2018-11-01

Analyzing Codon Usage and Coding Sequence Length Biases Across the Tree of Life

Justin B. Miller
Brigham Young University

Analyzing Codon Usage and Coding Sequence Length Biases Across the Tree of Life

Justin B. Miller

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Perry G. Ridge, Chair
Michael F. Whiting
Stephen R. Piccolo
John S.K. Kauwe
Mark J. Clement

Department of Biology

Brigham Young University

ABSTRACT

Analyzing Codon Usage and Coding Sequence Length Biases Across the Tree of Life

Justin B. Miller
Department of Biology, BYU
Doctor of Philosophy

Although codon usage bias has been shown to persist through non-random mutations and selection, many avenues of research into the applications of codon usage bias have remained unexplored. In this dissertation, we present several new applications of codon usage bias and their practical uses in a phylogenetic construct. We first review the literature and provide background into other software applications of codon usage bias in Chapter 1. In Chapter 2, we show that in tetrapods, codon aversion in orthologs is phylogenetically conserved. We further this analysis in Chapter 3 by exploring codon use and aversion across the Tree of Life, providing frameworks for other researchers to analyze different species subsets. We present a novel algorithm to recover species relationships using codon aversion, without regard to orthologous relationships in Chapter 4. We present several other algorithms in Chapter 5 to also recover species relationships using biases in codon pairing. Chapter 6 analyzes the relationship between codon usage bias in viruses that infect humans and proteins found in tissues that they infect. In Chapter 7, we present our discovery of a conservation in coding sequence lengths in orthologous genes that allowed us to accurately recover orthologous gene relationships and reduce overall ortholog identification runtime by over 96%. In Chapter 8 we discuss a novel algorithm for extracting a ramp of slowly-translated codons located at the beginning of gene sequences, allowing researchers to quickly identify translational bottlenecks. Finally, Chapter 9 touches on future applications of codon usage bias in phylogenetics. This dissertation represents a major vertical leap in phylogenetics by providing a framework and paradigm shift toward utilizing codon usage and coding sequence length biases in future analyses.

Keywords: codon usage bias, codon aversion, codon pairing, JustOrthologs, ExtRamp, phylogeny, tree of life, species relationships, phylogenetic systematics

ACKNOWLEDGEMENTS

My path to earning a PhD was not an individual effort. Sacrifices from family, friends, coworkers, and a university community who believe in my ability to succeed have enabled me to persevere and achieve my ambitions. Elisandra, my wife, is the main reason why I pursued a PhD. She has been immensely helpful and supportive by keeping me grounded as I began to have success in my research. Before completing my undergraduate degree in bioinformatics, I was debating if I wanted to continue my studies and earn a graduate degree. I seriously considered earning an MS degree and then going to work at a company. As I talked with Elisandra about my plans, I remember her looking at me and saying, "I know that you are capable of earning a PhD. Just go for it!" Her confidence in me made all the difference, and it helped support me through difficult projects, classes, and failed ideas. She never wavered in her support for me and for my work. After each manuscript was completed, she would beg me to read it to her, even if the paper was long and had scientific jargon. Her support was invaluable and inspired me to always think of different projects so that I could eventually read the completed publication to her.

My two children, Lunna and Oliver, have been exceptionally helpful on my journey. They help me to relax and decompress after working on difficult projects. I attribute a lot of my creativity to the joy and laughter that they bring into my life. They both help me to widen my focus beyond the work done in the lab.

I would like to emphasize the support that I have received from my parents, Kory and Trina. Throughout several difficult life circumstances, they took an oversized role in helping Elisandra, Lunna, and Oliver. They have always been an inspiration to me, and the lessons that they taught me as I was growing up have shaped me into the person who can defend this

dissertation. Their support, and the support of my brother, Kyle, is invaluable. Thank you for always being there for me and for helping me to see different applications for my research.

Throughout this process, my coworkers have become my friends and family. I continually bounce ideas off Brandon Pickett and other people in the lab. I trust them, and they make coming to work fun. I am grateful for their hard work that has led to many publications.

My mentor, Perry Ridge, has been instrumental in my career and my life. I look to him as more than a managing director, but as a close friend. He has dedicated countless hours to preparing me for my future and providing me with opportunities to develop into a competitive applicant. I feel prepared to become a professor or work in the private sector because of his mentoring, and I look forward to collaborating with him on many projects in the future.

My committee has also been very influential in my life. Every time that I meet with Mike Whiting, Keoni Kauwe, Steve Piccolo, and Mark Clement, I leave feeling more enthused about my projects. In fact, Mike Whiting is primarily responsible for the direction of my dissertation. During my first semester as a PhD student, Mike taught me the complexities of systematic theory in his class, Phylogenetic Systematics. He opened my eyes to areas of debate within systematics at a pivotal moment in my life by challenging my preconceived notions of biology. Without Mike's influence and fearless pushback on my ideas, I would undoubtedly be on a different life trajectory. I would also like to specifically thank Keoni for trusting me enough to hire me for a post-doctoral fellowship following my PhD. I look forward to continuing my work with him.

A dissertation requires the assistance of many other people and organizations, and I thank all other people who have contributed in any way to my success. I, like Bernard of Chartres and Sir Isaac Newton, realize that if my successes amount to anything, it is because I have stood on the shoulders of giants. I am grateful for each of you giants who have lifted me up to be like you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Codon Usage Bias in Phylogenetic Systematics: A Review

Justin B. Miller[1], Michael F. Whiting[1,2], 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*

Abstract

Phylogenetic systematics is the study of historical and hierarchical relationships among genes, individuals, populations, or taxa. Therefore, systematists uncover genetic or morphological traits that accurately separate species or individuals based on homology. As genetic data has become more widely accessible, various characteristics of DNA sequences have been used to establish species relatedness. One avenue of research centers on analyzing codon usage bias. Codon usage bias is based on a non-random distribution of synonymous codons between different species, different genes within the same species, and different locations within the same gene. These observations have led to two non-mutually exclusive hypotheses explaining codon usage bias: non-random mutations occur within codons, and selection for certain codons exists. We review codon usage bias as a phylogenetic character state, how it affects common phylogenomic techniques, and its future in phylogenetic systematics.

Introduction

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 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 in which morphology is combined with genetic markers is typically used in phylogenetic analyses (Bertolani et al., 2014). Although genetic data allow researchers to quickly analyze 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 are explored by Philippe et al. (2011).

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 explain how multiple codons encode the same amino acid and were presumably identical in function. However, it was soon noted that 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 was also revealed, which led to the wobble hypothesis: tRNA anticodons do not need to latch onto all three codon nucleotides during translation (Crick, 1966). It was also discovered that 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). An early review of synonymous codon usage and tRNA content was done by Ikemura (1985).

Overview of Common Phylogenomic Techniques

Homologous sequence comparisons are commonly used to identify species relationships. Homologous characters are often identified by aligning orthologous genes and detecting character state changes of amino acid residues 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 be discussed in this review.

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 of the compared species (Koonin, 2005). In contrast, paralogs may share the same function, but can arise from gene duplication or horizontal gene transfer. Paralogs are not 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 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. Some common multiple sequence aligners are T-coffee (Magis et al., 2014), MUSCLE (Edgar, 2004), CLUSTAL (Sievers and Higgins, 2014), CLUSTAL OMEGA (Sievers

and Higgins, 2018), and MAFFT (Katoh and Standley, 2014), and reviews of their capabilities can be examined in Daugelaite et al. (2013) and Pais et al. (2014).

Current Phylogenetic Tree Recovery Techniques

Maximum Parsimony

Maximum parsimony assumes that each character is equally important in determining phylogenetic relationships. Parsimony minimizes the number of homoplasious 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 for identifying phylogenies based on parsimony.

Maximum Likelihood

As opposed to parsimony, maximum likelihood requires 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 in recent years is the mathematical property of consistency, which states that as more data (phylogenetically informative characters) are added, the likelihood function will converge to a single output (Wald, 1949; Rogers, 1997). Furthermore, maximum likelihood can take 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 (Stamatakis, 2014), IQ-TREE (Nguyen et al., 2015) and PHYLIP (Retief, 2000) are commonly used to recover phylogenies using maximum likelihood.

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). Popular Bayesian techniques are implemented in the programs MrBayes (Ronquist et al., 2012; Ling et al., 2016) and BEAST2 (Bouckaert et al., 2014).

Distance-based and Alignment-free

Distance-based phylogenies, using techniques such as neighbor-joining (NJ), quickly produce relatively good trees and are often used as a starting point for phylogenetic analyses using other methods. NJ 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. Unfortunately, compressing the sequences into distances loses information

and reliable phylogenies are difficult to obtain from highly divergent sequences (Holder and Lewis, 2003). Although distance-based methods are not as optimal for aligned sequences, they have been 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 genomes (i.e., GC content, k-mer counts, codon usages, etc.) (Chan et al., 2014). These techniques are still being developed, and new software packages are constantly updated to recover more robust trees.

Bootstrapping

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).

Types of Codon Usage Bias

Codon usage bias has recently been used in phylogenomic studies with and without ortholog annotations. Various types of codon usage bias are documented as either increasing and decreasing gene expression (Quax et al., 2015). Characteristics of codon usage bias and their biological importance follow.

Measuring Bias

Since unequal distributions of codons were discovered, several measurements of codon usage preferences have been developed to facilitate the comparison of codon usages. Originally, the Codon Adaptation Index determined if two species shared the same codon usage biases by comparing 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 equally used (Wright, 1990). Because of their simplicity, the effective number of codons and codon adaptation index are still widely used techniques in measuring codon bias. However, those technique 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.

Biological Importance

Codon usage bias affects gene expression by both decreasing and increasing translational efficiency (Quax et al., 2015). See Table 1 for different causes of codon usage biases.

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 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 Chapter 8.

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). 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.

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, 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 is 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 is 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), more recent studies suggest that both co-tRNA and identical codon pairing are phylogenetically conserved in all domains of life (Chapter 5).

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).

Codon Usage Bias in Phylogenetic Systematics

As expected, random single nucleotide polymorphisms (SNPs) are less likely to occur in conserved genomic regions because they can adversely affect fitness (Castle, 2011). Furthermore, codon usage bias is less likely to be affected by SNPs than expected based on genomic mutation rates (Castle, 2011). Many studies attempt to account for codon usage biases in phylogenetic studies and determine how its usage is phylogenetically conserved. The results from these studies are outlined below.

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). As expected, using 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 rate 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 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 are important for uncovering the evolutionary history of species and can be recovered using this method.

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, meaning >50% of some mammalian genomes originally arose 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.

Codon Usage in Viruses

Another purpose of phylogenies is to describe 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 usage as proteins expressed in tissues that the viruses infect (Miller et al., 2017b). 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).

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; mitochondrial synonymous codon usage in plants was associated with intron number and 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) was able to successfully recover the phylogeny of various tetrapods, showing that complete codon aversion is also conserved (Miller et al., 2017a). That study also found that stop codon usage had the highest phylogenetic signal (Miller et al., 2017a), 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 (Chapter 4). 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 (Chapter 5).

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 (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 (Chapter 3).

Future Direction

Codon usage bias continues to be widely studied; however, its application in phylogenetic studies remains limited. 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. Alignment-free comparisons similar to work shown

in Chapter 4 can be conducted on the ramp sequence and the gene sequence excluding the ramp sequence to determine the strength of its phylogenetic signal.

In addition, although it is known that tRNA supply and demand is not equal to codon usage, a model does not currently exist to assess tRNA supply and demand and its effect in a maximum likelihood analysis. 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. Also, 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 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. We propose that more research into codon usage bias and its phylogenetic conservation will be beneficial to future phylogenetic studies by providing researchers with more robust phylogenetic trees.

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, C. H., Xie, D., Suchard, M. A., Rambaut, A., Drummond, A. J. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10, e1003537.

Brule, C. E., Grayhack, E. J. 2017. Synonymous Codons: Choose Wisely for Expression. Trends Genet 33, 283-297.

Butt, A. M., Nasrullah, I., Qamar, R., Tong, Y. 2016. Evolution of codon usage in Zika virus genomes is host and vector specific. Emerging Microbes & Infections; Infections 5, e107.

Camiolo, S., Melito, S., Porceddu, A. 2015. New insights into the interplay between codon bias determinants in plants. DNA Research, 10.1093/dnares/dsv027.

Cannarozzi, G., Schraudolph, N. N., Faty, M., von Rohr, P., Friberg, M. T., Roth, A. C., Gonnet, P., Gonnet, G., Barral, Y. 2010. A role for codon order in translation dynamics. Cell 141, 355-367.

Carbone, A., Kepes, F., Zinovyev, A. 2005. Codon bias signatures, organization of microorganisms in codon space, and lifestyle. Mol Biol Evol 22, 547-561.

Castle, J. C. 2011. SNPs occur in regions with less genomic sequence conservation. PLoS One 6, e20660.

Chan, C. X., Bernard, G., Poirion, O., Hogan, J. M., Ragan, M. A. 2014. Inferring phylogenies of evolving sequences without multiple sequence alignment. Sci Rep 4, 6504.

Chantawannakul, P., Cutler, R. W. 2008. Convergent host-parasite codon usage between honeybee and bee associated viral genomes. J Invertebr Pathol 98, 206-210.

Christianson, M. L. 2005. Codon usage patterns distort phylogenies from or of DNA sequences. American Journal of Botany. 92, 1221-1233.

Crick, F. H. 1966. Codon--anticodon pairing: the wobble hypothesis. J Mol Biol 19, 548-555.

Crick, F. H., Barnett, L., Brenner, S., Watts-Tobin, R. J. 1961. General nature of the genetic code for proteins. Nature 192, 1227-1232.

Cristina, J., Fajardo, A., Sonora, M., Moratorio, G., Musto, H. 2016. A detailed comparative analysis of codon usage bias in Zika virus. Virus Res 223, 147-152.

Dagan, T., Artzy-Randrup, Y., Martin, W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proceedings of the National Academy of Sciences 105, 10039-10044.

Das, S., Chakrabarti, J., Ghosh, Z., Sahoo, S., Mallick, B. 2005. A new measure to study phylogenetic relations in the brown algal order Ectocarpales: The "codon impact parameter". Journal of Biosciences 30, 699-709.

Daugelaite, J., O' Driscoll, A., Sleator, R. D. 2013. An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics. ISRN Biomathematics 2013, 14.

Dixon, G. B., Bay, L. K., Matz, M. V. 2016. Evolutionary Consequences of DNA Methylation in a Basal Metazoan. Molecular Biology and Evolution, 33, 2285-2293.

dos Reis, M., Savva, R., Wernisch, L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32, 5036-5044.

dos Reis, M., Wernisch, L., Savva, R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. Nucleic Acids Res 31, 6976-6985.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792-1797.

Farris, J. S. 2008. Parsimony and explanatory power. Cladistics 24, 825-847.

Felsenstein, J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Systematic Biology 27, 401-410.

Frenkel-Morgenstern, M., Danon, T., Christian, T., Igarashi, T., Cohen, L., Hou, Y.-M., Jensen, L. J. 2012. Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. Molecular Systems Biology 8, 572-572.

Friedman, R., Ely, B. 2012. Codon usage methods for horizontal gene transfer detection generate an abundance of false positive and false negative results. Curr Microbiol 65, 639-642.

