Bioinformatics 30:1312-1313.

767 Stephen, S., Pheasant M., Makunin I. V., and Mattick J. S. 2008. Large-scale appearance of
768 ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol
769 Biol Evol 25:402-408.
770 Sukumaran, J., and Holder M. 2010. DendroPy: a Python library for phylogenetic computing.
771 Bioinformatics 26:1569-1571.
772 Sukumaran, J., and Holder M. 2015.SumTrees: Phylogenetic Tree Summarization, version 4.0.0.
773 Swofford, D. L. 2002.Phylogenetic analysis using parsimony (* and other methods). Version
774 4.Sinauer Associates, Sunderland, MA.
775 Tan, G., Muffato M., Ledergerber C., Herrero J., Goldman N., Gil M., and Dessimoz C. 2015.
776 Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently
777 Worsen Single-Gene Phylogenetic Inference. Syst Biol 64:778-791.
778 Vinogradov, A. E. 2002. Growth and decline of introns. Trends in Genetics 18:232-236.
779 Vogler, A. P., Cardoso A., and Barraclough T. G. 2005. Exploring rate variation among and
780 within sites in a densely sampled tree: Species level phylogenetics of North American
781 tiger beetles (Genus Cicindela). Systematic Biology 54:4-20.
782 Waltari, E., and Edwards S. V. 2002. Evolutionary dynamics of intron size, genome size, and
783 physiological correlates in archosaurs. American Naturalist 160:539-552.
784 Zhang, G. J., Li C., Li Q. Y., Li B., Larkin D. M., Lee C., Storz J. F., Antunes A., Greenwold M.
785 J., Meredith R. W., Odeen A., Cui J., Zhou Q., Xu L. H., Pan H. L., Wang Z. J., Jin L. J.,
786 Zhang P., Hu H. F., Yang W., Hu J., Xiao J., Yang Z. K., Liu Y., Xie Q. L., Yu H., Lian J.
787 M., Wen P., Zhang F., Li H., Zeng Y. L., Xiong Z. J., Liu S. P., Zhou L., Huang Z. Y.,
788 An N., Wang J., Zheng Q. M., Xiong Y. Q., Wang G. B., Wang B., Wang J. J., Fan Y.,
789 Da Fonseca R. R., Alfaro-Nunez A., Schubert M., Orlando L., Mourier T., Howard J. T.,

790 Ganapathy G., Pfenning A., Whitney O., Rivas M. V., Hara E., Smith J., Farre M.,
791 Narayan J., Slavov G., Romanov M. N., Borges R., Machado J. P., Khan I., Springer M.
792 S., Gatesy J., Hoffmann F. G., Opazo J. C., Hastad O., Sawyer R. H., Kim H., Kim K. W.,
793 Kim H. J., Cho S., Li N., Huang Y. H., Bruford M. W., Zhan X. J., Dixon A., Bertelsen
794 M. F., Derryberry E., Warren W., Wilson R. K., Li S. B., Ray D. A., Green R. E., O'brien
795 S. J., Griffin D., Johnson W. E., Haussler D., Ryder O. A., Willerslev E., Graves G. R.,
796 Alstrom P., Fjeldsa J., Mindell D. P., Edwards S. V., Braun E. L., Rahbek C., Burt D. W.,
797 Houde P., Zhang Y., Yang H. M., Wang J., Jarvis E. D., Gilbert M. T. P., Wang J., and
798 Consortium A. G. 2014. Comparative genomics reveals insights into avian genome
799 evolution and adaptation. Science 346:1311-1320.

800 Zhang, Q., and Edwards S. V. 2012. The evolution of intron size in amniotes: a role for powered
801 flight? Genome Biology and Evolution.

802

803

804 Table 1. Differences between vertebrate CNEEs and UCEs for phylogenetic analysis. This table
805 focuses on the classical definition of UCEs as originally described by Bejerano et al. (2004) and
806 Faircloth et al. (2012).

