How Codon Usage Bias Affects Our Ability to Recover the Tree of Life

Justin B. Miller[1], Michael F. Whiting[1,2], John S.K. Kauwe[1], Perry G. Ridge[1,*]

[1]*Department of Biology, Brigham Young University, Provo, UT 84602, USA*

[2]*M.L. Bean Museum, Brigham Young University, Provo, UT 84602, USA*

[*]*Correspondence to be sent to: 4102 LSB, Department of Biology, Brigham Young University, Provo, UT 84602, USA;*

E-mail: perry.ridge@byu.edu

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2019 doi:10.20944/preprints201910.0086.v1

Abstract

Phylogenies depict shared evolutionary patterns and structures on a tree topology, enabling the identification of hierarchical and historical relationships. Recent analyses indicate that phylogenetic signals extend beyond the primary structure of protein or DNA, and various aspects of codon usage biases are phylogenetically conserved. Several functional biases exist within genes, including the number of codons that are used, the position of the codons, and the overall nucleotide composition of the genome. Codon usage biases can significantly affect transcription and translational efficiencies, leading to differential gene expression. Although systematic codon usage biases originate from the overall GC content of a species, ramp sequences, codon aversion, codon pairing, and tRNA competition also significantly affect gene expression and are phylogenetically conserved. We review recent advances in analyzing codon usage biases and their implications in phylogenomics. We first outline common phylogenomic techniques. Next, we identify several codon usage biases and their effects on secondary structure, gene expression, and implications in phylogenetics. Finally, we suggest how codon usage biases can be included in phylogenomics. By incorporating various codon usage biases in common phylogenomic algorithms, we propose that we can significantly improve tree inference. Since codon usage biases have significant biological implications, they should be considered in conjunction with other phylogenetic algorithms.

The Continued Importance of Phylogenetic Systematics

Phylogenetic systematics explores the historical and hierarchical relationships among genes, individuals, populations, and taxa. Phylogenies allow biologists to infer similar characteristics in closely related species and provide an evolutionary framework for analyzing biological patterns (Soltis and Soltis 2003). Furthermore, phylogenies are statements of homology and are used to organize shared structures or patterns between species (Haszprunar 1992). Originally, phylogenies were recovered using only morphological data. However, with the increased availability of molecular data, a combined approach using morphology and genetic markers is typically used in phylogenetic analyses (Bertolani, et al. 2014). Although genetic data provide researchers with access to more species, it typically requires large amounts of data cleaning (e.g., alignment and annotation) before it becomes useful. Some of the greatest difficulties in recovering phylogenetic trees from molecular data (e.g., multiple substitutions at the same position between ancient terminal branches or no substitutions in a gene in short internal tree branches) are explored by Philippe, et al. (2011). These issues have recently become more pertinent as sequencing costs have dropped and genomic data now spans the Tree of Life.

Codon Usage Biases Span the Tree of Life

Codon usage bias is present throughout molecular datasets. There are 61 canonical codons plus three stop codons that form and regulate the creation of 20 amino acids and the stop signal (Crick, et al. 1961). Since there are more codons than amino acids, the term synonymous codon is used to describe how multiple codons encode the same amino acid and were presumably identical in function. However, an unequal distribution of synonymous codons occurs within species, especially within highly expressed genes, suggesting that synonymous codons might play different roles in species fitness (Sharp and Li 1986). Furthermore, an unequal distribution of tRNA anticodons directly coupling codons also varies between species, leading to the wobble hypothesis: tRNA anticodons do not need to latch onto all three

codon nucleotides during translation (Crick 1966). Codon usage is highly associated with the most abundant tRNA present in the cell (Post, et al. 1979) and codon usage patterns affect gene expression (Gutman and Hatfield 1989). Non-random mutations or selection for phenotypic differences caused by differential gene expression could explain some of the phylogenetic differences in synonymous codon usages. Although codon usages directly affect phenotypes, common phylogenomic approaches typically ignore the influence of codon bias in tree inference.

Overview of Common Phylogenomic Techniques

Homologous sequence comparisons are commonly used to identify species relationships. Homologous characters are identified by aligning orthologous genes and detecting character state changes of amino acid residues or nucleotides across a tree topology. This multi-step process is time-consuming and requires orthologous gene annotations. Non-homologous sequence comparisons have also been explored in alignment-free methods and will subsequently be discussed.

1. Ortholog Identification

Orthologs are genes within two or more species that usually share the same function because they are derived from the same ancestral gene in the most recent common ancestor (Koonin 2005). In contrast, paralogs may share the same function, but can arise from gene duplication or horizontal gene transfer. Paralogs may not be under the same evolutionary pressures and should not be compared in a direct positional alignment because these comparisons are a poor indicator of phylogenetic relationships (Koonin 2005). An in-depth evaluation of ortholog identification techniques is presented by Tekaia (2016). Once an ortholog is identified, phylogenetic studies typically require a multiple sequence alignment to align homologous characters. Reviews of some common multiple sequence aligners such as T-coffee (Magis, et al. 2014), MUSCLE (Edgar 2004), Clustal (Sievers and Higgins 2014), Clustal Omega

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2019 doi:10.20944/preprints201910.0086.v1

(Sievers and Higgins 2018), and MAFFT (Katoh and Standley 2014) can be examined elsewhere (Daugelaite, et al. 2013; Pais, et al. 2014).