Goloboff, P. A., Farris, J. S., Nixon, K. C. 2005. TNT: Tree Analysis Using New Technology. Cladistics, 54, 176-178.

Goodman, D. B., Church, G. M., Kosuri, S. 2013. Causes and effects of N-terminal codon bias in bacterial genes. Science 342, 475-479.

Gutman, G. A., Hatfield, G. W. 1989. Nonrandom utilization of codon pairs in Escherichia coli. Proc Natl Acad Sci U S A 86, 3699-3703.

Haszprunar, G. 1992. The types of homology and their significance for evolutionary biology and phylogenetics. Journal of Evolutionary Biology 5, 13-24.

Hillis, D. M., Bull, J. J. 1993. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. Systematic Biology 42, 182-192.

Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K. A., Deng, J., Drew, B. T., Gazis, R., Gude, K., Hibbett, D. S., Katz, L. A., Laughinghouse, H. D. t., McTavish, E. J., Midford, P. E., Owen, C. L., Ree, R. H., Rees, J. A., Soltis, D. E., Williams, T., Cranston, K. A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci U S A 112, 12764-12769.

Holder, M., Lewis, P. O. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4, 275-284.

Huelsenbeck, J. P., Larget, B., Miller, R. E., Ronquist, F. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst Biol 51, 673-688.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2, 13-34.

Inagaki, Y., Roger, A. J. 2006. Phylogenetic estimation under codon models can be biased by codon usage heterogeneity. Mol Phylogenet Evol 40, 428-434.

Ingvarsson, P. K. 2008. Molecular evolution of synonymous codon usage in Populus. BMC Evol Biol 8, 307.

Ivancevic, A. M., Kortschak, R. D., Bertozzi, T., Adelson, D. L. 2018. Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. Genome Biol 19, 85.

Katoh, K., Standley, D. M. 2014. MAFFT: iterative refinement and additional methods. Methods Mol Biol 1079, 131-146.

Kloster, M., Tang, C. 2008. SCUMBLE: a method for systematic and accurate detection of codon usage bias by maximum likelihood estimation. Nucleic Acids Res 36, 3819-3827.

Kober, K. M., Pogson, G. H. 2013. Genome-Wide Patterns of Codon Bias Are Shaped by Natural Selection in the Purple Sea Urchin, Strongylocentrotus purpuratus. G3: Genes|Genomes|Genetics 3, 1069.

Koonin, E. V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39, 309-338.

Koonin, E. V., Wolf, Y. I. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Research 36, 6688-6719.

Koski, L. B., Morton, R. A., Golding, G. B. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol 18, 404-412.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35, 1547-1549.

Lal, D., Verma, M., Behura, S. K., Lal, R. 2016. Codon usage bias in phylum Actinobacteria : relevance to environmental adaptation and host pathogenicity. Research in Microbiologoy 167, 669-677.

Li, J., Zhou, J., Wu, Y., Yang, S., Tian, D. 2015. GC-Content of Synonymous Codons Profoundly Influences Amino Acid Usage. G3 (Bethesda) 5, 2027-2036.

Ling, C., Hamada, T., Gao, J., Zhao, G., Sun, D., Shi, W. 2016. MrBayes tgMC3++: A High Performance and Resource-Efficient GPU-Oriented Phylogenetic Analysis Method. IEEE/ACM Trans Comput Biol Bioinform 13, 845-854.

Magis, C., Taly, J. F., Bussotti, G., Chang, J. M., Di Tommaso, P., Erb, I., Espinosa-Carrasco, J., Notredame, C. 2014. T-Coffee: Tree-based consistency objective function for alignment evaluation. Methods Mol Biol 1079, 117-129.

Miller, J. B., Hippen, A. A., Belyeu, J. R., Whiting, M. F., Ridge, P. G. 2017a. Missing something? Codon aversion as a new character system in phylogenetics. Cladistics, 33,545-556.

Miller, J. B., Hippen, A. A., Wright, S. M., Morris, C., Ridge, P. G. 2017b. Human viruses have codon usage biases that match highly expressed proteins in the tissues they infect. Biomedical Genetics and Genomics 2(2):1-5.

Miyazawa, S. 2013. Superiority of a mechanistic codon substitution model even for protein sequences in Phylogenetic analysis. BMC Evolutionary Biology, 13: 257.

Mortazavi, M., Zarenezhad, M., Alavian, S. M., Gholamzadeh, S., Malekpour, A., Ghorbani, M., Torkzadeh Mahani, M., Lotfi, S., Fakhrzad, A. 2016. Bioinformatic Analysis of Codon Usage and Phylogenetic Relationships in Different Genotypes of the Hepatitis C Virus. Hepatitis Monthly 16, e39196.

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., Minh, B. Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32, 268-274.

Nielsen, R., Bauer DuMont, V. L., Hubisz, M. J., Aquadro, C. F. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. Mol Biol Evol 24, 228-235.

Pais, F. S., Ruy Pde, C., Oliveira, G., Coimbra, R. S. 2014. Assessing the efficiency of multiple sequence alignment programs. Algorithms Mol Biol 9, 4.

Paninski, L., Pillow, J. W., Simoncelli, E. P. 2004. Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model. Neural Comput 16, 2533-2561.

Pechmann, S., Frydman, J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol 20, 237-243.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., Baurain, D. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLOS Biology 9, e1000602.

Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H., Dennis, P. P. 1979. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. Proc Natl Acad Sci U S A 76, 1697-1701.

Prat, Y., Fromer, M., Linial, N., Linial, M. 2009. Codon usage is associated with the evolutionary age of genes in metazoan genomes. BMC Evolutionary Biology 9, 285.

Quax, T. E., Claassens, N. J., Soll, D., van der Oost, J. 2015. Codon Bias as a Means to Fine-Tune Gene Expression. Mol Cell 59, 149-161.

Retief, J. D. 2000. Phylogenetic analysis using PHYLIP. Methods Mol Biol 132, 243-258.

Rogers, J. S. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. Syst Biol 46, 354-357.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M. A., Huelsenbeck, J. P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61, 539-542.

Roy-Zokan, E. M., Dyer, K. A., Meagher, R. B. 2015. Phylogenetic Patterns of Codon Evolution in the Actin-depolymerizing factor/cofilin (ADF/CFL) Gene Family. PLoS One 10, e0145917.

Roychoudhury, S., Mukherjee, D. 2010. A detailed comparative analysis on the overall codon usage pattern in herpesviruses. Virus Res 148, 31-43.

Sanderson, M. J. 1995. Objections to Bootstrapping Phylogenies: A Critique. Systematic Biology 44, 299-320.

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., Plotkin, J. B. 2013. Rate-limiting steps in yeast protein translation. Cell 153, 1589-1601.

Shao, Z. Q., Zhang, Y. M., Feng, X. Y., Wang, B., Chen, J. Q. 2012. Synonymous codon ordering: a subtle but prevalent strategy of bacteria to improve translational efficiency. PLoS One 7, e33547.

Sharp, P. M., Li, W. H. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24, 28-38.

Siddall, M. E. 1998. Success of Parsimony in the Four-Taxon Case: Long-Branch Repulsion by Likelihood in the Farris Zone. Cladistics 14, 209-220.

Sievers, F., Higgins, D. G. 2014. Clustal omega. Curr Protoc Bioinformatics 48, 3 13 11-16.

Sievers, F., Higgins, D. G. 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci 27, 135-145.

Soltis, D. E., Soltis, P. S. 2003. The Role of Phylogenetics in Comparative Genetics. Plant Physiology 132, 1790-1800.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313.

Sun, Y., Tamarit, D., Andersson, S. G. E. 2017. Switches in Genomic GC Content Drive Shifts of Optimal Codons under Sustained Selection on Synonymous Sites. Genome Biol Evol 9, 2560-2579.

Tekaia, F. 2016. Inferring Orthologs: Open Questions and Perspectives. Genomics Insights 9, 17-28.

Trotta, E. 2013. Selection on codon bias in yeast: a transcriptional hypothesis. Nucleic Acids Res 41, 9382-9395.

Tuller, T. 2011. Codon bias, tRNA pools and horizontal gene transfer. Mob Genet Elements 1, 75-77.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., Pilpel, Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141, 344-354.

Wald, A. 1949. Note on the Consistency of the Maximum Likelihood Estimate. 595-601.

Wilgenbusch, J. C., Swofford, D. 2003. Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics Chapter 6, Unit 6 4.

Wright, F. 1990. The 'effective number of codons' used in a gene. Gene 87, 23-29.

Xu, W., Xing, T., Zhao, M., Yin, X., Xia, G., Wang, M. 2015. Synonymous codon usage bias in plant mitochondrial genes is associated with intron number and mirrors species evolution. PLoS One 10, e0131508.

Xu, Y., Ma, P., Shah, P., Rokas, A., Liu, Y., Johnson, C. H. 2013. Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. Nature 495, 116-120.

Yang, Z., Rannala, B. 2012. Molecular phylogenetics: principles and practice. Nat Rev Genet 13, 303-314.

Zhang, Y. M., Shao, Z. Q., Yang, L. T., Sun, X. Q., Mao, Y. F., Chen, J. Q., Wang, B. 2013. Non-random arrangement of synonymous codons in archaea coding sequences. Genomics 101, 362-367.

Zur, H., Tuller, T. 2016. Predictive biophysical modeling and understanding of the dynamics of mRNA translation and its evolution. Nucleic Acids Research 44, 9031-9049.

Tables and Figures

Chapter 1 Tables

Table 1.1. Causes of 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.
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.

Manuscript published in *Cladistics*. 2017; https://doi.org/10.1111/cla.12183

Chapter 2

Missing Something? Codon aversion as a new character system in phylogenetics

Justin B. Miller[1], Ariel A. Hippen[1], Jonathon R. Belyeu[1], Michael F. Whiting[1,2] and 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*

Abstract

Although many studies have documented codon usage bias in different species, the importance of codon usage in a phylogenetic framework remains largely unknown. We demonstrate that a phylogenetic signal is present in the codon usage and non-usage biases of 17 717 orthologues evaluated across 72 tetrapod species using a simple parsimony analysis of a binary matrix of codon characters. Phylogenies estimated using stop codons were more congruent with previous hypotheses than phylogenies based on any other single codon or a combination of codons. Although each codon is present in every species, specific genes have different codon preferences and may or may not use every possible codon. This observation allowed us to map the pattern of codon usage and non-usage across the topology. These results suggest that codon usage is phylogenetically conserved across shallow and deep levels within tetrapods.

Introduction

Although 64 *different codons* exist, only 20 amino acids and a stop codon are encoded by these codons (Crick, 1966). Synonymous codons encode the same amino acid (Crick, 1966); however, their usage is typically not random. Codon usage bias refers to the nonrandom codon preference observed in most species (Ikemura, 1985; Sharp and Li, 1986; Gutman and Hatfield, 1989; Zhang et al., 2013). In addition to codon preferences, DNA triplet preferences also are evolutionarily conserved in both intronic and exonic regions of plants, *Escherichia coli* and *Drosophila* (Akashi et al., 2007; Yang, 2007; Xu et al., 2015).

Two nonmutually exclusive hypotheses attempt to explain the presence of codon usage bias: (i) nonrandom mutations occur particularly at the third codon position, and (ii) selection for codon bias persists (Hershberg and Petrov, 2008; Quax et al., 2015). An unequal expression of optimal (directly complementing all three nucleotides) transfer RNA (tRNA) anticodons among tissues and species, as well as an incomplete set of tRNAs in each species, leads to evolutionary pressure for using certain codons, potentially explaining both hypotheses (Quax et al., 2015). Suboptimal codons, in this instance, are defined as codons that bind to one or two tRNA anticodon nucleotides, but do not form a traditional hydrogen bond with the other base(s). The normalized translation efficiency (nTE) metric was introduced to account for different tRNA-codon binding efficiencies by incorporating both the supply and demand rates of tRNA with the suboptimal codons vying for each tRNA (Pechmann and Frydman, 2013). Other normalization rates, such as the effective number of codons, also have been introduced and used to account for codon variance (Wright, 1990). Interestingly, two competing studies report translational speed as being either slower or faster for suboptimal codons, rendering the effects of codon usage bias on

translational efficiency unresolved (Quax et al., 2015). Xu et al. (2013) suggests that suboptimal codons might be preferential to some species for increased translational efficiency.

In this research, we sought to determine if complete codon aversion (i.e. if a species does not use a codon within a gene) is conserved in some lineages. Furthermore, we assessed the use of codon non-usage bias as a phylogenetic character, and compared the phylogenetic signal present for each unused codon. As a test case, we analyzed 17 717 orthologues across 72 tetrapod species, and compared our phylogenies to the Tree of Life project (Maddison et al., 2007; Hinchliff et al., 2015). Our results suggest that codon non-usage bias is an informative phylogenetic character. Surprisingly, stop codon non-usage displayed the most reliable phylogenetic signal of all codons.

Methodology

Data collection and processing

We extracted all coding sequences (CDS) from the annotated proteins in 72 tetrapods and an outgroup representing *Clupeocephata* found in the National Center for Biotechnology Information (NCBI) database using gene annotations found in General Feature Format 3 (GFF3) files (Ostell and McEntyre, 2007; Pruitt et al., 2014; Tatusova et al., 2014; NCBI Resource Coordinators, 2016). We downloaded all reference sequence data, including gene annotations, from NCBI in September 2014. A reference genome is the average assembly of many individuals in a species, and is continually updated to represent the most common nucleotides found in a given species (Ostell and McEntyre, 2007). Because we are looking at the evolution of species from which many individuals have been sequenced, the reference genome will most accurately represent an “average individual” in a given species. We report the species taxonomy for all

species used in this study in Table S1 (Ostell and McEntyre, 2007; Pruitt et al., 2014; Tatusova et al., 2014; NCBI Resource Coordinators, 2016). Similar to Camiolo et al. (2015), when multiple isoforms were annotated we used the longest isoform as representative for that gene. Next, we removed any protein with an annotated exception (translational, unclassified transcription discrepancy, suspected errors, etc.). These filters do not appear to change the overall coding sequence data because they eliminated < 5% of the sequences. Based on standards established by the Human Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC), in which they attempt to maintain the same GFF3 gene names between species when orthologues exist (Gray et al., 2015), an uppercase and lowercase insensitive review of the gene names yielded 872 274 unique genes across all species. However, upon closer inspection, the majority of those unique genes were CDS locations that were identified with an "LOC_" tag, and were present only in one species. Because a phylogenetically informative character requires that species be separable by that character, a filter was placed on all genes requiring that the genes be present in at least three ingroup species, limiting the number of orthologues to 17 717 and the total number of sequences to 473 685.

Codon usage matrix calculation

We created a binary-encoded matrix of 64 characters per gene—one for each codon. If a species used a codon for a given orthologue, it was coded as "1" in the matrix; conversely, if a species did not use a specific codon in an orthologue, then it was coded as "0" in the matrix. This process was repeated for all species across all orthologues. This process is depicted in Figure 1.