UCEs	CNEEs
Individual elements found throughout vertebrates, from fish to mammals/birds	CNEEs arise variably in evolution, and are not necessarily found in all vertebrate clades
Information in flanking regions principally used for phylogenetic analysis	Core region used for phylogenetic analysis
Can be coding or noncoding	Only noncoding
Core regions are ultraconserved	Core regions evolve slower than neutral regions, and thus can evolve faster than UCE core regions
Discovered via arbitrary conservation and length thresholds on synteny-aware whole-genome alignments	Discovered via tuned hidden Markov model

807

808

809 **Figure legends**

810

811 **Figure 1.** Hypothetical schematic of comparisons of evolutionary rates of different non-coding
812 markers discussed in this paper. The shading of each bar is meant to indicate the distribution of
813 rates within the range indicated by each bar. Thus, as found in this paper, introns are somewhat
814 more variable than UCEs, and CNEEs are less conserved than classically defined core regions of
815 UCEs but not as variable as introns or the neutral rate.

816

817 **Figure 2:** *Top row*: Distribution of aligned sequence lengths for a) CNEEs (3822 loci), b) introns
818 (3579 loci), and c) UCEs (3679 loci). *Bottom* row: Correlations between alignment length and
819 number of variable sites for d) CNEEs ($r= 0.2277$, $P< 0.00001$), e) introns ($r= 0.9918$, $P< 0.00001$)
820 and f) UCEs ($r= 0.6665$, $P< 0.0001$).

821

822 **Figure 3.** Variation in alignment gappiness among marker types. a) Average percentage of
823 undetermined bases per alignment for each marker type and taxon. Here, undetermined bases
824 indicates both gaps and Ns in each alignment. Alignments that were missing any species were
825 excluded before analysis. b) Same data as in a, expressed as a histogram. c) Distribution of gaps
826 per aligned base pairs, here including only genuine missing sequence. d) Distribution of the
827 percentage of each alignment remaining after trimming with TrimAl. See methods for further
828 discussion.

829

830 **Figure 4.** Patterns of GC content variation among markers and taxa. Alignments in which a
831 taxon is entirely absent are omitted from all calculations. a) Per-taxon values for mean GC
832 content for each marker type. b) Distribution of GC content among species for each marker type.

c) Variance in GC content among taxa for each marker type.

Figure 5. Species tree topologies discussed in this study. *Top row*: MP-EST species trees for a) CNEEs (3822 loci), b) introns (3579 loci), and c) UCEs (3679 loci), with support values < 100% indicated. *Bottom row*: Total evidence nucleotide trees (TENT), each built from 2516 introns, 3769 UCEs, and 8251 protein-coding genes, from Jarvis et al. (2014), pruned to the taxon set used in the current study. d) MP-EST unbinned analysis, e) MP-EST binned analysis and f) concatenated analysis. Support values are omitted from the pruned Jarvis et al. trees. The main tree presented in Prum et al. (2015) is identical to the tree depicted in panel d, assuming that Chilean Tinamou and Emu would fall where found in other studies.

Figure 6. Phylogenomic subsampling, with MP-EST species trees inferred for twelve datasets of increasing numbers of randomly chosen loci, and with ten replicates per dataset. In each row, left panels plot the mean bootstrap support among the ten MP-EST replicates for each marker type and data set size. At right are the two branches, indicated by stars, whose support is investigated by subsampling, with open and solid markers indicating support for one or the other branch. a) Trends in support with increasing numbers of loci for paraphyly of ratites. b) Trends in support for the branch uniting falcons as the sister group to passerine birds. c) Trends in support for competing hypotheses placing either core landbirds (songbirds+falcon+eagle+woodpecker) or (flamingo+grebe) as the sister group to (penguin+loon). d) Trends in support for either bald eagle or downy woodpecker as the sister to (songbirds+falcon).