2. Recovering the Phylogenetic Tree

i. Maximum Parsimony

Maximum parsimony assumes that each character is equally important and minimizes the number of character state changes to recover the relatedness of species. Proponents of parsimony point to its explanatory power and ability to minimize *ad hoc* hypotheses (Farris 2008). However, parsimony can be misleading if unequal evolutionary rates between lineages exist because longer evolutionary branches have a tendency to form monophyletic groups even if the species have different phylogenetic histories (Felsenstein 1978). PAUP (Wilgenbusch and Swofford 2003) and TNT (Goloboff, et al. 2005) are two popular software packages to identify phylogenies based on parsimony.

ii. Maximum Likelihood

Maximum likelihood requires specific models of evolution that show the probability of character state changes and can be used in the likelihood function. Maximum likelihood calculates the probability of obtaining the data given the model and tree topology. One of the main reasons that maximum likelihood estimates have gained traction is the mathematical property of consistency, which states that as more data (phylogenetically informative characters) are added, the likelihood function will converge to the correct tree (Wald 1949; Rogers 1997). Furthermore, maximum likelihood takes into account more complex modelling of datasets, and the modelling has become more computationally tractable through faster algorithmic design and faster computer processors (Paninski, et al. 2004). However, in exact opposition to maximum parsimony, maximum likelihood is more likely to separate highly divergent species, leading to long branch repulsion (Siddall 1998). MEGA X (Kumar, et al. 2018), RAxML

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2019 doi:10.20944/preprints201910.0086.v1

(Stamatakis 2014), IQ-TREE (Nguyen, et al. 2015) and PHYLIP (Retief 2000) are commonly used to recover phylogenies using maximum likelihood.

iii. Bayesian Inference

Bayesian phylogenetic estimates use posterior probabilities of a distribution of trees calculated with Markov Chain Monte Carlo (MCMC) techniques to evaluate tree probabilities. Bayesian inference adds statistical support to phylogenies and empirically produces more accurate trees in simulations. However, Bayesian inference is highly sensitive to prior probabilities (Huelsenbeck, et al. 2002). How Bayesian techniques compare to other phylogenetic methods is addressed by Yang and Rannala (2012) and popular Bayesian techniques are implemented in MrBayes (Ronquist, et al. 2012; Ling, et al. 2016) and BEAST2 (Bouckaert, et al. 2014).

iv. Distance-based and Alignment-free

Distance-based phylogenies use techniques such as neighbor-joining to quickly produce relatively good trees and are often used as a starting point for phylogenetic analyses using other methods. Neighbor-joining decomposes a star tree by taking the two closest taxa based on the number of character changes between them, pairing them together, recalculating weights based on the shortest distance between the paired species and all other species, and repeating this process until all taxa are paired. Although this technique is computationally fast, compressing the sequences into distances loses information and phylogenetic reliability is difficult to ascertain from highly divergent sequences (Holder and Lewis 2003). However, distance-based methods are frequently used when sequence alignments are not available or in whole genome comparisons. Since genome assembly and multiple sequence alignment affect phylogenies more than the technique used to recover the phylogeny, alignment-free methods attempt to recover shared phylogenetic history without an alignment by comparing basic characteristics of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2019 doi:10.20944/preprints201910.0086.v1

genomes (i.e., GC content, k-mer counts, codon usages, etc.) (Chan, et al. 2014). Broadly, alignment-free approaches can be classified into three main groups. The first group analyzes the frequency of words with a certain length (e.g., FFP (Sims, et al. 2009; Jun, et al. 2010) and CVTree (Zuo and Hao 2015)). The second group matches lengths of overlapping sequences (e.g., ACS (Ulitsky, et al. 2006), KMACS (Leimeister and Morgenstern 2014), and Kr (Haubold, et al. 2009)). The last group calculates informational content between sequences (e.g., Co-phylog (Yi and Jin 2013), FSWM (Leimeister, et al. 2017), andi (Haubold, et al. 2015), CAM (Miller, McKinnon, et al. 2019a), and codon pairing (Miller, McKinnon, et al. 2019b)). These techniques are still being developed, and new software packages are constantly updated to recover more robust trees.

3. Assessing the Phylogenetic Tree

Bootstrapping is a common technique to assess the robustness of a phylogeny by randomly sampling characters with replacement and determining if the recovered phylogenetic tree changes. Proponents of bootstrapping point to its ability to uncover the phylogenetic signal under the noise of phylogenetically uninformative characters. Bootstrapping also has statistical properties that allow a confidence value to be placed on clades (Sanderson 1995). On the other hand, critics of bootstrapping point to the statistical assumptions that are violated in DNA characters because DNA characters cannot be considered independently and identically distributed (Sanderson 1995). Furthermore, a bootstrap proportion is generally unbiased but highly imprecise, meaning the bootstrap number can give high confidence that the data support a clade even if the clade is not real (Hillis and Bull 1993).

Biological Construct of Codon Usage Bias

Phylogenomic studies have recently used codon usage bias to recover species relationships with or without ortholog annotations. Various codon usage biases appear to track speciation events and can

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2019 doi:10.20944/preprints201910.0086.v1

cause gene expression to either increase or decrease (Quax, et al. 2015). Furthermore, codon usage biases affect protein and RNA folding, which affects transcription and translational efficiency, as well as gene expression. Although genetic drift drives global codon usages, the majority of codon usage bias within individual genes is influenced by translational selection (Labella, et al. 2019). Figure 1 outlines how codon biases affect protein levels.