Matrices were created for each gene and combined to form a super-matrix of all codons across all orthologues for each of 72 species. We only included each gene once in the character matrix, regardless of how many copies were present in an average individual. Parsimony-uninformative characters were removed from the data set. We removed parsimony-uninformative characters by first identifying and removing genes that were not present in at least three ingroup species. Next, we removed characters for which a "0" and a "1" were not both present across the species that had sequence data for the gene. We then eliminated species that did not have annotated data for at least 5% of the informative characters and repeated the process of removing parsimony-uninformative characters and eliminating species with < 5% of informative characters until no changes were made to the data set. Seventy-two species and 473 685 characters across 17 717 orthologues passed all filters. Species that lacked an annotated orthologue were coded as missing for that orthologue (See Table S1 for a list of missing data per species). Of the 72 species, there were 13 birds, 11 even-toed ungulates, 8 rodents, 6 bony fishes, 6 primates, 5 bats, 5 carnivores, 5 other placentals, 3 turtles, 2 lizards, 2 odd-toed ungulates, 2 other vertebrates, 1 insectivore, 1 monotreme, 1 marsupial and 1 rabbit (see Table S1 for a complete taxonomy). In total, there were 24 226 112 instances of 0/1 codon usage that were coded in the matrix for this analysis.

Phylogeny estimation

All trees were estimated using Tree Analysis Using New Technology (TNT) (Goloboff et al., 2008). We allowed up to 5000 trees to be held using the mult and bbreak=tbr commands. The resample command produced bootstrap support values. Although Bremer supports were also calculated for each tree, due to the large number of characters used, all Bremer support values

were very high, and are not included in the figures. The analysis was done as a pure parsimony analysis that generally analyzed tens of millions of potential trees per run.

Results

We constructed a phylogeny based on codon nonusage of all 64 codons using TNT (Goloboff et al., 2008). The maximum parsimony analysis produced a single most-parsimonious tree (Figure 2) with high average bootstrap support. Based on this phylogeny there appeared to be a strong phylogenetic signal in nonusage of all 64 codons together. Thirty-eight clades were recovered when compared with the Open Tree of Life project. Some of the main clades that were recovered include: Euarchontoglires, Laurasiatheria, Boreoeutheria, Muroidea, Passeriformes and Archelosauria. However, Rodentia was not correctly recovered, being polyphyletic in two distinct clades. To determine whether or not we recovered clades just by chance or if there is actually a phylogenetic signal, we generated ten random phylogenies through the web interface developed by Alix et al. (2012), and we recovered an average of only 1.3 clades per random tree.

Because the original tree recovered from all codons (Figure 2) appeared to have a phylogenetic signal, we wanted to determine if the phylogenetic signal is stronger in some codons compared to others. We partitioned the character matrix into 64 character matrices (one for each codon) and built a phylogeny for each of the 64 matrices (Figure S1). The number of clades in each of the 64 trees correctly corresponding to the Open Tree of Life project are shown in Table 1. The average number of clades recovered using a single codon was 31.36. However, the stop codons TAG, TAA and TGA recovered the most clades when compared with the Open Tree of Life at 41, 40, and 39, respectively. We then combined all codons that encoded each of the 20 amino acids plus

the stop codon, and we determined that the number of clades recovered by the stop codons is considered a true outlier when compared with all other amino acids. We also divided the codons into groups based on polarity and charge, and the stop codons again reported the highest phylogenetic signal and were true outliers compared with the other groups (Table 1). Because the stop codons displayed a significantly higher phylogenetic signal compared with the other codons, we built a phylogeny encoding all three possible stop codons as a single multistate character (0 = TAA, 1 = TGA, 2 = TAG) and the most-parsimonious tree (Figure 3) also had high average bootstrap support. To determine if the phylogenetic signal would be lost when excluding the stop codons, we removed all stop codon characters from the original matrix and estimated a new phylogeny. The resulting tree was identical to that in Figure 2.

We observed that a higher number of genes was used to reconstruct phylogenies based on the stop codons than for other codons (Table 1). To determine if the phylogenetic signal observed for stop codons was due simply to differences in the number of characters available for phylogeny reconstruction, we graphed the number of genes used in phylogeny reconstruction versus phylogenetic signal (i.e. number of clades recovered). Trend lines (linear and exponential) suggest that the signal measured by stop codons is in line with the predicted signal based on the number of genes used. However, we note that TTC (green data point in Figure 4) has the highest per character signal and possibly has a lower phylogenetic signal compared to the stop codons because of low usage (2382 total genes compared to > 15 000 for each of the stop codons).

Next, we sought to determine if stop codons maintained a strong phylogenetic signal with fewer data. We randomly sampled 2382 genes and reconstructed phylogenies 10 times based on stop

codons alone (red points in Figure 5). Even with the reduced sample sizes, the stop codons still demonstrated a stronger phylogenetic signal compared to other codons, recovering 37– 42 clades, with an average of 39 clades recovered.

Discussion

The recovered phylogenies in this work generally were congruent with the currently accepted phylogeny from the Open Tree of Life project. Each of the clades congruent with the Open Tree of Life is labelled on Figs 2 and 3. Interestingly, we recovered a bird phylogeny that was more congruent with the phylogeny proposed by Jetz et al. (2012) than the Open Tree of Life. There is also a debate regarding the correct placement of *Tupaia chinensis* (tree shrews) as a primate or a rodent. The phylogeny we recovered using just the stop codons (Figure 3) supports the assertion made by Xu et al. (2012) and Wu et al. (2013) that tree shrews could reasonably be considered a primate based on their phylogenetic history, albeit with low nodal support. However, when we used all codon nonusages to construct the phylogeny (Figure 2), *Tupaia chinensis* is depicted as a sister taxon to *Oryctolagus cuniculus* (European rabbit) and *Erinaceus europaeus* (European hedgehog) with high nodal support. This placement of the tree shrew most closely mimics the phylogenetic placement proposed by Murphy et al. (2001), as well as the maximum likelihood trees of third codon positions and all codons proposed by Lin et al. (2014). Although either tree could be correct, we propose that because the phylogeny recovered from the stop codons in Figure 3 is typically more congruent with the Open Tree of Life, successfully recovering 42 clades compared with 38 clades in Figure 2, the phylogeny recovered from stop codons should be favored over the phylogeny recovered from all codons. Favoring the phylogeny in Figure 3

lends support to the hypothesis that the tree shrew is more closely related to primates than rodents.

We found that a codon's usage within a given orthologue is sometimes constrained to a few clades with limited use across other clades. We depicted this phenomenon in Figure 6 using three codons as examples of a wider trend, and offer this explanation for why codon non-usage is phylogenetically informative when given a sufficient number of annotated orthologues. For example, in Figure 6 the codon ACG in the PPARG orthologue is present only within Rodentia, CCC in the PLN orthologue is present primarily in Sauria, and GCT in PLN is present primarily in Laurasiatheria.

Perhaps the most interesting aspect of this research is that a stronger phylogenetic signal was discovered in the stop codons than the other codons. We show that the phylogenetic signal present in codon usage is robust and is not affected by differences in the number of characters used to recover the phylogeny. Although the trend lines in Figure 4 appear to show a potential bias in recovering clades based solely on the number of genes used in the character matrix, when we extracted random samples of characters to match the number of genes used in TTC, which appeared to have the highest phylogenetic signal in Figure 4, the stop codons still recovered an average of three more clades than TTC. Furthermore, by limiting the number of stop codon characters from an average of 16 259 to 2382, the phylogenetic signal only decreased by recovering one less clade (Figure 5).

Additional research is needed to determine why the phylogenetic signal exists, but there are several possible explanations. Unlike other codons, stop codons are recognized directly by ribosomal protein release factors (Trotta, 2013), implying that unequally distributed tRNA anticodons might not be the only reason for codon usage bias. Castle (2011) also suggests that stop codons are more highly conserved than other codons. Another potential explanation is that compared to other codons, stop codons are guaranteed to have only one instance of any of the three stop codons. Although other codons are sometimes present in a binary manner (either having only one instance or zero instances), the number of instances across all 73 taxa where this occurs is quite low, ranging from 2 to 805 with an average of 196. Because 196 characters are insufficient to accurately represent the phylogeny of 73 species, we were not able to test directly if the higher phylogenetic signal was based on instances where the codon is either present with one instance or absent. However, because each codon by itself displayed a high phylogenetic signal, and the majority of codon instances had at least one species with more than one instance of the codon, it is reasonable to expect that the main factor in the phylogenetic signal is complete codon aversion as opposed to single codon instances.

We recognize limited taxon sampling due to the difficulty and expense of annotating species, and it may be several years before a sufficient number of genes are annotated to make this phylogenetic tree reconstruction method viable for more diverse taxa. We predict that as taxon sampling increases, our tree will become more robust because this method recovers both shallow and deep phylogenies of the species used. Future research should focus on the use and nonuse of codons within orthologues to determine which orthologues contain the highest phylogenetic signal, and in what ways codon usage might be integrated with other phylogenetic tree

reconstruction methods. Because both shallow and deep phylogenies were recovered through these simple parsimony analyses, we believe that continued research will provide us with realistic weights that might be added to a codon's use when incorporated in a Bayesian or maximum-likelihood setting. Furthermore, the simplicity of our analysis has the potential to identify conserved codons in genes and contribute to our knowledge of where phylogenetic signals exist within orthologues.

Acknowledgements

We appreciate the contributions of Brigham Young University for sponsoring our research and providing a facility in which to work and for the Fulton Supercomputing Laboratory at Brigham Young University, without which these analyses would not have been possible. Duke Rogers, Brandon Pickett and Anton Suvorov provided expert suggestions. The free use of the software, TNT, was made available to us through the sponsorship of the Willi Hennig Society, and we appreciate its efforts in keeping science accessible to everyone.

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.

Alix, B., Boubacar, D.A., Vladimir, M., 2012. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Res. 40, W573–W579.

Camiolo, S., Melito, S., Porceddu, A., 2015. New insights into the interplay between codon bias determinants in plants. DNA Res. 22, 461–470.

Castle, J.C., 2011. SNPs occur in regions with less genomic sequence conservation. PLoS One 6, e20660.

Crick, F.H., 1966. Codon–anticodon pairing: the wobble hypothesis. J. Mol. Biol. 19, 548–555.

Goloboff, P.A., Farris, S., Nixon, K., 2008. TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786.

Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., Bruford, E.A., 2015. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 43, D1079–D1085.

Gutman, G.A., Hatfield, G.W., 1989. Nonrandom utilization of codon pairs in *Escherichia coli*. Proc. Natl Acad. Sci. USA 86, 3699–3703.

Hershberg, R., Petrov, D.A., 2008. Selection on codon bias. Annu. Rev. Genet. 42, 287–299.

Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., Crandall, K.A., Deng, J., Drew, B.T., Gazis, R., Gude, K., Hibbett, D.S., Katz, L.A., Laughinghouse IV, H.D., McTavish, E., Midford, J., Owen, P.E., Ree, C.L., Rees, R.H., Soltis, J.A., Douglas, E., 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc. Natl Acad. Sci. USA 112, 12764–12769.

Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 13–34.

Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K., Mooers, A.O., 2012. The global diversity of birds in space and time. Nature 491, 444–448.

Lin, J., Chen, G., Gu, L., Shen, Y., Zheng, M., Zheng, W., Hu, X., Zhang, X., Qiu, Y., Liu, X., Jiang, C. 2014. Phylogenetic affinity of tree shrews to Glires is attributed to fast evolution rate. Mol. Phylogenet. Evol. 71, 193–200.

Maddison, D.R., Schulz, K.S., Maddison, W.P., 2007. The tree of life web project. Zootaxa 40, 19–40.

Michonneau, F., Brown, J.W., Winter, D.J., 2016. rotl: an R package to interact with the Open Tree of Life data. Methods Ecol. Evol. 7, 1476–1481.

Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A., O'Brien, S.J., 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409, 614–618.

NCBI Resource Coordinators 2016. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 44, 7–19.

Ostell, J., McEntyre, J., 2007. The NCBI Handbook. NCBI Bookshelf 1–8. doi:10.4016/12837.01

Pechmann, S., Frydman, J., 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat. Struct. Mol. Biol. 20, 237–243.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., Murphy, M.R., O'Leary, N.A., Pujar, S., Rajput, B., Rangwala, S.H., Riddick, L.D., Shkeda, A., Sun, H., Tamez, P., Tully, R.E., Wallin, C., Webb, D., Weber, J., Wu, W, DiCuccio, M., Kitts, P., Maglott, D.R., Murphy,

T.D., Ostell, J.M., 2014. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 42, 756–763.

Quax, T.E.F., Claassens, N.J., So€ll, D., van der Oost, J., 2015. Codon bias as a means to fine-tune gene expression. Mol. Cell 59, 149–161.

Sharp, P.M., Li, W.H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28–38.

Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K., Tolstoy, I., 2014. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res. 42, 553–559.

Trotta, E., 2013. Selection on codon bias in yeast: a transcriptional hypothesis. Nucleic Acids Res. 41, 9382–9395.

Wright, F., 1990. The "effective number of codons" used in a gene. Gene 87, 23–29.

Wu, X., Chang, Q., Zhang, Y., Zou, X., Chen, L., Zhang, L., Lv, L., Liang, B., 2013. Relationships between body weight, fasting blood glucose concentration, sex and age in tree shrews (*Tupaia belangeri chinensis*). J. Anim. Physiol. Anim. Nutr. 97, 1179– 1188.

Xu, L., Chen, S.Y., Nie, W.H., Jiang, X.L., Yao, Y.G., 2012. Evaluating the phylogenetic position of Chinese tree shrew (*Tupaia belangeri chinensis*) based on complete mitochondrial genome: implication for using tree shrew as an alternative experimental animal to primates in biomedical research. J. Genet. Genomics 39, 131–137.

Xu, Y., Ma, P., Shah, P., Rokas, A., Liu, Y., Johnson, C.H., 2013. Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. Nature 495, 116–120.

Xu, W., Xing, T., Zhao, M., Yin, X., Xia, G., Wang, M., 2015. Synonymous codon usage bias in plant mitochondrial genes is associated with intron number and mirrors species evolution. PLoS One 10, e0131508.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.