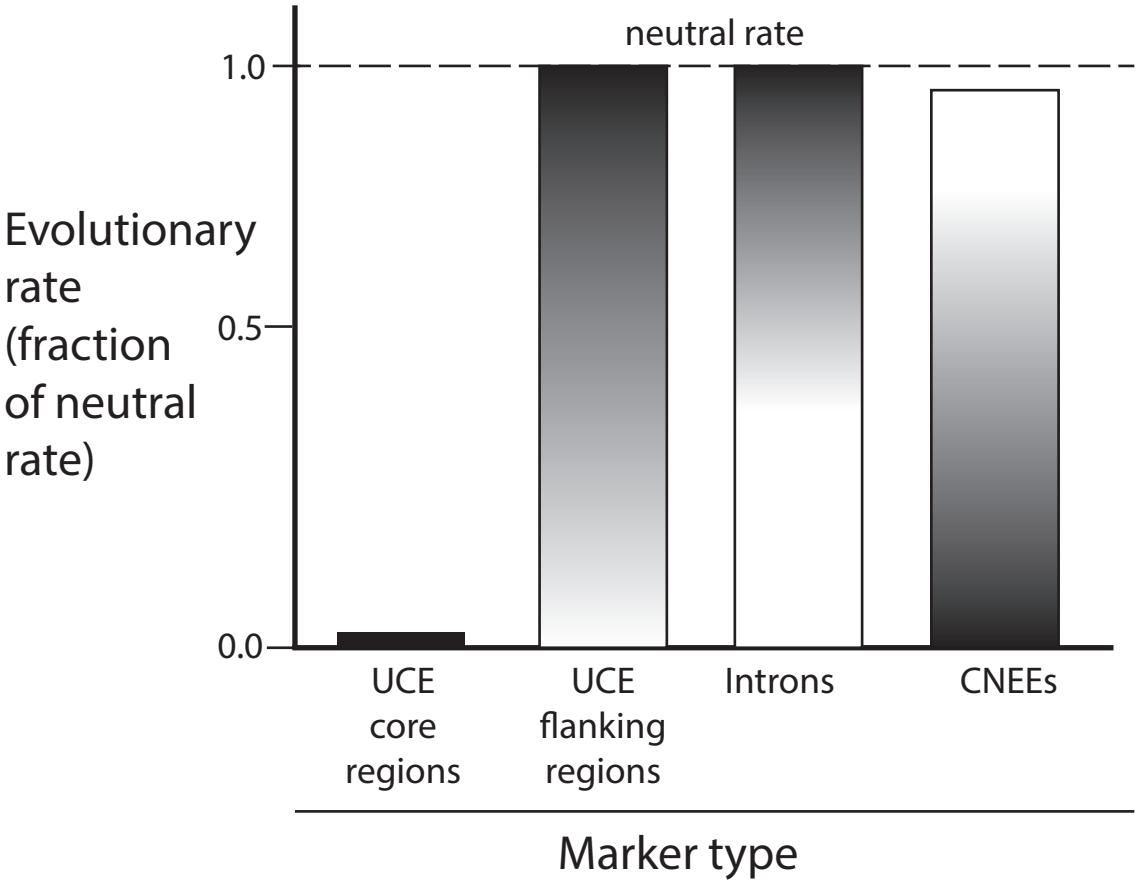

Fig. 1