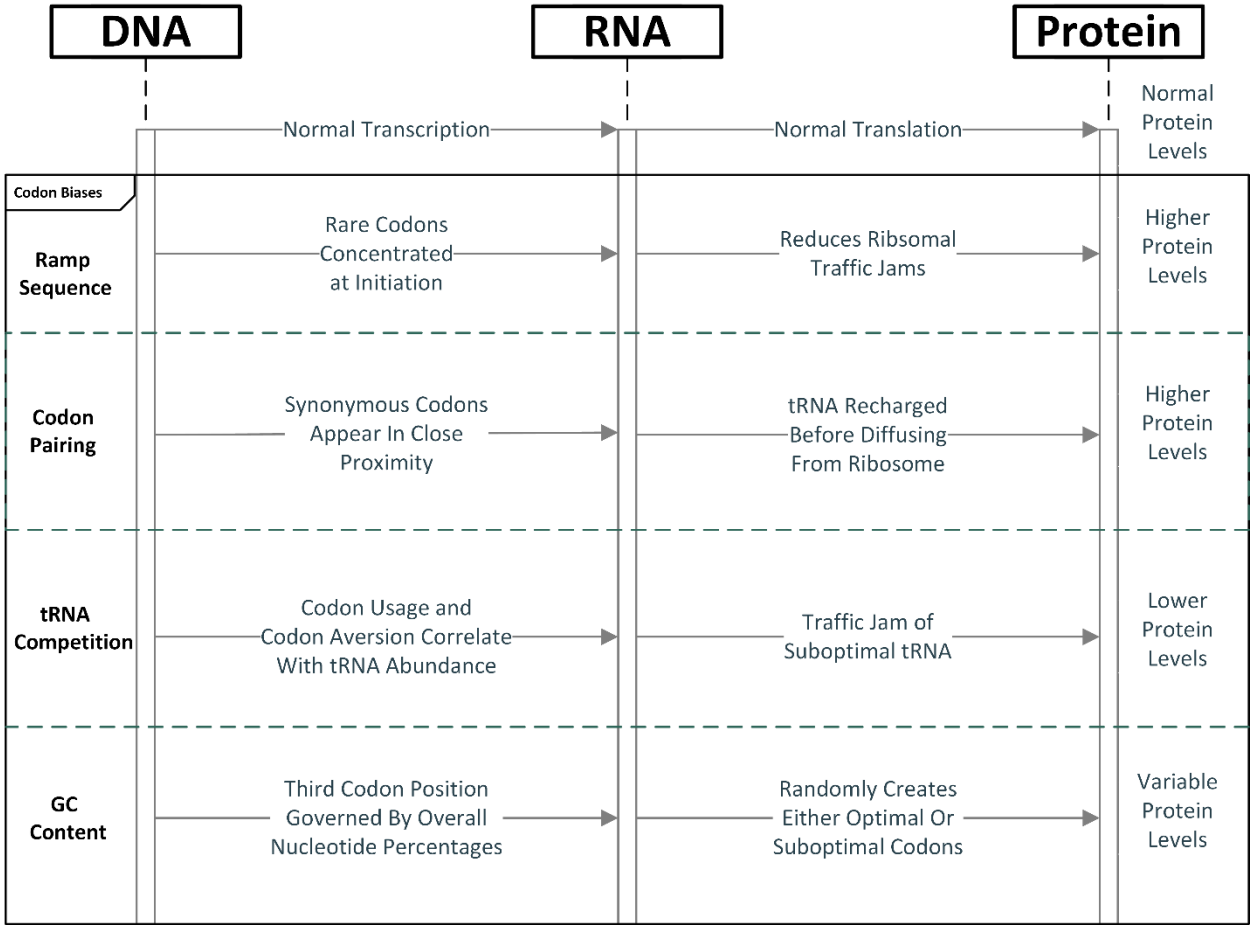

Figure 1: How Codon Usage Biases Affect Protein Levels. Many types of codon usage biases directly affect DNA, RNA, and protein secondary structure. They also affect transcription and translational efficiency. The mechanisms by which ramp sequences, codon pairing, tRNA competition, and the GC nucleotide composition affect protein levels are depicted.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2019 doi:10.20944/preprints201910.0086.v1

1. Codon Usage Metrics

Several measurements of codon usage preferences facilitate comparing codons. Originally, the Codon Adaptation Index compared the relative codon usage of the most commonly used codons within highly expressed genes (Sharp and Li 1986). Soon thereafter, the effective number of codons quantified the difference in codon usage versus the expected usage if all synonymous codons were used equally (Wright 1990). Because of their simplicity, the effective number of codons and codon adaptation index are still widely used techniques. However, those methods oversimplify the dynamics of codon usage. The tRNA adaptation index (tAI) takes into account the complex relationship between tRNA and codons by using tRNA copy number, gene length, number of codons, and the preponderance of tRNA wobble to determine codon optimality (dos Reis, et al. 2003; dos Reis, et al. 2004). Building on tAI, the normalized translational efficiency (nTE) measurement balances tRNA supply and demand on codon usage and considers cellular tRNA dynamics. A codon is considered "optimal" if the relative supply of its cognate tRNAs exceeds the codon's usage (Pechmann and Frydman 2013). Unfortunately, tAI and nTE require data that are not always available in a species or gene, thus limiting their use across the Tree of Life.

2. Biological Implications of Codon Usage Bias

a. Selection toward decreased translational efficiency

Occasionally, suboptimal codons are more beneficial to cells because they slow translation and allow for more precise, deliberate gene translation. Codon usage bias affects mRNA secondary structure so strongly that local mRNA secondary structure can be used to predict codon usage in highly expressed genes (Trotta 2013). Highly expressed genes also have a ramp of 30-50 slowly-translated, rare codons at the 5' end of most protein coding sequences (Tuller, et al. 2010) that serves to evenly space ribosomes (Shah, et al. 2013) and reduce mRNA secondary structure (Goodman, et al. 2013) at translation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2019 doi:10.20944/preprints201910.0086.v1

initiation. A comprehensive analysis of ramp sequences from all domains of life, as well as a method to extract ramp sequences from individual genes is presented in Miller, Brase, et al. (2019).

Suboptimal codons are also used in genes that are regulated by the cell cycle. Since tRNA expression levels are highest during the G2 phase, suboptimal codon usage for genes expressed during this phase is also highest. The G1 phase has the lowest tRNA expression, and genes expressed during G1 have a tendency toward optimal codon usage (Frenkel-Morgenstern, et al. 2012).