Tables and Figures

Chapter 2 Tables

Table 2.1. Sixty-four Phylogenetic Strict Consensus Trees Recovered Using TNT were Created Using Just the Presence or Absence of a Single Codon

Codon/Amino Acid/Property	# Clades Recovered	# Total Genes Used	# Genes with exactly one instance of codon	# Genes with more than one instance of codon
AAA	31	4908	67	4841
AAC	29	2612	75	2537
AAG	24	1057	33	1024
AAT	31	8449	183	8266
ACA	33	7553	140	7413
ACC	33	3697	61	3636
ACG	29	14056	403	13653
ACT	32	8790	128	8662
AGA	33	8445	178	8267
AGC	26	3275	80	3195
AGG	30	6177	134	6043
AGT	31	8720	217	8503
ATA	33	12380	641	11739
ATC	26	2705	72	2633
ATG	18	291	2	289
ATT	32	8263	203	8060
CAA	32	9416	203	9213
CAC	32	4576	130	4446
CAG	27	1040	21	1019
CAT	35	10187	312	9875
CCA	35	6615	126	6489
CCC	29	5116	95	5021
CCG	33	14803	481	14322
CCT	33	6019	102	5917
CGA	32	12874	543	12331
CGC	31	10253	354	9899
CGG	34	9936	366	9570
CGT	37	14558	741	13817
CTA	32	13654	356	13298
CTC	32	3528	92	3436
CTG	24	1271	25	1246
CTT	34	9295	138	9157

GAA	31	4427	82	4345
GAC	30	1768	62	1706
GAG	21	993	25	968
GAT	32	5084	112	4972
GCA	34	6273	90	6183
GCC	23	2595	39	2556
GCG	28	15111	463	14648
GCT	26	4672	55	4617
GGA	30	5455	87	5368
GGC	30	3529	52	3477
GGG	29	5029	129	4900
GGT	33	8830	164	8666
GTA	31	13695	462	13233
GTC	34	4653	90	4563
GTG	29	1561	32	1529
GTT	34	9698	208	9490
TAA	40	16822	NA	16822
TAC	32	3707	154	3553
TAG	41	15017	NA	15017
TAT	33	9432	354	9078
TCA	34	9457	235	9222
TCC	34	4041	68	3973
TCG	31	15605	710	14895
TCT	35	7544	116	7428
TGA	39	16939	NA	16939
TGC	31	5761	177	5584
TGG	29	3229	219	3010
TGT	31	8738	303	8435
TTA	33	13502	805	12697
TTC	36	2382	69	2313
TTG	35	7699	123	7576
TTT	35	5918	159	5759
F	35.5			
L	31.66666667			
I	30.33333333			
M	18			
V	32			
S	31.83333333			
P	32.5			
T	31.75			
A	27.75			
Y	32.5			
H	33.5			
Q	29.5			

N	30			
K	27.5			
D	31			
E	26			
C	31			
W	29			
R	32.83333333			
G	30.5			
Nonpolar, aliphatic R groups	28.375			
Nonpolar, aromatic R groups	32.33333333			
Polar, Uncharged R groups	31.09722222			
Positively charged R groups	31.27777778			
Negatively charged R groups	28.5			
STOP	40			

Each of these trees was then scored based on the number of clades successfully recovered when compared with the Open Tree of Life project. The outgroup was considered a single clade and had to be correctly positioned with all fish included in the clade in order for it to be scored. All other clades were scored as the most specific clade name without containing any member that did not belong to that clade. We then averaged the number of clades successfully recovered across all codons in each of the 20 amino acids plus the stop codon, as well as amino acids based on specific properties. We found that the stop codons were outliers when compared with the other amino acids, recovering on average 40 of 43 clades.

Chapter 2 Figures

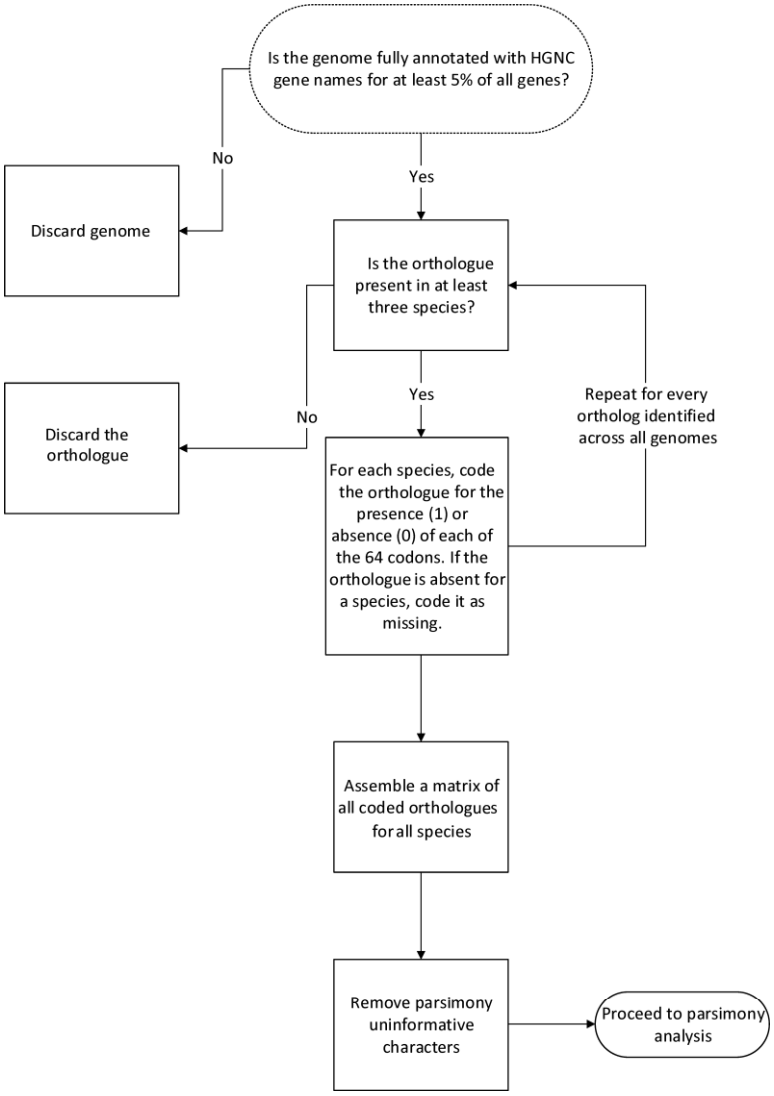


Figure 2.1. Flowchart Demonstrating How the Character Matrix was Coded We started with 45 234 orthologues in 72 tetrapods and outgroup fishes. First, we counted the number of times each codon was used in each gene, in each species. Species that did not have annotated data for a particular gene were denoted with an "X." Next, we coded a binary matrix, in which "1" means that the codon was present in an orthologue, and "0" means that it was not present. Species for which data were not available received a "?" in that field. Finally, we condensed the matrix to include only parsimony-informative characters, which had two distinct phenotypes—thus, both a "1" and a "0" were present for the character among the species with the given gene and at least three ingroup species had sequence data for the gene. All other characters were removed from the analysis. These steps were repeated for each gene in each species, and all the matrices were combined into a single super matrix. After all steps, there were 473 685 parsimony-informative characters across 17 717 orthologues.

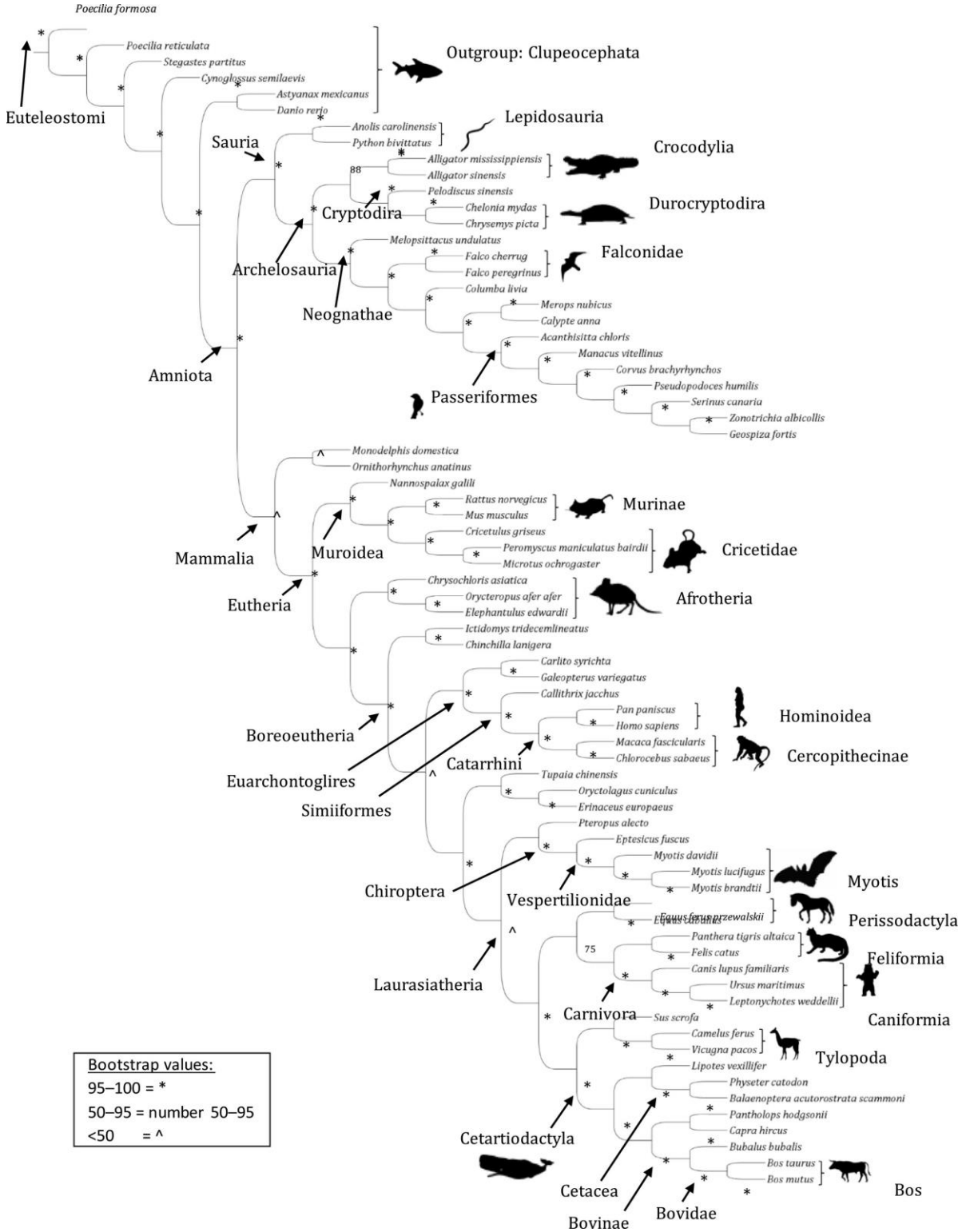


Figure 2.2. Most-parsimonious Tree Produced from TNT Using All 473 685 Codon Usage Characters The character matrix was created for each codon as outlined in Figure 1. Average bootstrap support for this tree is 94.1. Clades were labelled based on The Open Tree of Life project by ensuring that each labelled clade contains a majority of potential species belonging to that clade without including any species that do not belong to that clade.

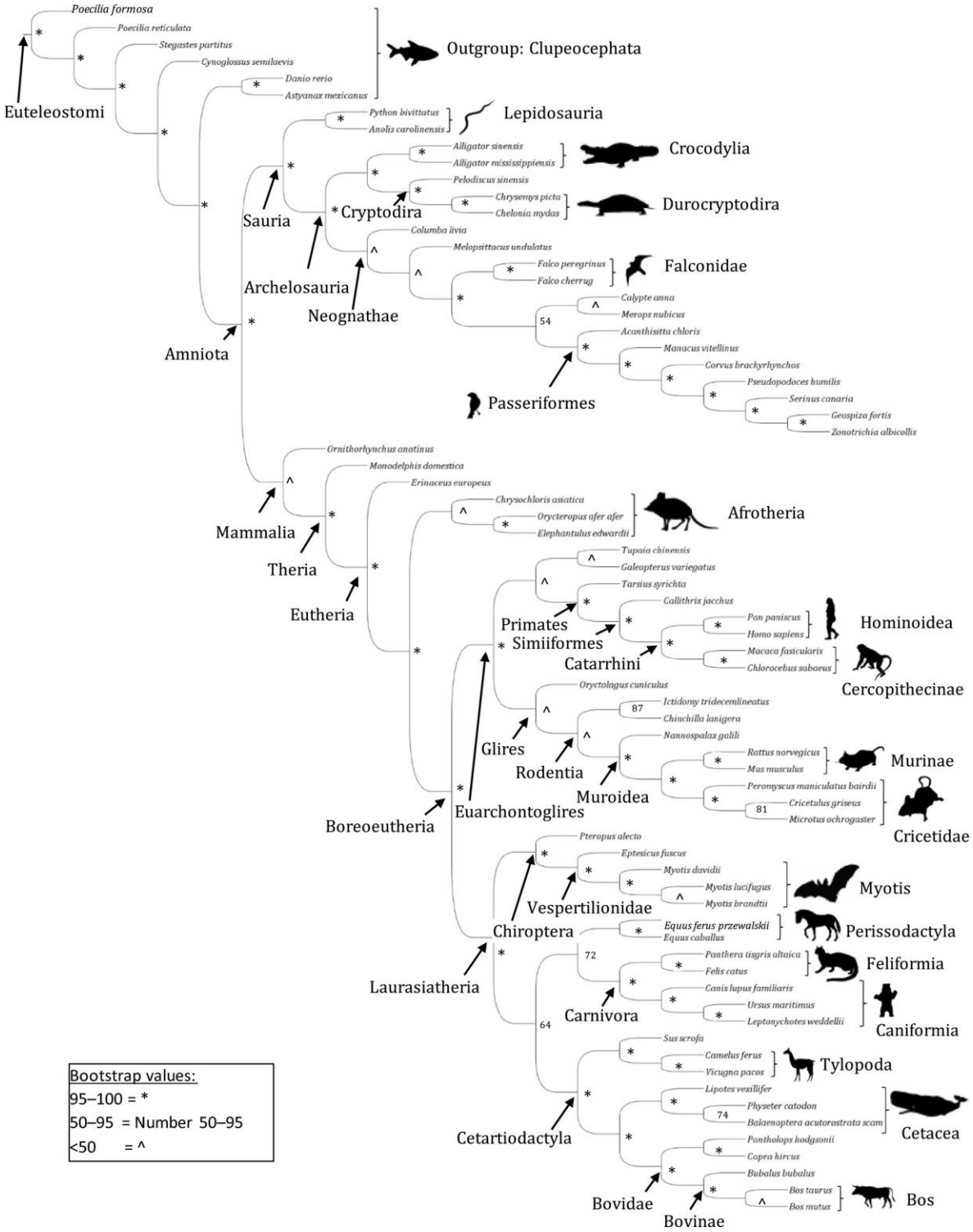


Figure 2.3. Most-parsimonious Tree Produced from TNT Using 48 778 Stop Codons The character matrix was created for each codon as outlined in Figure 1. Average bootstrap support for this tree is 85.6. Clades were labelled based on The Open Tree of Life project by ensuring that each labelled clade contains a majority of potential species belonging to that clade without including any species that do not belong to that clade.