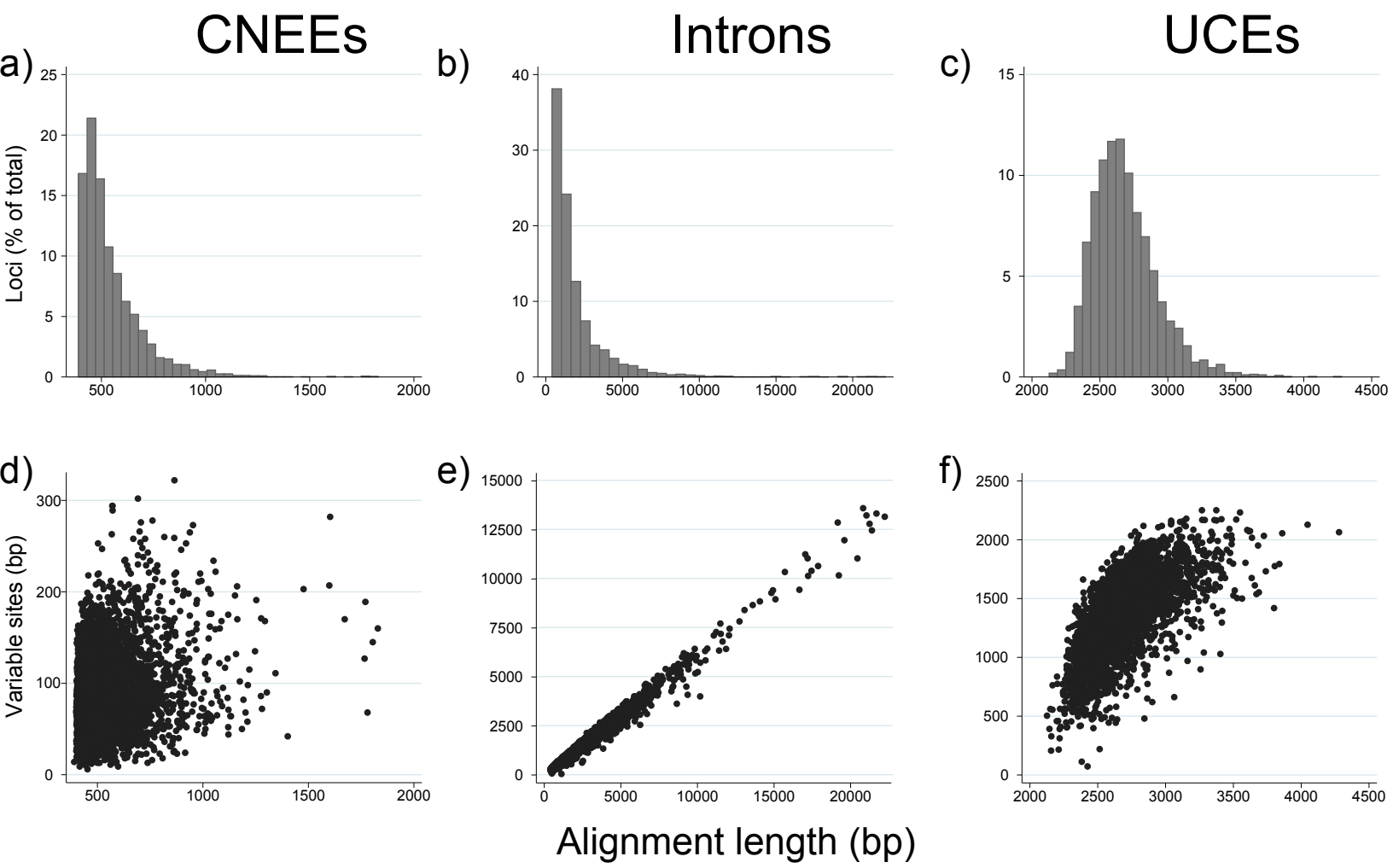

Fig. 2