Codon usage bias in various bacteria is also associated with species lifestyle (Carbone, et al. 2005; Botzman and Margalit 2011). For cyanobacteria (photosynthetic bacteria), selection toward sub-optimal codon usage produces the circadian clock conditionality, where the circadian clock is expressed only under certain environmental conditions where cyanobacteria are not intrinsically robust (Xu, et al. 2013). Similarly, the pathogenicity and habitat of *Actinobacteria* (High GC gram positive bacteria important for soil systems) also influence codon usage, with aerobic species varying significantly from anaerobic species, and pathogenic species varying significantly from non-pathogenic species (Lal, et al. 2016). In each case, codon usage explains bacterial adaptation to their environment.

b. Selection toward increased translational efficiency

Highly expressed genes tend to use more optimal codons after the ramp sequence to increase gene translation because optimal codons are translated faster (Quax, et al. 2015). Faster translation is due to decreased wobble interactions, increased optimal tRNA composition, and decreased competition from synonymous codons within a gene. (Brule and Grayhack 2017) Selective pressures for protein expression also act on mRNA sequences to optimize co-translational folding within polypeptides in over 90% of high expression genes and about 80% of low expression genes (Pechmann and Frydman 2013). Furthermore,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2019 doi:10.20944/preprints201910.0086.v1

gene body methylation is strongly correlated with codon bias, and appears to systematically replace CpG bearing codons, potentially influencing optimal codon establishment (Dixon, et al. 2016).

Recharging a tRNA while the ribosome is still attached to the mRNA strand is another strategy used to increase translational efficiency and decrease overall resource utilization. Co-tRNA codon pairing occurs when two non-identical codons that encode the same amino acid are located in close proximity to each other in a gene. Identical codon pairing occurs when identical codons are located in close proximity in a gene. Co-tRNA and identical codon pairing are mechanisms that a cell uses to reuse a tRNA by recharging the tRNA with an amino acid before the tRNA diffuses, and increases translational speed by approximately 30% (Cannarozzi, et al. 2010). Although co-tRNA codon pairing occurs more prominently in eukaryotes and identical codon pairing occurs prominently in bacteria (Shao, et al. 2012) and archaea (Zhang, et al. 2013), both co-tRNA and identical codon pairing are phylogenetically conserved in all domains of life (Miller, McKinnon, et al. 2019b).

Other systematic biases also influence codon choice. Background dinucleotide substitution biases from GC to AT and AT to GC often coincide with shifts in optimal codons (Sun, et al. 2017). Even under sustained selective pressure, GC content at the third codon position is highly correlated with overall GC content in a gene, suggesting that optimal codons are affected by overall GC content (Sun, et al. 2017). In an analysis of 65 eukaryotes and prokaryotes, GC content accounted for 76.7% of amino acid variation (Li, et al. 2015). A summary of mechanisms that affect codon usage bias are shown in Table 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2019 doi:10.20944/preprints201910.0086.v1

Table 1: Mechanisms Affecting Codon Usage Bias

Name	Location/ Domain	Description
Ramp Sequence	30-50 nucleotides downstream of start codon	The ramp sequence consists of rare, slowly translated codons that increase ribosomal spacing, reduce mRNA secondary structure, and slow initial translation.
Co-tRNA pairing	More prominent in eukaryotes. Phylogenetically conserved in all domains of life	tRNA are recharged with amino acids for synonymous codon translation when synonymous codons are in close proximity to each other. Recharging allows the tRNA to stay attached to the ribosome and significantly increases translation efficiency.
Identical Codon Pairing	All domains of life	tRNA are recharged with amino acids for identical codon translation when identical codons are in close proximity to each other. Recharging allows the tRNA to stay attached to the ribosome and significantly increases translation efficiency.
tRNA competition	Eukarya, bacteria, and archaea	Cognate, near-cognate, and non-cognate tRNA may attempt to bind to an mRNA codon. If relatively few cognate tRNA are available, translation will slow because other tRNA attempt to bind to the same codon. This process is essential for translation elongation, efficiency, and accuracy (Zur and Tuller 2016).
GC Content	All domains of life	Overall GC content in a gene is highly correlated with GC content at the third codon position. GC content influences over two-thirds of codon variation.

Codon Usage Bias in Phylogenetic Systematics

As expected, random mutations are less likely to occur in conserved genomic regions because they can adversely affect fitness, and codon usage bias is less likely to be affected by random mutations than expected based on genomic mutation rates (Castle 2011). Many phylogenetic studies attempt to account for codon usage biases by determining its importance in species relatedness.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2019 doi:10.20944/preprints201910.0086.v1

1. Codon Usage in Maximum Likelihood

Limited codon substitution models have been used for decades in maximum likelihood estimates. However, until recently, a full 61 x 61 codon matrix was too computational intensive to apply to more than a few species and genes (Anisimova and Kosiol 2009). Somewhat surprisingly, after a 61 x 61 codon matrix became computationally viable, it was determined that the full matrix is not always optimal because models that use a fixed codon mutation rate for phylogenetic tree reconstruction fit the data better than a variable codon substitution rate. The apparent variation in codon substitution is actually caused by variable selection against amino acid substitutions in the regions used to develop the model, specifically mitochondria, chloroplast, and hemagglutinin proteins (Miyazawa 2013). Maximum likelihood estimates that use codon models outperform a parsimony analysis only when codon usage is highly skewed and is not affected by asymmetry in substitution rates (approach validated using *Drosophila*) (Akashi, et al. 2007).

Because full codon models are computationally intensive and do not always elucidate more information than simpler models, common likelihood approaches use nonsynonymous to synonymous mutation rates per site (d_N/d_S) instead of the complete codon model. If the codon usage bias is strongly conserved, then d_S will decrease and d_N/d_S will increase within a population. The d_N/d_S ratio was used in *Drosophila* lineages, and helped determine that the *Notch* locus had evolved to include suboptimal codons (Nielsen, et al. 2007). Using 158 orthologous genes, maximum likelihood also detected a strong shift from suboptimal to optimal codons in two lineages of *Populus* (Ingvarsson 2008). Detecting the cause of such shifts in codon usage is important for determining the biological significance of mutations. SCUMBLE (Synonymous Codon Usage Bias Maximum Likelihood Estimation) uses a model inspired by statistical physics to identify different sources of codon bias including selection and mutation (Kloster and Tang 2008). SCUMBLE is also used as a filter to identify regions with insufficient information for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2019 doi:10.20944/preprints201910.0086.v1

analysis. This technique helped determine that natural selection shaped codon biases in *Strongylocentrotus purpuratus* (purple sea urchin) by limiting the analysis to only regions with sufficient support (Kober and Pogson 2013). Shifts in mutation and selection rates allow the evolutionary history of species to be recovered using this method.