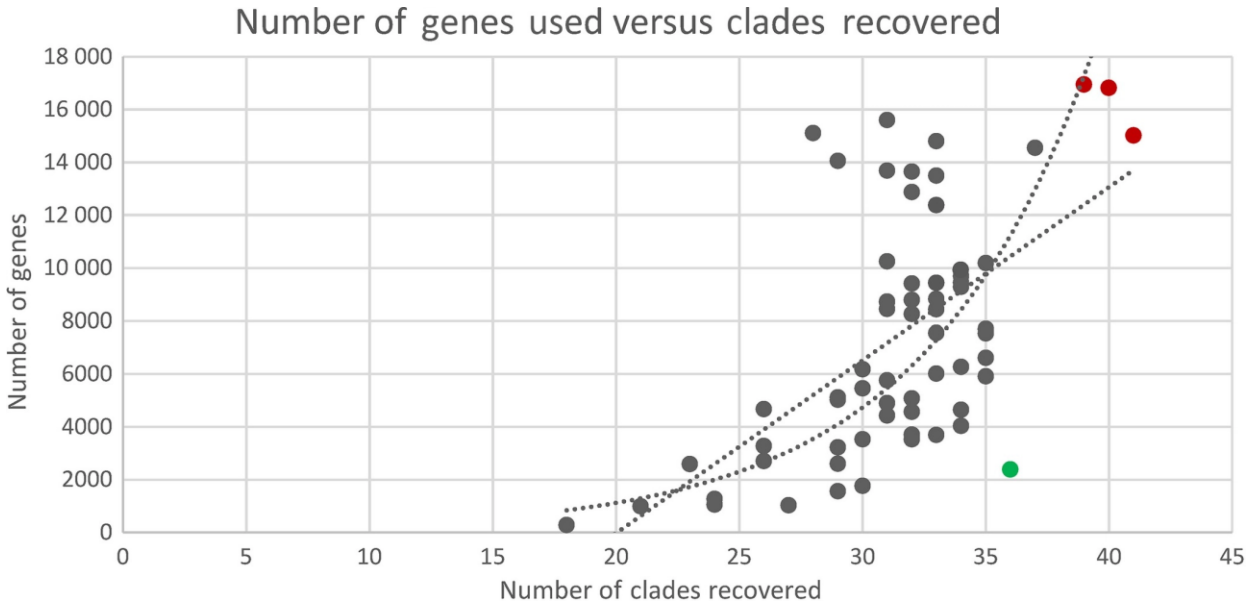


Figure 2.4. The Number of Genes with a Parsimony-informative Codon Plotted Against the Number of Clades Successfully Recovered when Compared with the Open Tree of Life
Each point represents a different codon. The three red points represent the three stop codons, and the green point represents Phenylalanine encoded by TTC. The linear and exponential trend lines show that although the stop codons display the highest phylogenetic signal, it is in line with the expectation based on the number of genes used to recover the phylogeny. Although TTC does not display as high a phylogenetic signal as the stop codons, the low usage (2382 instances) means that per character, TTC displays the highest phylogenetic signal.

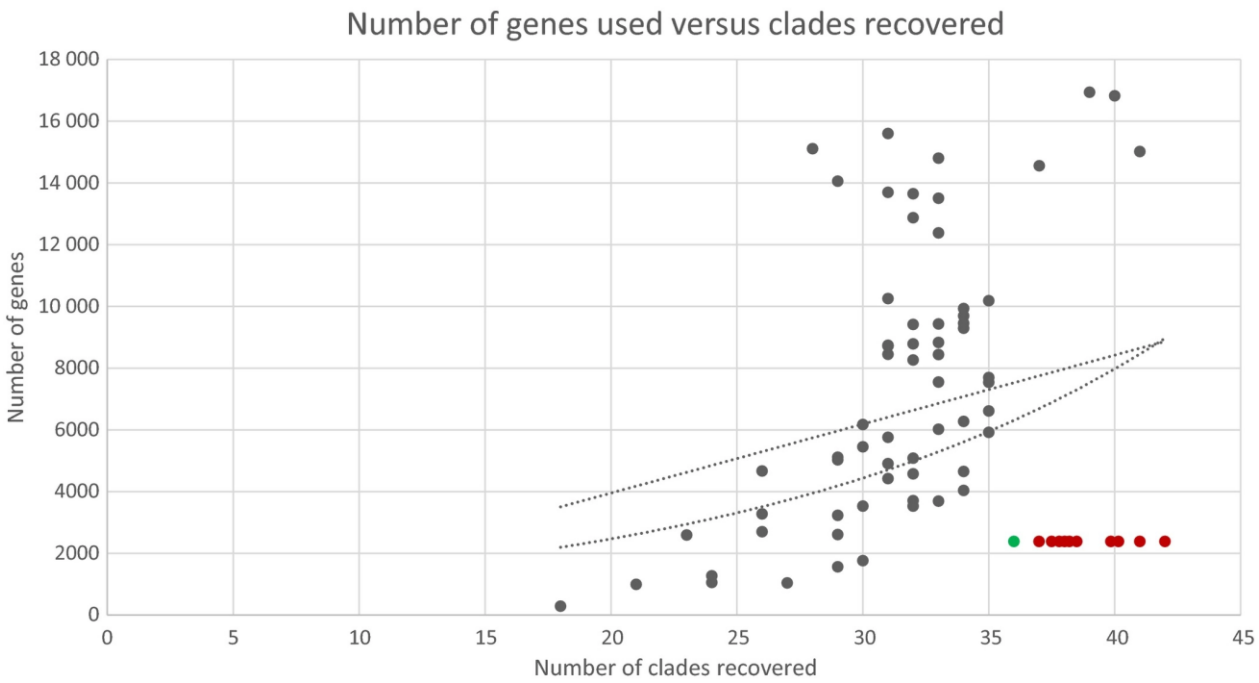


Figure 2.5. The Number of Genes with a Parsimony-informative Codon Plotted against the Number of Clades Successfully Recovered when Compared with the Open Tree of Life
Each point represents a different codon. The ten red points represent ten random samples of 2382 instances of the stop codons. The green point represents Phenylalanine encoded by TTC. The ten random phylogenies recovered an average of 39 clades with a range of 37–42. All random stop codon phylogenies recovered more clades than the phylogeny recovered by TTC using the same number of characters. We added jitter to the 10 red points to make them all visible.

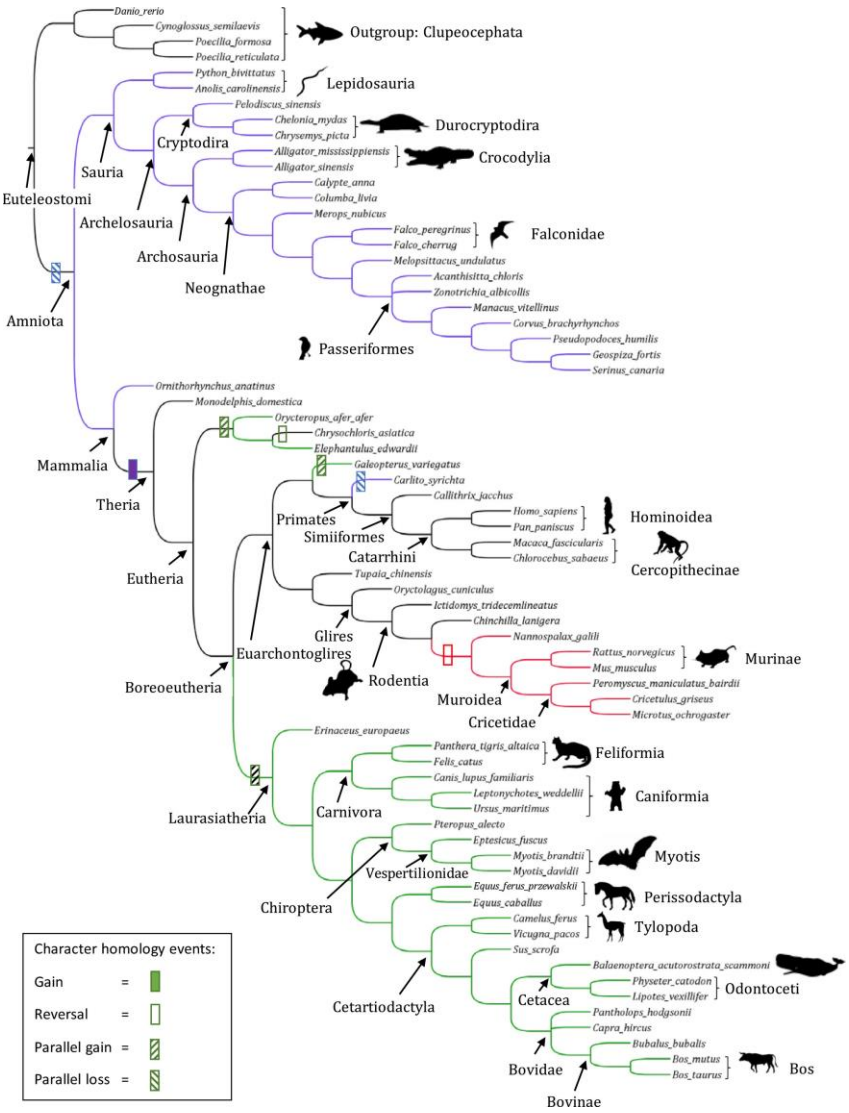


Figure 2.6. A Phylogeny Recovered from the Open Tree of Life Project (Hinchliff et al., 2015) This tree was constructed using an R package named ROTL (Michonneau et al., 2016). ROTL extracts the phylogeny from the Open Tree of Life, and then allows users to induce a subtree from the larger phylogeny. The subtree that we induced contained all species in our analysis, with the exception of *Astyanax mexicanus*, *Myotis lucifugus* and *Stegastes partitus*, because those species' phylogenies were not inducible by the software package. Three different characters are mapped on this tree. The first character (red) shows all species that did not have the codon ACG in the Peroxisome Proliferator-Activated Receptor Gamma (PPARG) gene. The second character (purple) shows all species that did not have the CCC codon in the Phosholamban (PLN) gene. The third character (green) shows all species that had at least one GCT codon in the Phosholamban (PLN) gene. These genes were chosen because > 95% of the species had annotated data for these genes, and the codons depict a conserved phylogenetic component in a species' use or nonuse of these particular codons.

Manuscript Submitted to *Molecular Phylogenetics and Evolution*

Chapter 3

Codon Use and Aversion is Largely Phylogenetically Conserved Across the Tree of Life

Justin B. Miller[1], Lauren M. McKinnon[1], Michael F. Whiting[1,2], and 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*

Abstract

Using parsimony, we analyzed codon usages across 12 337 species and 25 727 orthologous genes to rank specific genes and codons according to their phylogenetic signal. We examined each codon within each ortholog to determine the codon usage for each species. In total, 890 814 codons were parsimony informative. Next, we compared species that used a codon with species that did not use the codon. We assessed each codon's congruence with species relationships provided in the Open Tree of Life (OTL) and determined the statistical probability of observing these results by random chance. We determined that 25 771 codons had no parallelisms or reversals when mapped to the OTL. Codon usages from orthologous genes spanning many species were 1 109x more likely to be congruent with species relationships in the OTL than would be expected by random chance. Using the OTL as a reference, we show that codon usage is phylogenetically conserved within orthologous genes in archaea, bacteria, plants, mammals, and other vertebrates. We also show how to use our provided framework to test different tree hypotheses by confirming the placement of turtles as sister taxa to archosaurs.

Availability: All scripts, a README, and necessary test files are freely available on GitHub at https://github.com/ridgelab/codon_congruence

Key Words: codon aversion; tree of life; species classification; maximum likelihood; parsimony; phylogeny.

Contact: perry.ridge@byu.edu

Supplementary Information: All supplemental files are available at *Molecular Phylogenetics and Evolution* online

Introduction

The genetic code is degenerate because 64 canonical codons encode 20 amino acids and the stop codon, meaning multiple synonymous codons encode the same amino acid (Crick, 1970; Crick, 1966, 1968; Crick et al., 1961). Codon usage bias refers to the unequal distribution of synonymous codons between species, genes, or locations within the same gene, and can be used to regulate gene expression (Quax et al., 2015), suggesting that codon choice, even when synonymous, has biological implications. Typically, more closely related species share more similar patterns of codon usage and codon aversion (i.e., when a species does not use a codon within an ortholog), and these patterns are phylogenetically conserved (Miller et al., 2017). However, similar to other genetic characters (Rokas and Carroll, 2008), parallelism is present in the usage or aversion to many codons, resulting in homoplasy. In codon data, homoplasy may occur by parallelism, convergence, or reversal, resulting in identical character states that were not directly inherited from the most recent common ancestor. The presence of homoplasy is the greatest challenge in phylogenetic estimation, and nearly all characters, whether morphological or molecular, display homoplasy of some form at some level (Sanderson and Hufford, 1996).

To limit the effects of contradictory signals from homoplasy, the commonly-used maximum likelihood statistical method for estimating phylogenies approximates rates of evolution (e.g., transition and transversion ratios, evolutionary clock, evolutionary distance of species, etc.) and tree topography (Felsenstein, 1981). The basis of maximum likelihood is the proportionality of the likelihood function to the multinomial probability of observing the data given the tree and model (Huelsenbeck and Crandall, 1997; Yang et al., 1994). Maximum likelihood also uses the statistical property of consistency, which shows that as the number of data points approaches

infinity, the maximum likelihood estimators will converge on the same estimate (Wald, 1949). In contrast, parsimony does not use a model to recover phylogenies, which potentially limits the consistency of the method (Felsenstein, 1978); however, *ad hoc* hypotheses of homoplasy are also limited (Farris, 1983).

From a likelihood standpoint, a model of codon usage and codon aversion requires understanding how codon usages change throughout evolutionary time. Since no models of codon aversion evolution are currently available, we must first start with parsimony. From a parsimony perspective, a tree hypothesis is meant to minimize the number of similarities left unexplained (Farris, 2008). We aimed to determine the extent that codon usage and codon aversion within orthologous genes is congruent with species relationships as presented in the Open Tree of Life (OTL) (Hinchliff et al., 2015).

Encoding the codon matrix based on codon usage anywhere in the gene was first proposed by Miller et al. (2017), and categorizes homology on a genic scale, instead of positional homology from a multiple sequence alignment. Since this method characterizes codon usage with a binary representation, it essentially determines if a species "chooses" to use a given codon within a gene. Within genes, codon usage regulates gene expression in various ways. Using an equal number of codons to the supply of cognate tRNA anticodons maintains optimal codon usage that increases translational efficiency (Sharp and Li, 1987). Using multiple instances of the same codon (identical codon pairing) or synonymous codons (co-tRNA codon pairing) within a ribosomal window also increases translational efficiency and speed (Cannarozzi et al., 2010). Furthermore, mRNA structural folding and differential protein production are affected by codon

usage bias within a gene (Gingold et al., 2014; Pechmann and Frydman, 2013; Quax et al., 2015). Therefore, codon aversion within a gene, although potentially not homologous at a given position, is homologous on a genic level and can influence mRNA folding, protein production, and tRNA translational efficiency. We show that this method is phylogenetically conserved using 12 337 reference genomes across all domains of life.

Materials and Methods

Data Collection and Processing

All reference genomes were downloaded from the National Center for Biotechnology Information (NCBI) (Coordinators, 2013; Pruitt et al., 2014; Pruitt et al., 2000; Tatusova et al., 2014) in September, 2017 from ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/. A reference genome represents the consensus genome for a species based on the most complete genome assemblies (Pruitt et al., 2014). We extracted all coding sequence (CDS) data from the reference genomes, and we assigned each of the 12 337 species to the following groups: 362 archaea, 11 227 bacteria, 214 fungi, 147 invertebrates, 105 mammals, 120 other vertebrates, 87 plants, and 75 protozoa based on species annotations in NCBI. Since viruses are not included in the OTL, they were not included in our analysis. We recognize that several of these taxonomic groups do not represent monophyletic clades, but we opted to keep the groups outlined in the NCBI database to facilitate comparisons between studies that also use these annotations. We required that CDS regions be annotated with a gene name from the HUGO Gene Nomenclature Committee (HGNC) (Gray et al., 2015) to ensure that orthologous comparisons of codon usage were used. Although we do not perform any formal analysis to verify the orthologous relationships proposed by the HGNC, the HGNC standardizes various gene studies with gene annotations in SWISS-PROT (UniProt Consortium, 2018), and facilitates ortholog comparisons between species.