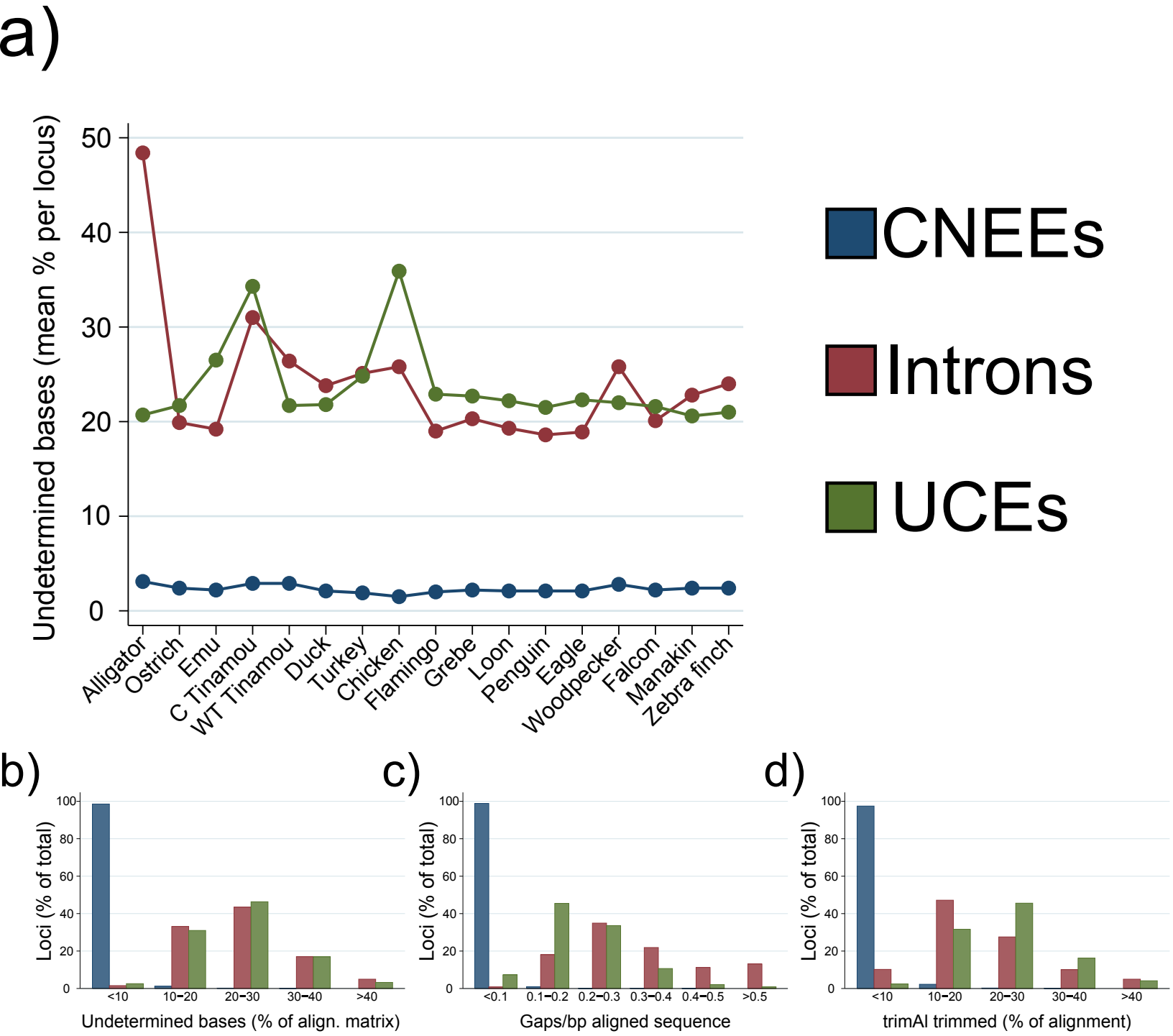

Fig. 3

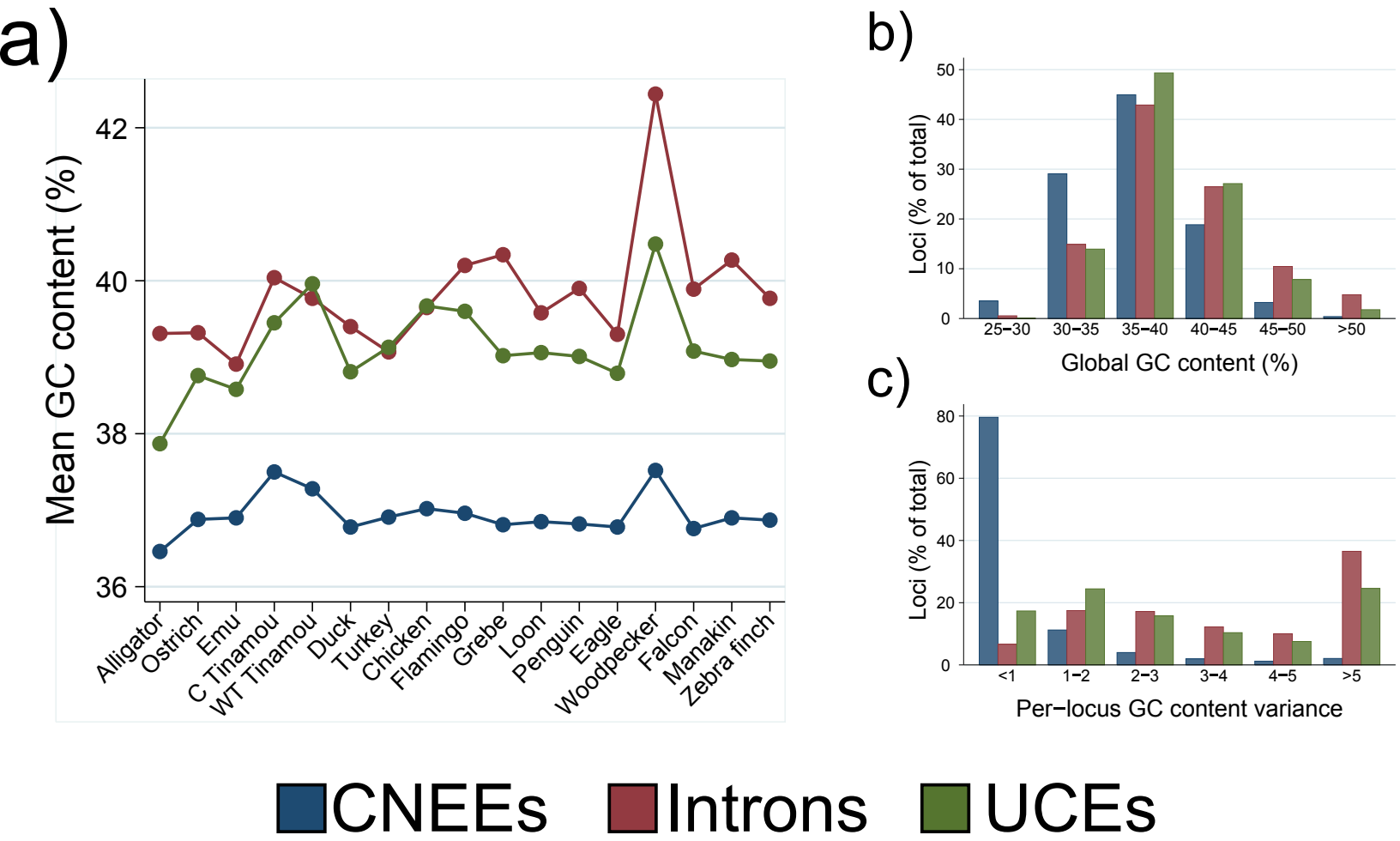