2. Violations of Maximum Likelihood Statistical Properties in a Codon Model

Many of the assumptions of the statistical properties in maximum likelihood are violated by a codon model. For instance, species are constrained to taxon-specific pools of tRNA, and triplets in coding sequences are not independent. Algorithms with statistical properties that require character independence, such as maximum likelihood, violate that rule for genetic data (Christianson 2005). Furthermore, the codon model assumption of homogeneity of codon composition leads to seriously biased phylogenetic estimations when that assumption is violated (Inagaki and Roger 2006).

Horizontal gene transfer is another important mechanism in evolution and complicates phylogenetic analyses in bacteria because 81 ± 15% of genes have been laterally transferred among bacteria at some point in their evolutionary history (Dagan, et al. 2008). Common transposable elements in eukaryotes also arose from horizontal gene transfer, with over 50% of some mammalian genomes originally arising from horizontal gene transfer (Ivancevic, et al. 2018). Detecting horizontal gene transfer has been challenging, and codon bias is a poor indicator of horizontal transmission, normally underestimating the effects of lateral transfer (Koski, et al. 2001; Tuller 2011; Friedman and Ely 2012). However, codon composition is an excellent indicator of whether a gene will become fixed in a species after a lateral transfer event (Tuller 2011). The concept of horizontal gene transfer not only complicates a general phylogenetic analysis, but suggests that a standard bifurcating tree might not be the best choice in analyses of bacteria or archaea (Koonin and Wolf 2008). Although it is known that codons (and DNA in

general) do not strictly follow many of the assumptions of phylogenetic analyses, the bifurcating tree is still the most widely used phylogenetic representation, and generally depicts statements of homology even when some assumptions are violated.

3. Codon Usage in Viruses

Another purpose of phylogenies is to predict the pathogenicity of viruses and viral interactions with their hosts. Bee-infecting viruses have strong correlations in their codon usages with their hosts, and the infected insects' codon usage similarity follows the insect phylogeny (Chantawannakul and Cutler 2008). Furthermore, human-host viruses tend to share the same codon usages as proteins expressed in tissues that the viruses infect (Miller, Hippen, Wright, et al. 2017). More specifically, the key determinant in codon patterns within herpesviruses were the overall GC content, GC content at the 3rd codon position, and gene length (Roychoudhury and Mukherjee 2010). In contrast, mutation played a larger role in Zika viruses, with higher frequencies of A-ending codons (Cristina, et al. 2016). However, evidence of natural selection in Zika viruses also suggest that they evolved host- and vector-specific codon usage patterns to successfully replicate in various hosts and vectors (Butt, et al. 2016). In hepatitis C, preferred codon usages did not always match the phylogenetic histories of the viruses as determined by sequence similarity, indicating that codon usage might provide additional information not identified in common phylogenomic approaches (Mortazavi, et al. 2016).

4. Successful Implementations of Codon Usage Bias in Phylogenetics

Beyond analyzing pathogenicity, phylogenetic inferences using codon usage bias from all domains of life have successfully uncovered several interesting biological principles. One study found compositional differences in codon usage between monocots (flowering plants whose seeds contain one embryonic leaf) and dicots (flowering plants whose seeds contains two embryonic leaves), where monocots had

lower DNA background compositional bias, but higher codon usage bias than dicots (Camiolo, et al. 2015). Another technique used a distance-based clustering method of codon usage weighted by nucleotide base bias per position (i.e., the frequency of a codon over the product of the frequency of the nucleotide at the first, second, and third positions) to recover the phylogeny of closely related *Ectocarpales* (brown algae) (Das, et al. 2005). The phylogenetic signal of codon usage was not limited to nuclear DNA, and mitochondrial synonymous codon usage in plants was associated with intron number that mirrored species evolution (Xu, et al. 2015).

Creative attempts at analyzing codon usage have also proven fruitful. A binary representation of codon aversion (i.e., creating a character matrix based on codons which are not used in an ortholog) successfully recover the phylogeny of various tetrapods, showing that complete codon aversion is also conserved (Miller, Hippen, Belyeu, et al. 2017). That study also found that stop codon usage had the highest phylogenetic signal (Miller, Hippen, Belyeu, et al. 2017), meaning a codon matrix of 64 x 64 (the probability of all codons including the stop codons transitioning to all other codons) might be better than the traditional 61 x 61 codon matrix in a likelihood framework. Codon aversion has also been used in an alignment-free context by comparing sets of codon tuples found in a genome, where each tuple is a list of codons not used in a gene (Miller, McKinnon, et al. 2019a). A similar technique used codon pairing and codon pairs (i.e., the same codon being used within a ribosomal window) and was phylogenetically informative in both alignment-free and parsimony frameworks (Miller, McKinnon, et al. 2019b).

Other studies map codon usage in a particular gene across a reference phylogeny. This technique can produce meaningful representations of codon transitions across genes. Mapping the codon usage bias of a gene tree to a species tree revealed purifying selection among the actin-depolymerizing factor/cofilin

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2019 doi:10.20944/preprints201910.0086.v1

(ADF/CFL) gene family (Roy-Zokan, et al. 2015). This technique also showed that codon usage is significantly correlated with gene age within metazoan genomes (Prat, et al. 2009). Codon aversion in all domains of life was also mapped to the Open Tree of Life (OTL) (Hinchliff, et al. 2015) and showed that codon aversion follows established species relationships more closely than expected by random chance (Miller, McKinnon, et al. 2019c).

Concluding Remarks

Codon usage bias continues to be widely studied in a phylogenetic construct. However, its application in phylogenomics remains limited by its applicability in current phylogenomic techniques. While some applications attempt to incorporate codon usage bias as a novel character state in phylogenetics or in a maximum likelihood framework, many of the key attributes of codon bias remain unexplored. For instance, although the ramp of slowly translated codons has been identified, it is unknown if the ramp sequence is more or less phylogenetically conserved than the rest of the gene sequence.