Next, we filtered the CDS regions to remove any annotated exceptions (e.g., translational exceptions, unclassified transcription discrepancies, suspected errors, etc.). We used the longest isoform of each gene when multiple isoforms were annotated in order to include all codons that are used in the gene. We also included partial gene sequences where the orthologous relationship was annotated in order to include as many orthologous gene comparisons as possible. Finally, we required each ortholog to be present in at least four species to ensure that the codon usage could be parsimony informative.

For each codon within each ortholog, we encoded its usage in a binary matrix (i.e., if the codon was used, it was given a "1" and if it was not used, it was given a "0"). After all codons within all orthologs for all species were included in the binary matrix, we filtered out parsimony uninformative characters (i.e., when all species with an ortholog either used or did not use a given codon). For each remaining codon, we divided the species sampled into two partitions based on their character state for that codon: species that use a codon within an ortholog and species that do not use that codon within the same ortholog. This process is depicted in Figure 1.

After encoding the binary matrix for each codon within each ortholog, we evaluated each bipartition against the OTL to determine if parallelisms or reversals occurred. Parallelisms occur when the same codon independently arises in different lineages not due to a common ancestor. Reversals occur when a codon reverts back to an ancestral state (e.g., if a species uses a codon that its most resent ancestor did not use, but the codon was used by a more distant ancestor). For each orthologous gene and codon state in each bipartition, we report the number of gains, losses, unknown gain/loss at the root node, number of species in the smaller partition, total species with

that ortholog, percent of species in the smaller partition, and total number of gain/loss divided by the number of species in the smaller partition (see Supplementary Tables 1-9). A codon was classified as separating species according to taxonomic groups reported in the OTL if the smaller group had at least two species and the total number of gain/loss and unknown gain/loss at the root node equaled one. This process was used because a singular gain/loss event would indicate that no reversals or parallelisms occurred for that codon character state and its state is unique to that lineage.

Statistical Validation

Because autapomorphies are not parsimony informative, we required that the smaller partition include at least two species before it was mapped to the OTL. We then determined where on the OTL species gain or lose the usage of each codon character. Initially mapping the codon usage from a single species to the OTL has a probability of 1.0 of mapping to a taxonomic group that is congruent with the OTL because autapomorphies provide no evidence of species relationships. If the remaining character-state distributions randomly separate the other species (i.e., the null hypothesis), then we can use conditional probabilities to calculate the probability that a monophyletic group of more than one species is obtained by random chance. In this case, the probability that another species from the same taxonomic group as the first species would be correctly added to the same taxonomic group as the first, or subsequent, species is given in equation (1).

Equation 1:

$$\frac{number\ of\ species\ in\ a\ taxonomic\ group\ -\ number\ of\ species\ already\ assigned\ to\ that\ taxonomic\ group}{total\ number\ of\ species\ not\ yet\ assigned\ to\ a\ taxonomic\ group}$$

Using conditional probability, we calculate the probability of correctly assigning all members of a species partition to the taxonomic groups outlined in the OTL by random chance, using Equation (2). In Equation (2) *s*=the number of species in the smaller taxonomic group, and *t*=the total number of species sampled. We start with *s*-1 and *t*-1 to account for the initial species that was mapped to the phylogeny, which will always have a probability of 1.0 that it is correctly placed in a monophyletic group.

Equation 2: $\left(\frac{s-1}{t-1}\right)\left(\frac{s-2}{t-2}\right)(\ldots)\left(\frac{1}{t-s+1}\right)$

This equation simplifies to equation (3).

Equation 3: $\frac{\prod_{i=1}^{s-1} i}{\prod_{j=t-s+1}^{t-1} j}$

A taxonomic distribution is defined as the number of species in the sets separated by the codon character state (i.e., the number of species that use a given codon and the number of species that do not use a given codon within an ortholog), without regard to the OTL. Using equation (3), we calculated the expected number of significant character states for each taxonomic distribution (e.g., if five species were sampled, with two species in the smaller group, then, using equation (3), the probability that they were correctly divided by random chance is 0.25). We then multiplied the probability of that taxonomic distribution correctly agreeing with the OTL by the number of instances of that taxonomic distribution in our dataset (e.g., if 12 instances of five species dividing into groups of two and three occurred in our dataset, then the expected number of species partitions agreeing with the OTL taxonomic groups would be 0.25 * 12 = 3). We performed a chi square analysis for the taxonomic distributions using these expected values and the observed values from our analysis.

Random Permutations

The statistical validation gives a theoretical probability of obtaining these results by random chance. However, it does not take into account the degree of homoplasy in the dataset or the overall congruence within the homoplastic character states with respect to other trees than the OTL. For these reasons, random permutations are needed to ensure that the probability of character states mapping to random trees does not exceed the probability of the observed character states mapping to the OTL.

Random permutations were conducted 100 times for each taxonomic group. Each permutation maintained the tree structure as hypothesized in the OTL to not bias our results based on artificial tree structures. We then randomized the distribution of the species in each taxonomic group, creating 100 different species relationships using the same tree structure. Next, we conducted the same statistical validation on each of these trees with randomly distributed species, as outlined above. We calculated the number of permutations where the *p-value* of the mapped characters was less than or equal to the observed *p-value*. Where random permutations obtained a smaller *p-value* than the observed, we concluded that there was not support for codon usage as being more congruent with the OTL than expected by random chance.

Visualizing Homology on the Tree of Life

We inferred the reference phylogenies from the OTL for each pre-defined taxonomic group using tools available in the OTL documentation. We then mapped each character state to the inferred subtree from the OTL and determined how many gains, losses, parallelisms, and reversals occurred. The entire process of mapping character states from the original coding

sequences to the phylogeny in Newick format is outlined in Figure 2. Visualizations of the phylogenies were created using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Dealing with Limitations in Ortholog Annotations

While the HGNC gene annotations often span hundreds of species, many genes are annotated in fewer than 10 species, with the smaller species partition (i.e., species either using or not using a codon in an ortholog) containing two or three species. These taxonomic distributions were also included in the main analysis. However, the statistical probability (outlined above) of changes in each codon's usage being congruent with the phylogeny outlined in the OTL for these groups allows for many false positives. For instance, if an ortholog is annotated in four species, with two species using a codon and two species not using that codon, the statistical probability of that codon usage being congruent with the OTL by random chance is 0.33333. Across all species, 9 990 codons fall under this taxonomic distribution, meaning 3 330 of these codons are expected to agree with the OTL by random chance. Although the observed congruence is much higher (4 915), we wanted to ensure that the signal was not simply due to missing ortholog annotations. So, we excluded taxonomic distributions where the probability of obtaining congruence with the OTL was less than one divided by the number of parsimony informative characters. By doing this analysis, we limited the maximum total number of expected congruent codons to one, while ensuring that all observed congruences were statistically unlikely to occur by random chance (i.e., not due to missing data).

Results

Statistical Test

We report the number of different taxonomic distributions, the number of codons with no parallelisms or reversals on the OTL, the *t*-statistic for each group of species used in our analysis, and the *p*-value in Table 1. All taxonomic distributions, expected values, and observed values for each group of species are found in Supplementary Tables 10-18. All 64 codons had similar proportions in the group of codons that mapped to a single gain/loss event on the OTL (*t*-statistic=0.17907, *p*-value=1.0). The ratio of each codon to the total number of codons with a single gain/loss event is depicted in Figure 3.

Permutations

Random permutations for each taxonomic group show that codon usages in archaea, bacteria, plants, mammals, and other vertebrates are not likely to be congruent with the OTL by random chance. The *t-statistics* for the codons with no parallelisms or reversals within these taxonomic groups proposed by the OTL were orders of magnitude larger than the highest *t-statistic* obtained by the random permutations. The largest difference occurred in other vertebrates, where the observed *t-statistic* was 1.50814×10^{25} and the highest *t-statistic* from random permutations was 1.16996×10^{3}.

Although most taxonomic groups had observed *t-statistics* for codon usage that were much greater than those *t-statistics* calculated from random permutations, fungi, invertebrates, and protozoa did not. The *t-statistic* obtained for protozoa was within one order of magnitude of the *t-statistic* of the most improbable random permutation. For fungi, 16% of random permutations

had *t-statistics* greater than or equal to observed *t-statistics* from mapping codons to the OTL. Permutations for invertebrates produced *t-statistics* greater than or equal to mapping codons to the OTL 3% of the time.

Missing Ortholog Annotations

Table 2 shows the number of codons within each taxonomic group that have ortholog annotations spanning many species and are unlikely to be completely congruent with the OTL by random chance. Using the statistical validation outlined above, we set a cutoff of one divided by the total number of parsimony informative codons. This ensured that if all codons had a probability less than or equal to the cutoff, at most one codon will be completely congruent with the OTL by random chance. However, as shown in column 4 of Table 2, the maximum number of codons expected to be congruent with the OTL assuming each codon had the maximum probability (i.e., column 3 divided by column 2) was always less than one. By dividing the observed number of codons that agree with the OTL (column 5) by the maximum expected number of codons agreeing with the OTL (column 4), we see a substantial difference in the observed versus the expected in most taxonomic groups (column 5). No orthologs spanned a sufficient number of species in fungi or protozoa to make this analysis possible for those taxonomic groups. Furthermore, few orthologs were annotated across a sufficient number of invertebrates to assess the quality of codon homology in that taxonomic group. The observed number of codons congruent with the OTL in orthologs spanning many species of archaea, bacteria, mammals, other vertebrates, all species, or plants were 37x, 42x, 641x, 985x, 1 109x, and 1 795x larger than the expected values, respectively (column 6).

Character States that are Completely Congruent with the OTL

We report the Newick formatted phylogeny from the OTL with all codons that have a singular gain/loss mapped to the trees for each set of species (Supplementary Files 1-9), with the respective character state files showing which codons were gained or lost (Supplementary Files 10-18). Visualizations of the codons that are completely congruent with fungi, invertebrates, plants, protozoa, mammals, and other vertebrates are shown in Supplementary Figures 1-6.

Very Unlikely Character State Distributions

Of the 890 814 codon states analyzed, 25 771 codons had no homoplasy when mapped to the OTL (see Table 1). We further explored a fraction of these character state changes by choosing a subset of codon state changes with a *p*-value $\leq 1 \times 10^{-25}$ of being congruent with the OTL by random chance. Using this arbitrary threshold of 1×10^{-25}, 52 854 codon characters had taxonomic distributions with a *p*-value $\leq 1 \times 10^{-25}$. Of those characters, the usages of 12 codons were completely congruent with species relationships in the OTL. In Table 3 we report the 12 codons with a p-value $\leq 1 \times 10^{-25}$, and for each codon we report the probability of the taxonomic distribution, the name of the ortholog, and a short description of the species division.

Discussion

Using the species relationships reported in the OTL, we identified codons that, once lost or gained, continued in the same character state to all leaf nodes. Two examples of stop codons that persist through evolutionary time from deep nodes to shallow nodes are in the TNFAIP8 (Tumor Necrosis Factor) and RHOA (encodes small GTPase) genes. Both genes play a role in tumor

progression, and the specific stop codons used separate most mammals from other vertebrates. Other codons with a singular gain/loss that occurs in deep nodes are outlined in Table 3.

Since each gain/loss event outlined in Table 3 has a probability of occurring by random chance that is less than 1×10^{-25} and we studied only 5.2854×10^{4} codons that could be congruent with the OTL at that *p*-value threshold, if codon congruence with the OTL were due to random chance, it would be highly unlikely to identify any groups congruent with the OTL. We identified 12 codon usages that were congruent with taxonomic groups found in the OTL. In contrast to the overall analysis where all codons were equally likely to be included as completely congruent with the OTL, in these deep nodal comparisons, nine of the reported codons are stop codons. Since nonsense or nonstop mutations often affect gene function and the stop codon usage persists through time, it is not unreasonable to expect these orthologs to be crucial for species fitness. These codons also lend support to deep species relationships for which these codons map. In conjunction with other methods, codon usage can add support to proposed species trees.

For instance, several controversial nodes were recently analyzed by Shen et al. (2017). In their analysis, they concluded that turtles are the sister taxa to archosaurs (birds and crocodiles) instead of being sister taxa to only crocodiles, and the OTL was updated to reflect this taxonomic relationship. We evaluated this change to the OTL by determining the probability of each tree (with turtles as sister taxa to archosaurs and with turtles as sister taxa to crocodiles) based on codon usages. We used the other vertebrates taxonomic group and changed only the location of turtles on the tree. We found that the probability of observed codon usages randomly mapping to

the other vertebrates tree with turtles as sister taxa to crocodiles was 7.3306×10^{16} higher than the probability of turtles being sister taxa to archosaurs. This analysis lends support to keeping turtles as sister taxa to archosaurs on the OTL because the probability of the codons randomly mapping to that tree is smaller than the probability of the codons randomly mapping to the other tree.

We recognize a bias toward recovering shallow nodes using this method because many orthologs are not yet annotated for all species. To overcome this bias, we looked at codon usages that were statistically unlikely to be congruent with the OTL and found that across all species, codons were 1 109x more likely to be congruent with the OTL than expected by random chance. Of the 890 814 parsimony informative codons, 590 366 (66.27%) had ortholog annotations for at least 100 species, and 6 688 (25.95%) of the 25 771 codons that were congruent with the OTL taxonomic relationships were from orthologs annotated in at least 100 species. Furthermore, 11 codons whose usage was congruent with the OTL were identified from orthologous genes that were annotated in more than 1 000 species. Identifying complete codon congruence with the OTL in thousands of groups of at least 100 species shows that homology in codon usage can exist in larger taxonomic groups. By performing random permutations of our dataset, we also show that there is less congruence between the codons that were not congruent with the species relationships in the OTL than with the original codon dataset. This analysis shows that although the majority of codon usages do not have a singular gain/loss when mapped to the OTL, codon usages are more likely to follow species relationships in the OTL than in a random phylogeny.

Although we do not have sufficient ortholog annotations to conclude codon congruence with the OTL in fungi, invertebrates, or protozoa, this analysis shows that codon usage is maintained in archaea, bacteria, plants, mammals, and other vertebrates. We also propose that the framework that we provide for performing this analysis will reveal a phylogenetic signal in the other taxonomic groups when more orthologs are annotated across those species. Looking forward, we anticipate that codon usage will become another tool for evaluating different species trees, similar to our evaluation of the placement of turtles on the OTL.

Acknowledgements

We appreciate the contributions of Brigham Young University and the Fulton Supercomputing Laboratory at Brigham Young University for supporting our research.

Competing Interests

The authors declare no competing interests.

References

Cannarozzi, G., Schraudolph, N.N., Faty, M., von Rohr, P., Friberg, M.T., Roth, A.C., Gonnet, P., Gonnet, G., Barral, Y., 2010. A role for codon order in translation dynamics. Cell 141, 355-367.

Coordinators, N.R., 2013. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 41, D8-D20.