Fig. 4

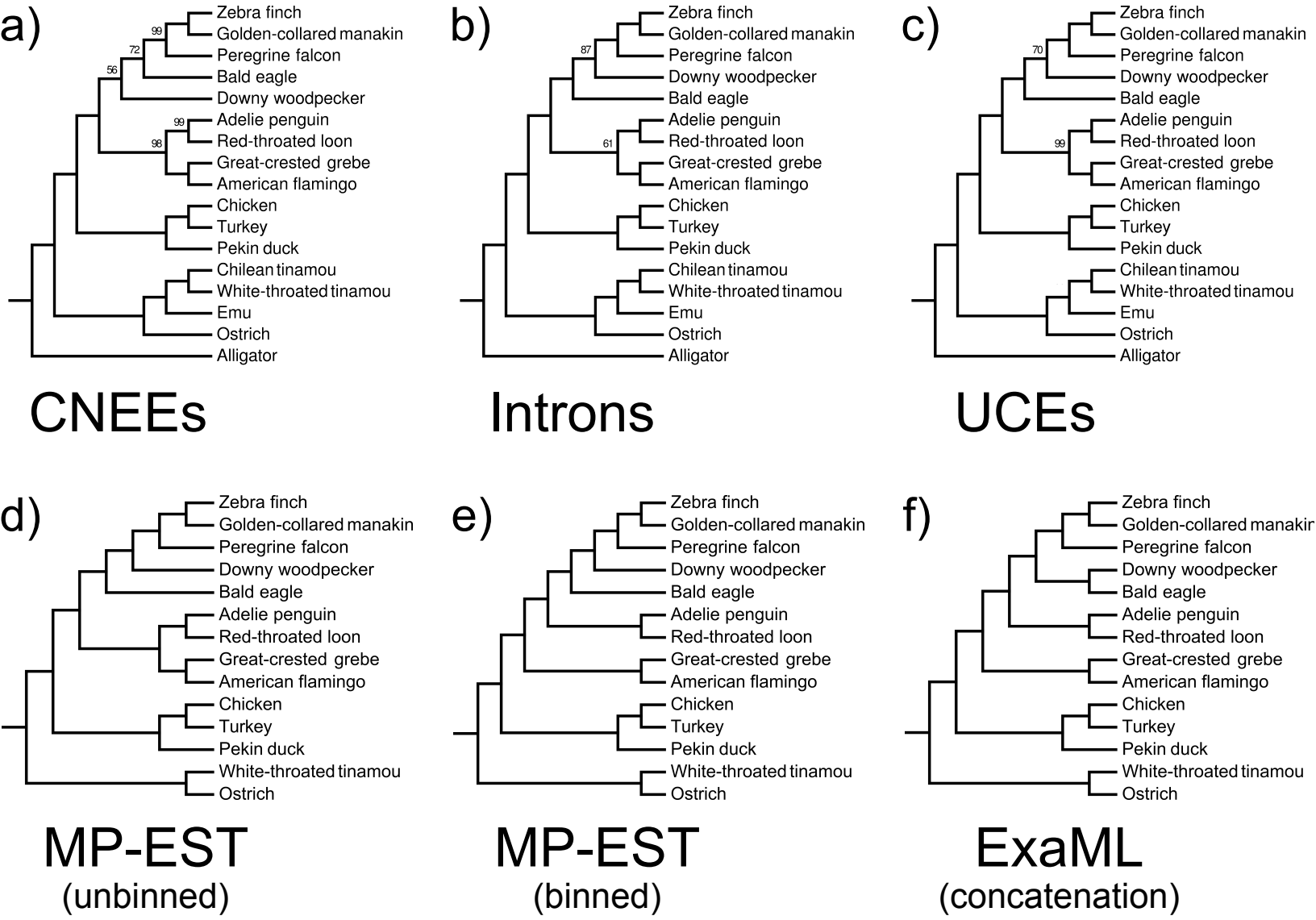