In addition, although it is known that tRNA supply and demand is correlated to codon usage, a model does not currently exist to assess tRNA supply and demand in a maximum likelihood framework. Future codon analyses will necessitate more complete datasets with accurate tRNA expression values in different tissues and species. A more robust dataset of tRNA expression values would also facilitate codon model analyses. Furthermore, since codons are used to regulate gene translational efficiency, codon models might require gene expression data in addition to the full (or reduced) codon matrix.

Codon usage bias is an exciting biological principle that has not been fully utilized in phylogenetic systematics. Few likelihood methods use codon bias, and many aspects of the ramp sequence, co-tRNA codon pairing, gene expression, and tRNA expression have yet to be explored. Although codon usage

bias has been shown to be phylogenetically conserved, many of the biological principles surrounding codon usage bias have yet to be fully utilized in phylogenomics. We propose that more research into codon usage bias and its phylogenetic conservation will be beneficial to future phylogenomic studies by providing researchers with more robust phylogenetic trees.

Acknowledgements

We appreciate the continued support of Brigham Young University.

Authors' Contributions

JM and PR conceived the idea. JM led the writing of the manuscript. All authors contributed critically to the drafts, edited the drafts, and gave final approval for publication.

References

Akashi H, Goel P, John A. 2007. Ancestral inference and the study of codon bias evolution: implications for molecular evolutionary analyses of the Drosophila melanogaster subgroup. PLoS One 2:e1065.

Anisimova M, Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 26:255-271.

Bertolani R, Guidetti R, Marchioro T, Altiero T, Rebecchi L, Cesari M. 2014. Phylogeny of Eutardigrada: New molecular data and their morphological support lead to the identification of new evolutionary lineages. Molecular Phylogenetics and Evolution 76:110-126.

Botzman M, Margalit H. 2011. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. Genome Biol 12:R109.

Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10:e1003537.

381 Brule CE, Grayhack EJ. 2017. Synonymous Codons: Choose Wisely for Expression. Trends Genet 33:283-
382 297.
383 Butt AM, Nasrullah I, Qamar R, Tong Y. 2016. Evolution of codon usage in Zika virus genomes is host and
384 vector specific. Emerging Microbes &Amp; Infections 5:e107.
385 Camiolo S, Melito S, Porceddu A. 2015. New insights into the interplay between codon bias
386 determinants in plants. DNA Research:dsv027.
387 Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y.
388 2010. A role for codon order in translation dynamics. Cell 141:355-367.
389 Carbone A, Kepes F, Zinovyev A. 2005. Codon bias signatures, organization of microorganisms in codon
390 space, and lifestyle. Mol Biol Evol 22:547-561.
391 Castle JC. 2011. SNPs occur in regions with less genomic sequence conservation. PLoS One 6:e20660.
392 Chan CX, Bernard G, Poirion O, Hogan JM, Ragan MA. 2014. Inferring phylogenies of evolving sequences
393 without multiple sequence alignment. Sci Rep 4:6504.
394 Chantawannakul P, Cutler RW. 2008. Convergent host-parasite codon usage between honeybee and bee
395 associated viral genomes. J Invertebr Pathol 98:206-210.
396 Christianson ML. 2005. Usuage patterns distort phylogenies from or of DNA sequences. 92:1221-1233.
397 Crick FH. 1966. Codon--anticodon pairing: the wobble hypothesis. J Mol Biol 19:548-555.
398 Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. 1961. General nature of the genetic code for proteins.
399 Nature 192:1227-1232.
400 Cristina J, Fajardo A, Sonora M, Moratorio G, Musto H. 2016. A detailed comparative analysis of codon
401 usage bias in Zika virus. Virus Res 223:147-152.
402 Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer
403 in prokaryote genome evolution. Proceedings of the National Academy of Sciences 105:10039-10044.

404 Das S, Chakrabarti J, Ghosh Z, Sahoo S, Mallick B. 2005. A new measure to study phylogenetic relations
405 in the brown algal order Ectocarpales: The "codon impact parameter". Journal of Biosciences 30:699-
406 709.
407 Daugelaite J, O' Driscoll A, Sleator RD. 2013. An Overview of Multiple Sequence Alignments and Cloud
408 Computing in Bioinformatics. ISRN Biomathematics 2013:14.
409 Dixon GB, Bay LK, Matz MV. 2016. Evolutionary Consequences of DNA Methylation in a Basal Metazoan.
410 33:2285-2293.
411 dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for
412 translational selection. Nucleic Acids Res 32:5036-5044.
413 dos Reis M, Wernisch L, Savva R. 2003. Unexpected correlations between gene expression and codon
414 usage bias from microarray data for the whole Escherichia coli K-12 genome. Nucleic Acids Res 31:6976-
415 6985.
416 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic
417 Acids Res 32:1792-1797.
418 Farris JS. 2008. Parsimony and explanatory power. Cladistics 24:825-847.
419 Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading.
420 Systematic Biology 27:401-410.
421 Frenkel-Morgenstern M, Danon T, Christian T, Igarashi T, Cohen L, Hou Y-M, Jensen LJ. 2012. Genes
422 adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels.
423 Molecular Systems Biology 8:572-572.
424 Friedman R, Ely B. 2012. Codon usage methods for horizontal gene transfer detection generate an
425 abundance of false positive and false negative results. Curr Microbiol 65:639-642.
426 Goloboff PA, Farris JS, Nixon KC. 2005. TNT: Tree Analysis Using New Technology. 54:176-178.