Crick, F., 1970. Central dogma of molecular biology. Nature 227, 561-563.

Crick, F.H., 1966. Codon--anticodon pairing: the wobble hypothesis. J Mol Biol 19, 548-555.

Crick, F.H., 1968. The origin of the genetic code. J Mol Biol 38, 367-379.

Crick, F.H., Barnett, L., Brenner, S., Watts-Tobin, R.J., 1961. General nature of the genetic code for proteins. Nature 192, 1227-1232.

Farris, J., 1983. The logical basis of phylogenetic analysis.

Farris, J.S., 2008. Parsimony and explanatory power. Cladistics 24, 825-847.

Felsenstein, J., 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Systematic Biology 27, 401-410.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17, 368-376.

Gingold, H., Tehler, D., Christoffersen, N.R., Nielsen, M.M., Asmar, F., Kooistra, S.M., Christophersen, N.S., Christensen, L.L., Borre, M., Sorensen, K.D., Andersen, L.D., Andersen, C.L., Hulleman, E., Wurdinger, T., Ralfkiaer, E., Helin, K., Gronbaek, K., Orntoft, T., Waszak, S.M., Dahan, O., Pedersen, J.S., Lund, A.H., Pilpel, Y., 2014. A dual program for translation regulation in cellular proliferation and differentiation. Cell 158, 1281-1292.

Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., Bruford, E.A., 2015. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res 43, D1079-1085.

Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., Crandall, K.A., Deng, J., Drew, B.T., Gazis, R., Gude, K., Hibbett, D.S., Katz, L.A., Laughinghouse, H.D.t., McTavish, E.J., Midford, P.E., Owen, C.L., Ree, R.H., Rees, J.A., Soltis, D.E., Williams, T., Cranston, K.A., 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci U S A 112, 12764-12769.

Huelsenbeck, J.P., Crandall, K.A., 1997. Phylogeny Estimation and Hypothesis Testing Using Maximum Likelihood. Annual Review of Ecology and Systematics 28, 437-466.

Miller, J.B., Hippen, A.A., Belyeu, J.R., Whiting, M.F., Ridge, P.G., 2017. Missing something? Codon aversion as a new character system in phylogenetics. Cladistics, n/a-n/a.

Pechmann, S., Frydman, J., 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol 20, 237-243.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., Murphy, M.R., O'Leary, N.A., Pujar, S., Rajput, B., Rangwala, S.H., Riddick, L.D., Shkeda, A., Sun, H., Tamez, P., Tully, R.E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D.R., Murphy, T.D., Ostell, J.M., 2014. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42, D756-763.

Pruitt, K.D., Katz, K.S., Sicotte, H., Maglott, D.R., 2000. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet 16, 44-47.

Quax, T.E., Claassens, N.J., Soll, D., van der Oost, J., 2015. Codon Bias as a Means to Fine-Tune Gene Expression. Mol Cell 59, 149-161.

Rokas, A., Carroll, S.B., 2008. Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol 25, 1943-1953.

Sanderson, M.J., Hufford, L., 1996. Homoplasy : the recurrence of similarity in evolution. Academic Press, San Diego.

Sharp, P.M., Li, W.H., 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15.

Shen, X.-X., Hittinger, C.T., Rokas, A., 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nature Ecology &Amp; Evolution 1, 0126.

Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K., Tolstoy, I., 2014. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res 42, D553-559.

UniProt Consortium, T., 2018. UniProt: the universal protein knowledgebase. Nucleic Acids Res 46, 2699.

Wald, A., 1949. Note on the Consistency of the Maximum Likelihood Estimate. The Annals of Mathematical Statistics 20, 595-601.

Yang, Z., Goldman, N., Friday, A., 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol 11, 316-324.

Tables and Figures

Chapter 3 Tables

Table 3.1. The Probability of Codons Mapping to the OTL Tree Topology Due to Random Chance Assuming No Phylogenetic Signal in Codon Usage

Taxonomic Group	Number of Different Taxonomic Distributions	Number of Codons with no Parallelisms or Reversals OTL	*t*-statistic	*p*-value	Best Random Permutation *t-statistic*	Best Random Permutation *p-value*	Number of Random Permutations with *p-value* is less than or equal to the observed
All	62,416	25,771	4.12488×10^{46}	0	1.77991×10^{3}	1.0	0
Archaea	2,320	925	8.28913×10^{20}	0	2.93471×10^{3}	3.26479×10^{-17}	0
Bacteria	58,368	6,639	1.53961×10^{13}	0	1.93163×10^{3}	1.0	0
Fungi	16	2,019	2.55445×10^{2}	9.37392×10^{-46}	5.99657×10^{2}	4.19847×10^{-118}	16
Invertebrates	182	124	2.78440×10^{2}	4.34253×10^{-6}	6.54011×10^{2}	9.54888×10^{-55}	3
Plants	477	1,702	3.90146×10^{9}	0	1.39101×10^{3}	1.89724×10^{-90}	0
Protozoa	21	2,449	2.14626×10^{3}	0	6.80127×10^{2}	3.53259×10^{-131}	0
Mammals	2,162	10,029	1.32695×10^{22}	0	4.04197×10^{3}	3.34856×10^{-117}	0
Other Vertebrates	2,770	11,877	1.50814×10^{25}	0	1.16996×10^{3}	1.0	0

The first column shows the species divisions, with the first row being a combination of all species. The second column shows the number of taxonomic distributions. The third column shows the number of codon characters that completely follow species relationships shown in the OTL. The fourth column shows the t-statistic obtained from performing a chi-square test on the expected number of congruent characters versus the actual number of congruent characters, with respect to the OTL. The fifth column shows the p-value of the data, obtained from the t-statistic and the degrees of freedom from the number of different taxonomic distributions. The sixth column is the best t-statistic obtained from 100 random permutations of the species while maintaining the same tree structure. The seventh column is the p-value obtained from the highest t-statistic from 100 random permutations of the species while maintaining the same tree structure as the OTL. The eighth column shows the number of random permutations where the permutated p-value is ≤ the observed p-value.

Table 3.2. Phylogenetic Signal in Orthologs Spanning Many Species

Taxonomic Group	Maximum Probability	Number of Codons with Probabilities Less Than Maximum	Maximum Number of Codons Expected to Agree with the OTL	Number of Codons that Agree with the OTL	Number Observed Divided by Maximum Expected
All	1.47453×10^{-6}	470 784	0.69419	770	1 109
Archaea	6.97058×10^{-5}	8 108	0.56517	21	37
Bacteria	7.29309×10^{-6}	93 740	0.68365	29	42
Fungi	1.02155×10^{-5}	0	0	0	0
Invertebrates	1.07875×10^{-4}	512	0.055232	0	0
Plants	1.63371×10^{-5}	1 944	0.031759	57	1 795
Protozoa	1.78508×10^{-5}	0	0	0	0
Mammals	2.64704×10^{-6}	247 714	0.65571	420	641
Other Vertebrates	2.83905×10^{-6}	218 129	0.61928	610	985

The first column is the taxonomic group analyzed. The second column shows the maximum probability of a codon being completely congruent with the OTL by random chance based on one divided by the total number of parsimony informative characters within that taxonomic group. The third column shows the number of codon characters with a probability less than or equal to the second column. The fourth column shows the maximum number of codons expected to agree with the OTL, assuming all codons in column three had the maximum probability shown in the second column. The fourth column is the product of columns two and three. The fifth column is the number of observed codons that agree with the OTL and have a probability of being congruent with the OTL less than or equal to the maximum probability from the second column. The sixth column is the quotient of the fifth and fourth columns, showing the magnitude difference between the observed and expected codon congruence with the OTL.

Table 3.3. Taxonomic Distributions with a p-value ≤ $1x10^{-25}$

Probability by random chance	Ortholog name	Codon	Description of bipartitions
2.47×10^{-47}	TNFAIP8	TGA	91 mammals use TGA while 75 non-mammalian vertebrates including *Ornithorhynchus anatinus* (mammal) do not use TGA
1.22×10^{-45}	RHOA	TAA	74 non-mammalian vertebrates and marsupials use TAA while 85 mammals do not
2.04×10^{-43}	DSTN	AGA	58 species starting at alligators and birds do not use AGA, while 103 mammals and other vertebrates use AGA
2.03×10^{-37}	CD164	TAG	43 species starting at geckos, turtles, and birds, use TAG. 113 mammals and other vertebrates do not use TAG
1.55×10^{-29}	MSANTD1	GGC	29 bird species do not use GGC, while 123 mammals and other vertebrates do use GGC
3.22×10^{-29}	PARD6B	TGA	26 fish do not use TGA. 154 mammals and other vertebrates do use TGA
2.42×10^{-28}	SGK1	TAG	27 fish use TAG. 127 mammals and other vertebrates do not use TAG
4.46×10^{-28}	GABRQ	TGA	32 other vertebrates, including *Ornithorhynchus anatinus,* use TGA. 89 mammals and other vertebrates do not use TGA.
7.17×10^{-27}	PNPO	AGT	24 birds do not use AGT. 141 other vertebrates and mammals use AGT
2.03×10^{-25}	BCORL1	TAG	23 fish use TAG. 132 mammals and other vertebrates do not use TAG
6.44×10^{-25}	FLRT2	TAA	21 small rodents use TAA. 157 mammals and other vertebrates do not use TAA
6.44×10^{-25}	FLRT2	TGA	21 small rodents do not use TGA. 157 mammals and other vertebrates use TGA

The first column is the probability of the taxonomic distribution randomly separating the species according to the OTL classifications. The second column is the name of the orthologous gene. The third column is the codon. The fourth column is a short explanation of how the species were separated.

Chapter 3 Figures

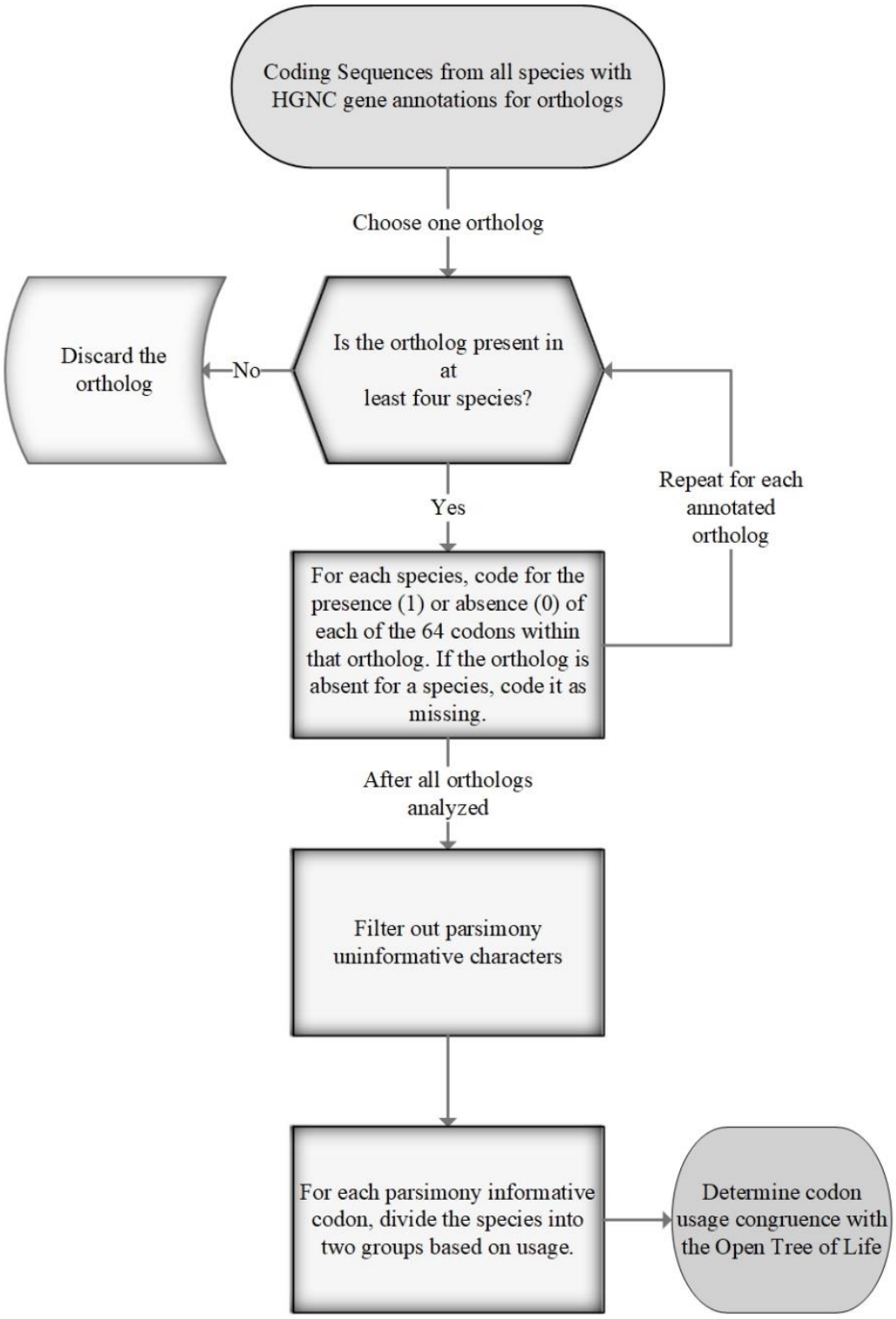


Figure 3.1. The Process for Encoding Codon Usage Codon characters are encoded as either present (1) or absent (0) if they are used or not used in an ortholog, respectively.

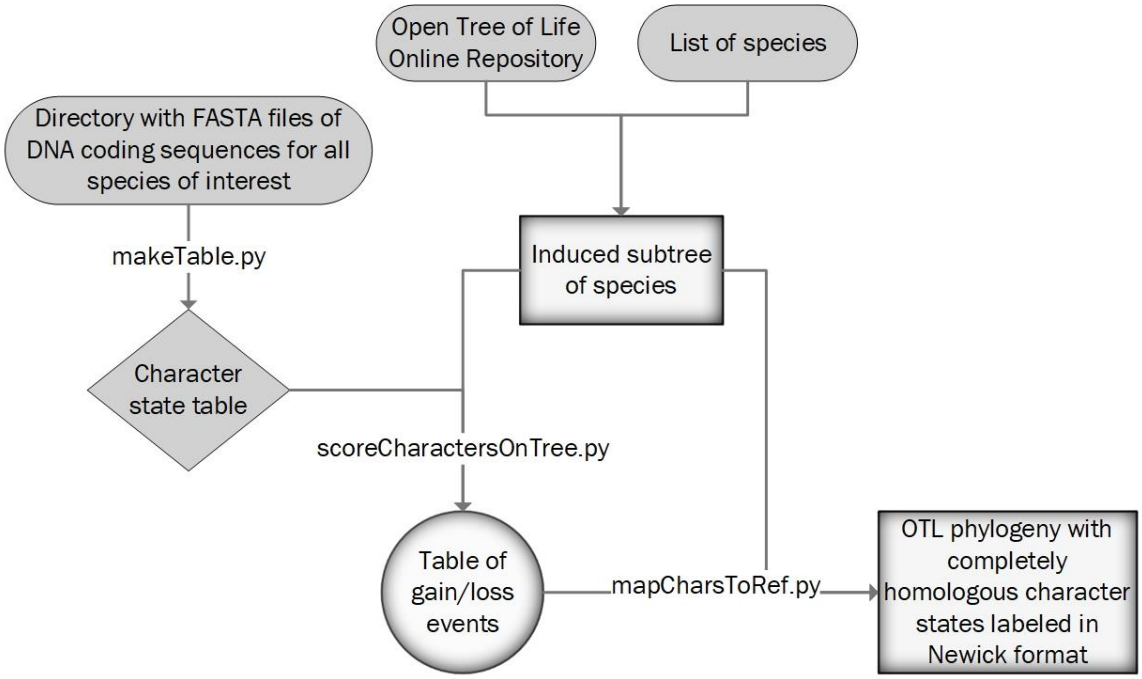


Figure 3.2. Process for Mapping Completely Congruent Character States to the OTL Starting with a directory where each species has a single FASTA file, a character state table is made, which annotates binary codon usages for all species. This table is passed to an induced subtree from the OTL and creates a table of codon usage transition events. From there, these gains/losses are plotted to the OTL induced subtree, and the phylogeny is reported in Newick format.