Fig. 5

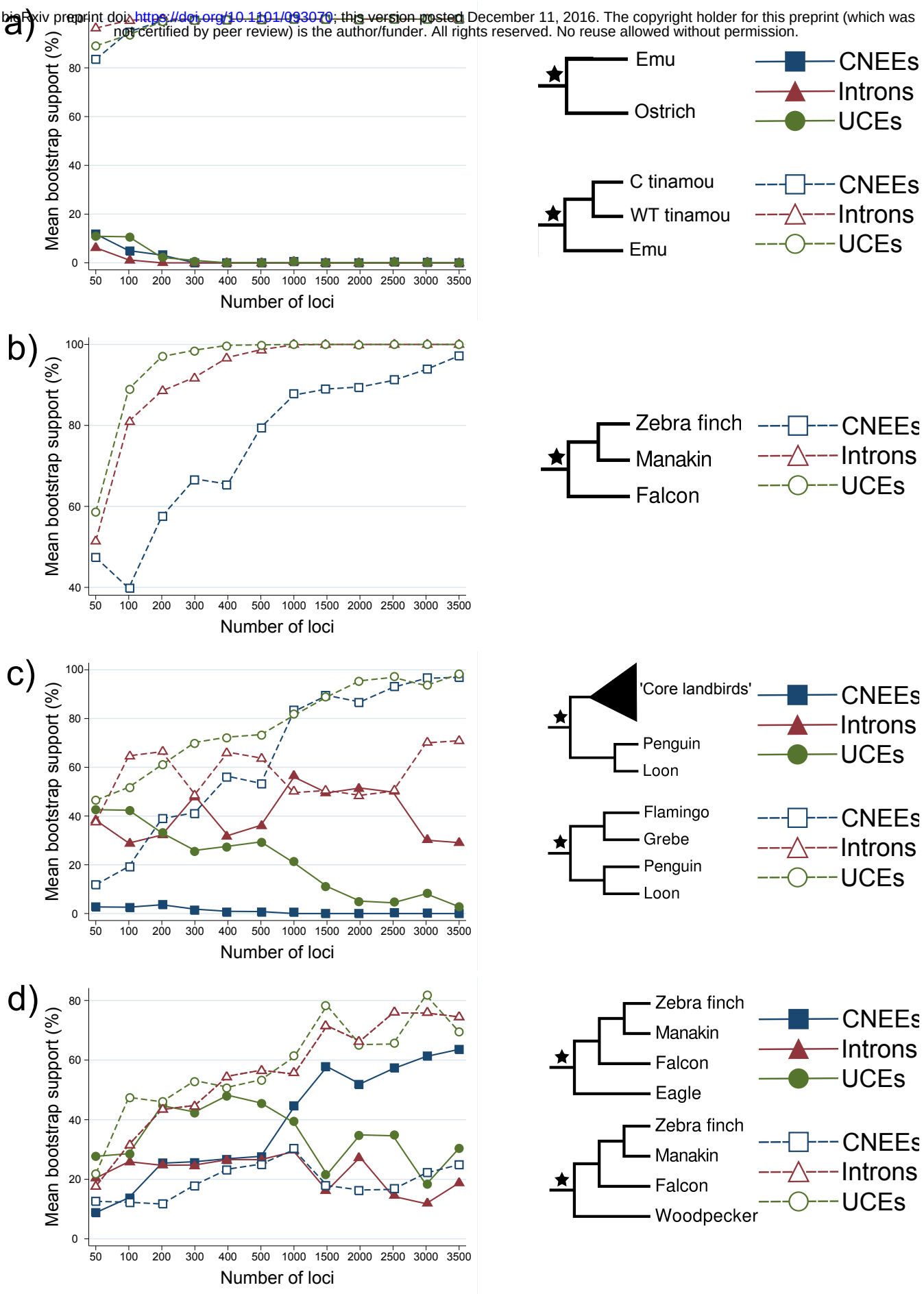

Fig. 6