427 Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-terminal codon bias in bacterial
428 genes. Science 342:475-479.
429 Gutman GA, Hatfield GW. 1989. Nonrandom utilization of codon pairs in Escherichia coli. Proc Natl Acad
430 Sci U S A 86:3699-3703.
431 Haszprunar G. 1992. The types of homology and their significance for evolutionary biology and
432 phylogenetics. Journal of Evolutionary Biology 5:13-24.
433 Haubold B, Klotzl F, Pfaffelhuber P. 2015. andi: fast and accurate estimation of evolutionary distances
434 between closely related genomes. Bioinformatics 31:1169-1175.
435 Haubold B, Pfaffelhuber P, Domazet-Loso M, Wiehe T. 2009. Estimating mutation distances from
436 unaligned genomes. J Comput Biol 16:1487-1500.
437 Hillis DM, Bull JJ. 1993. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in
438 Phylogenetic Analysis. Systematic Biology 42:182-192.
439 Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT,
440 Gazis R, et al. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl
441 Acad Sci U S A 112:12764-12769.
442 Holder M, Lewis PO. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet
443 4:275-284.
444 Huelsenbeck JP, Larget B, Miller RE, Ronquist F. 2002. Potential applications and pitfalls of Bayesian
445 inference of phylogeny. Syst Biol 51:673-688.
446 Inagaki Y, Roger AJ. 2006. Phylogenetic estimation under codon models can be biased by codon usage
447 heterogeneity. Mol Phylogenet Evol 40:428-434.
448 Ingvarsson PK. 2008. Molecular evolution of synonymous codon usage in Populus. BMC Evol Biol 8:307.
449 Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. 2018. Horizontal transfer of BovB and L1
450 retrotransposons in eukaryotes. Genome Biol 19:85.

451 Jun S-R, Sims GE, Wu GA, Kim S-H. 2010. Whole-proteome phylogeny of prokaryotes by feature
452 frequency profiles: An alignment-free method with optimal feature resolution. Proceedings of the
453 National Academy of Sciences 107:133-138.
454 Katoh K, Standley DM. 2014. MAFFT: iterative refinement and additional methods. Methods Mol Biol
455 1079:131-146.
456 Kloster M, Tang C. 2008. SCUMBLE: a method for systematic and accurate detection of codon usage bias
457 by maximum likelihood estimation. Nucleic Acids Res 36:3819-3827.
458 Kober KM, Pogson GH. 2013. Genome-Wide Patterns of Codon Bias Are Shaped by Natural Selection in
459 the Purple Sea Urchin, <em>Strongylocentrotus purpuratus</em&gt. G3:
460 Genes|Genomes|Genetics 3:1069.
461 Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309-338.
462 Koonin EV, Wolf YI. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the
463 prokaryotic world. Nucleic Acids Research 36:6688-6719.
464 Koski LB, Morton RA, Golding GB. 2001. Codon bias and base composition are poor indicators of
465 horizontally transferred genes. Mol Biol Evol 18:404-412.
466 Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis
467 across Computing Platforms. Mol Biol Evol 35:1547-1549.
468 Labella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A. 2019. Variation and selection on codon
469 usage bias across an entire subphylum. PLOS Genetics 15:e1008304.
470 Lal D, Verma M, Behura SK, Lal R. 2016. Codon usage bias in phylum Actinobacteria : relevance to
471 environmental adaptation and host pathogenicity. Research in Microbiologoy 167:669-677.
472 Leimeister CA, Morgenstern B. 2014. Kmacs: the k-mismatch average common substring approach to
473 alignment-free sequence comparison. Bioinformatics 30:2000-2008.

474 Leimeister CA, Sohrabi-Jahromi S, Morgenstern B. 2017. Fast and accurate phylogeny reconstruction
475 using filtered spaced-word matches. Bioinformatics 33:971-979.
476 Li J, Zhou J, Wu Y, Yang S, Tian D. 2015. GC-Content of Synonymous Codons Profoundly Influences Amino
477 Acid Usage. G3 (Bethesda) 5:2027-2036.
478 Ling C, Hamada T, Gao J, Zhao G, Sun D, Shi W. 2016. MrBayes tgMC3++: A High Performance and
479 Resource-Efficient GPU-Oriented Phylogenetic Analysis Method. IEEE/ACM Trans Comput Biol Bioinform
480 13:845-854.
481 Magis C, Taly JF, Bussotti G, Chang JM, Di Tommaso P, Erb I, Espinosa-Carrasco J, Notredame C. 2014. T-
482 Coffee: Tree-based consistency objective function for alignment evaluation. Methods Mol Biol 1079:117-
483 129.
484 Miller JB, Brase LR, Ridge PG. 2019. ExtRamp: a novel algorithm for extracting the ramp sequence based
485 on the tRNA adaptation index or relative codon adaptiveness. Nucleic Acids Res.
486 Miller JB, Hippen AA, Belyeu JR, Whiting MF, Ridge PG. 2017. Missing something? Codon aversion as a
487 new character system in phylogenetics. Cladistics:n/a-n/a.
488 Miller JB, Hippen AA, Wright SM, Morris C, Ridge PG. 2017. Human viruses have codon usage biases that
489 match highly expressed proteins in the tissues they infect. Biomedical Genetics and Genomics 2.
490 Miller JB, McKinnon LM, Whiting MF, Ridge PG. 2019a. CAM: An alignment-free method to recover
491 phylogenies using codon aversion motifs. PeerJ Preprints 7:e27756v27751.
492 Miller JB, McKinnon LM, Whiting MF, Ridge PG. 2019b. Codon Pairs are Phylogenetically Conserved:
493 Codon pairing as a new class of phylogenetic characters. bioRxiv:654947.
494 Miller JB, McKinnon LM, Whiting MF, Ridge PG. 2019c. Codon Use and Aversion is Largely
495 Phylogenetically Conserved Across the Tree of Life. bioRxiv:649590.
496 Miyazawa S. 2013. Superiority of a mechanistic codon substitution model even for protein sequences in
497 Phylogenetic analysis.1-10.

498 Mortazavi M, Zarenezhad M, Alavian SM, Gholamzadeh S, Malekpour A, Ghorbani M, Torkzadeh Mahani
499 M, Lotfi S, Fakhrzad A. 2016. Bioinformatic Analysis of Codon Usage and Phylogenetic Relationships in
500 Different Genotypes of the Hepatitis C Virus. Hepatitis Monthly 16:e39196.
501 Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic
502 algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268-274.
503 Nielsen R, Bauer DuMont VL, Hubisz MJ, Aquadro CF. 2007. Maximum likelihood estimation of ancestral
504 codon usage bias parameters in Drosophila. Mol Biol Evol 24:228-235.
505 Pais FS, Ruy Pde C, Oliveira G, Coimbra RS. 2014. Assessing the efficiency of multiple sequence alignment
506 programs. Algorithms Mol Biol 9:4.
507 Paninski L, Pillow JW, Simoncelli EP. 2004. Maximum likelihood estimation of a stochastic integrate-and-
508 fire neural encoding model. Neural Comput 16:2533-2561.
509 Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures
510 of cotranslational folding. Nat Struct Mol Biol 20:237-243.
511 Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving
512 Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLOS Biology 9:e1000602.
513 Post LE, Strycharz GD, Nomura M, Lewis H, Dennis PP. 1979. Nucleotide sequence of the ribosomal
514 protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. Proc Natl
515 Acad Sci U S A 76:1697-1701.
516 Prat Y, Fromer M, Linial N, Linial M. (Prat2009 co-authors). 2009. Codon usage is associated with the
517 evolutionary age of genes in metazoan genomes. BMC Evolutionary Biology 9:285.
518 Quax TE, Claassens NJ, Soll D, van der Oost J. 2015. Codon Bias as a Means to Fine-Tune Gene
519 Expression. Mol Cell 59:149-161.
520 Retief JD. 2000. Phylogenetic analysis using PHYLIP. Methods Mol Biol 132:243-258.

521 Rogers JS. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from
522 nucleotide sequences. Syst Biol 46:354-357.
523 Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA,
524 Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a
525 large model space. Syst Biol 61:539-542.
526 Roy-Zokan EM, Dyer KA, Meagher RB. 2015. Phylogenetic Patterns of Codon Evolution in the ACTIN-
527 DEPOLYMERIZING FACTOR/COFILIN (ADF/CFL) Gene Family. PLoS One 10:e0145917.
528 Roychoudhury S, Mukherjee D. 2010. A detailed comparative analysis on the overall codon usage
529 pattern in herpesviruses. Virus Res 148:31-43.
530 Sanderson MJ. 1995. Objections to Bootstrapping Phylogenies: A Critique. Systematic Biology 44:299-
531 320.
532 Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. 2013. Rate-limiting steps in yeast protein translation.
533 Cell 153:1589-1601.
534 Shao ZQ, Zhang YM, Feng XY, Wang B, Chen JQ. 2012. Synonymous codon ordering: a subtle but
535 prevalent strategy of bacteria to improve translational efficiency. PLoS One 7:e33547.
536 Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular
537 organisms. J Mol Evol 24:28-38.
538 Siddall ME. 1998. Success of Parsimony in the Four-Taxon Case: Long-Branch Repulsion by Likelihood in
539 the Farris Zone. Cladistics 14:209-220.
540 Sievers F, Higgins DG. 2014. Clustal omega. Curr Protoc Bioinformatics 48:3 13 11-16.
541 Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences.
542 Protein Sci 27:135-145.
543 Sims GE, Jun SR, Wu GA, Kim SH. 2009. Alignment-free genome comparison with feature frequency
544 profiles (FFP) and optimal resolutions. Proc Natl Acad Sci U S A 106:2677-2682.

545 Soltis DE, Soltis PS. 2003. The Role of Phylogenetics in Comparative Genetics. Plant Physiology 132:1790-
546 1800.
547 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
548 phylogenies. Bioinformatics 30:1312-1313.
549 Sun Y, Tamarit D, Andersson SGE. 2017. Switches in Genomic GC Content Drive Shifts of Optimal Codons
550 under Sustained Selection on Synonymous Sites. Genome Biol Evol 9:2560-2579.
551 Tekaia F. 2016. Inferring Orthologs: Open Questions and Perspectives. Genomics Insights 9:17-28.
552 Trotta E. 2013. Selection on codon bias in yeast: a transcriptional hypothesis. Nucleic Acids Res 41:9382-
553 9395.
554 Tuller T. 2011. Codon bias, tRNA pools and horizontal gene transfer. Mob Genet Elements 1:75-77.
555 Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010.
556 An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell
557 141:344-354.
558 Ulitsky I, Burstein D, Tuller T, Chor B. 2006. The average common substring approach to phylogenomic
559 reconstruction. J Comput Biol 13:336-350.
560 Wald A. 1949. Note on the Consistency of the Maximum Likelihood Estimate.595-601.
561 Wilgenbusch JC, Swofford D. 2003. Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics
562 Chapter 6:Unit 6 4.
563 Wright F. 1990. The 'effective number of codons' used in a gene. Gene 87:23-29.
564 Xu W, Xing T, Zhao M, Yin X, Xia G, Wang M. 2015. Synonymous codon usage bias in plant mitochondrial
565 genes is associated with intron number and mirrors species evolution. PLoS One 10:e0131508.
566 Xu Y, Ma P, Shah P, Rokas A, Liu Y, Johnson CH. 2013. Non-optimal codon usage is a mechanism to
567 achieve circadian clock conditionality. Nature 495:116-120.
568 Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. Nat Rev Genet 13:303-314.

569 Yi H, Jin L. 2013. Co-phylog: an assembly-free phylogenomic approach for closely related organisms.
570 Nucleic Acids Res 41:e75.
571 Zhang YM, Shao ZQ, Yang LT, Sun XQ, Mao YF, Chen JQ, Wang B. 2013. Non-random arrangement of
572 synonymous codons in archaea coding sequences. Genomics 101:362-367.
573 Zuo G, Hao B. 2015. CVTree3 Web Server for Whole-genome-based and Alignment-free Prokaryotic
574 Phylogeny and Taxonomy. Genomics Proteomics Bioinformatics 13:321-331.
575 Zur H, Tuller T. 2016. Predictive biophysical modeling and understanding of the dynamics of mRNA
576 translation and its evolution. Nucleic Acids Research 44:9031-9049.

577