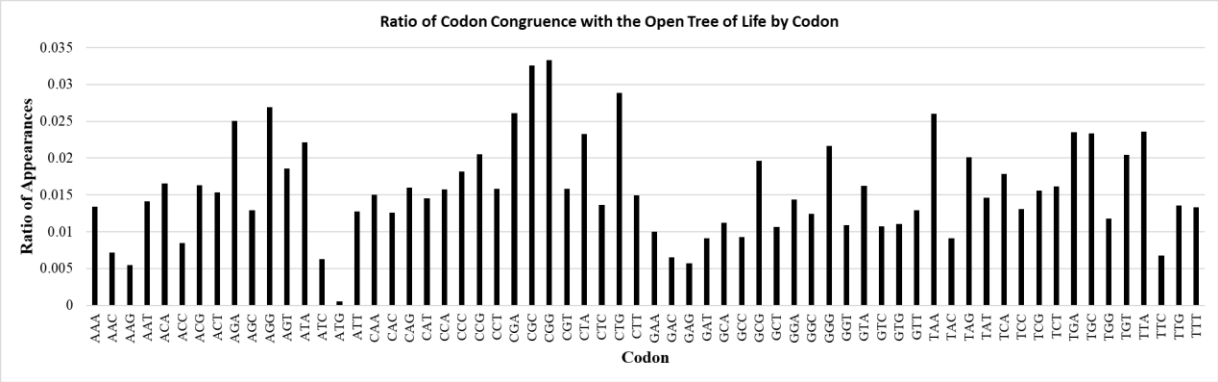


Figure 3.3. The Ratio that Each Codon with a Usage Congruent to the OTL If all codons were given equal weight, the null ratio would be 1.0 / 64 = 0.015625. Observed ratios do not statistically vary from the null, meaning that if a codon usage is congruent with the species relationships outlined in the OTL, it is equally likely for it to be any of the codons.

Manuscript submitted to *PeerJ*

Chapter 4

CAM: An alignment-free method to recover phylogenies using codon aversion motifs

Justin B. Miller[1], Lauren M. McKinnon[1], Michael F. Whiting[1,2], and 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

Abstract

Common phylogenomic approaches for recovering phylogenies are often time-consuming and require annotations for orthologous gene relationships that are not always available. In contrast, alignment-free phylogenomic approaches typically use structure and oligomer frequencies to calculate distances between species. Utilizing a novel alignment-free character state, we present CAM, an alignment-free approach to recover phylogenies using differences in codon aversion motifs (i.e., the set of unused codons within gene sets) between species. Synonymous codon usage is non-random and differs between organisms, between genes, and even within a gene. Many genes do not use all codons. We report a comprehensive analysis of codon aversion within 229 742 339 genes from 23 428 species across all kingdoms of life, and provide an alignment-free framework for its use in a phylogenetic construct. For each species, we constructed a set of tuples, where each tuple contains an ordered set of unused codons for a given gene. We define the pairwise distance between two species, A and B, as one minus the number of direct overlaps over the total possible overlaps. Total possible overlaps is the number of tuples in the set, for A or B, containing the fewest tuples, and direct overlaps is the intersection of tuples in the two sets. This approach allows us to calculate pairwise distances, even with substantial differences in the number of genes for each species. Finally, we use neighbor-joining to recover phylogenies. Using the Open Tree of Life and NCBI Taxonomy Database as expected phylogenies, our approach compares well, recovering phylogenies that largely match expected trees and are comparable to trees recovered using maximum likelihood and a k-mer based alignment-free approach. However, our technique is much faster than maximum likelihood and more congruent with the Open Tree of Life than the k-mer based approach.

Availability: CAM, documentation, and test files are freely available on GitHub at https://github.com/ridgelab/cam

Key Words: alignment-free phylogeny; codon aversion; tree of life; species classification; maximum likelihood.

Contact: perry.ridge@byu.edu

Supplementary Information: Supplementary information are available at Molecular Phylogenetics and Evolution online

Introduction

Phylogenies allow biologists to analyze similar characters between species by providing an evolutionary framework to infer homology (Haszprunar, 1992; Soltis and Soltis, 2003). Although Next Generation Sequencing (NGS) facilitates placement of novel species on the Tree of Life, many regions of the genome display contradictory phylogenetic signals (Philippe et al., 2011). Furthermore, typical alignment-based phylogenetic methods require ortholog annotations to recover the phylogeny, and assembled genes without orthologous pairs provide no information for species relatedness using a traditional approach (Pais et al., 2014b). Annotating a genome with orthologous relationships can often be costly and time-consuming, and some genes are currently impossible to annotate (Yandell and Ence, 2012). As complete genomes of more non-model organisms become available, correctly identifying orthologs will continue to impede the correct identification of taxa relationships. Common errors in recovering phylogenies include incorrect ortholog identification, erroneous alignments, and model violations for the phylogenetic tree reconstruction method (Philippe et al., 2011). To address these issues, alignment-free methods were developed to recover phylogenies based on oligomer frequency and Chaos Theory across the whole genome, without being subject to potential errors in orthology (Vinga and Almeida, 2003). These methods claim to recover phylogenetic relationships even when genetic recombination renders an alignment impossible (Vinga and Almeida, 2003). More recently, proteomes have been used to construct frequency profiles of amino acids or DNA k-mers, which are then used to recover phylogenies (Jun et al., 2010a). In our analysis, we limit our search space to coding sequences and compare the codon usages between species, ignoring all gene name annotations.

In the Central Dogma of biology, three consecutive nucleotides of coding DNA, called codons, are used as a template for protein translation, where each codon encodes a single amino acid (Crick, 1970). The genetic code is degenerate because 64 canonical codons are used to form 20 amino acids and the stop signal (Crick et al., 1961). Gene expression is fine-tuned, in part, by the skewed occurrence of certain codons over others, called codon usage bias, because some codons are translated more efficiently than others (Quax et al., 2015). Differences in codon translational efficiencies are explained by unequal tRNA expression within different species and tissues, limiting the supply of anticodons directly complementing the codons (Quax et al., 2015). Complete codon aversion (i.e., when a codon is not used in a gene) can also be advantageous in certain genes, and is phylogenetically conserved in orthologs (Miller et al., 2017a).

Our research explores the conservation of codon aversion and determines if codon aversion motifs (i.e., sets of codons not used in a gene) are phylogenetically conserved. We present a novel alignment-free algorithm, CAM, which we use to recover a phylogeny using the codon aversion of 229 742 339 genes from 23 428 species across the Open Tree of Life (OTL) (Hinchliff et al., 2015) and the NCBI taxonomy (Sayers et al., 2012; Sayers et al., 2011; Sayers et al., 2010; Sayers et al., 2009). Our results suggest that codon aversion is conserved and can be utilized to reconstruct phylogenetic trees without a sequence alignment.

Materials and Methods

Data Collection and Processing

We downloaded all coding sequences (CDS) from the National Center for Biotechnology Information (NCBI) in September, 2017 (Pruitt et al., 2014; Pruitt et al., 2000; Wheeler et al.,

2007). The CDS regions of the reference genomes were derived from the most common allele from multiple samples of different individuals within each species (Pruitt et al., 2000; Wheeler et al., 2007). When multiple transcript isoforms were annotated, we used the longest isoform in order to include the most possible codons used in a gene. Additionally, we removed any annotated exceptions from the gene dataset (i.e., translational exceptions, unclassified transcription discrepancies, suspected errors, etc.). Most sequences do not have annotated exceptions, and these filters removed fewer than 5% of sequences from each species. Partial gene annotations were included in the analysis. Although not present in most species, some species included large numbers of partial gene sequences, so we included partial gene sequences in the main analysis (See Supplementary Figure 1 for the percentage of partial protein sequences in each taxonomic group). We also compared the phylogenies recovered with and without partial gene sequences to determine the robustness of this method to partial gene inclusion.

Data Analyzed

Our analysis included 23 428 species, which were divided into the following taxonomic groups based on annotations within the NCBI database: 418 archaea, 15 068 bacteria, 234 fungi, 149 invertebrates, 89 plants, 75 protozoa, 107 mammalian vertebrates, 123 other vertebrates, and 7 233 viruses. Sixty-eight species are included in both bacteria and viruses.

Codon Aversion Motif Calculation

We define a codon aversion motif as an ordered set of codons that are not present in a gene. We represent these codons as tuples so they can be added to an unordered set of unique codon motifs for a species and compared using fast set operators (i.e., intersection) to find shared motifs

between species. Although some tuples overlap within the same species, we use a strict definition of a set, where only a single instance of each tuple is stored in the motif. We calculate the pairwise distance between two species, A and B, by one minus the relative similarity of codon aversion motifs between the species. The relative similarity of codon aversion motifs is calculated by dividing the number of directly overlapping motifs between the two species by the total number of possible overlapping motifs. We define directly overlapping motifs as the intersection of tuples in the two sets. The total possible overlapping motifs is defined as the number of tuples in the set, for A or B, containing the fewest tuples. This approach allows us to calculate pairwise distances (with a maximum distance of one), where closely related species have smaller distances than distantly related species, even when substantial differences in the number of genes for each species exist. We also require that 5% of motifs between species overlap to limit small genome bias (e.g., it would not be unusual if a species with five genes has at least one codon usage motif that randomly overlaps with a motif from a species with 20 000 genes without directly inheriting 20% of its genes from the same most recent common ancestor). This process is depicted in Figure 1.

The most common way to run CAM in Python 2.7 is using the following command, where `${DIR}` is a directory with all compressed or uncompressed species FASTA files, one for each species, and ${MATRIX} is the path to a distance matrix that will be created:

```
python cam.py -i ${DIR}/* > ${MATRIX}
```

After the distance matrix was created, we used a Biopython (Talevich et al., 2012) implementation of neighbor-joining to recover the phylogenetic tree. Neighbor-joining was used to combine the pairwise species distances because each pairwise distance represented a distance

based on codon aversion motifs present in a species, not homologous locations of the codon aversion motifs. We provide a python script, makeNewick.py, that calculates the phylogenetic tree from the output matrix created by CAM using the following command:

```
python makeNewick.py -i ${MATRIX} -o ${OUTPUT}
```

All algorithms, with accompanying README and test files, are freely available from GitHub at: https://github.com/ridgelab/cam.

Amino Acid Aversion Motifs

Similar to codon aversion motifs and the steps outlined in Figure 1, we also determined if amino acid aversion is phylogenetically conserved in an alignment-free framework. First, we translated the DNA/RNA sequences to protein sequences. Next, we made a tuple of unused amino acids within that sequence, following each of the steps outlined in Figure 1 by substituting amino acids for codons. This automated process is included in cam.py with usage details in the accompanying README.

Summary of Options

Several additional options are available for cam.py that allow users greater flexibility to run CAM and recover a distance matrix based on their preferences. An input FASTA file must be provided either using a list (standard bash expansion) through the `-i` option, or by providing the name of a directory through the `-id` option. Compressed files (gzip) are accepted and automatically handled with the `.gz` file extension. By default, all processing cores are used by CAM, although any number of cores can be specified by using the `-t` option. By default, the output is written to standard out, although an output file path can be supplied by using the `-o`

option. If memory constraints are an issue, the distance matrix can be calculated species-by-species through the `-w` option; however, the header line will be written at the end of the file instead of the beginning if this option is used. By default, DNA sequences are expected by CAM. For convenience, we also provide the `-rna` flag if the FASTA files are RNA sequences and the `-a` flag if they are protein sequences. If the user desires to run amino acid motifs from DNA or RNA sequences, we also provide the `-aa` option which translates DNA or RNA (if the `-rna` flag is set) to amino acids. Finally, by default species must share at least 5% of their usage motifs to not be given the maximum distance (1.0). This option can be modified using the `-p` option, although it is not recommended to change this option if the species have few genes because the 5% threshold prevents false positives from small genome bias.

Reference Phylogenies

In order to determine the accuracy of our phylogenetic trees, we compared them to reference trees from both the OTL and the NCBI Taxonomy Browser. Although the NCBI Taxonomy Browser is not considered a primary source for taxonomic phylogenetic information because it gathers phylogenetic annotations from many sources, it provides useful information for our analysis because it includes more species than the OTL. Both databases combine research from various studies to construct a tree. We assessed the accuracy of trees reconstructed via codon aversion by comparing our recovered trees to trees from each of these databases.

Extracting Phylogenies from the Open Tree of Life

We used the OTL documentation for programmatically inferring subtrees to develop a Python 2.7 program, getOTLtree.py, that retrieves subtrees from the OTL. Although other OTL parsers,

such as ROTL (Michonneau et al., 2015), are available, getOTLtree allows users to obtain a subtree of any number of species from the OTL in a single step. Inferring subtrees from a set of species requires accessing the OTL database twice: first to retrieve OTL Taxonomy Identifiers (OTT ids) for each species, and second to retrieve the phylogenetic tree. getOTLtree does both commands in a single step at runtime, prompting the user to manually select the correct domain of life when duplicates are found in the OTL database (e.g., *Nannospalax galili* is listed as a eukaryote [OTT id: 207281] and as a bacterium [OTT id: 5909124]). Furthermore, we account for the OTL command, match_names, which limits identical matching of species to 1 000 names, by combining results from multiple queries of fewer than 1 000 species. This process makes large-scale species analyses easier and takes only a few seconds to extract a phylogeny of 2 000 species on a single processing core. If each species is listed on a different line (or CSV or Newick format) in a file called `${INPUT}`, the typical usage for extracting the tree from the OTL is:

```
python getOTLtree.py -i ${INPUT}
```

getOTLtree, accompanying test files, and a README with more detailed explanations of how to run the program with different options are also available in the GitHub repository at https://github.com/ridgelab/cam. A summary of the process behind getOTLtree is depicted in Figure 2.

Extracting Phylogenies from the Open Tree of Life

The NCBI Taxonomy Browser (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) has many tools to enable large queries of its database. We opted to include unranked taxa in our analysis to maximize the

number of included species. We then downloaded the phylogeny in PHYLIP (Felsenstein, 1989) format directly from the website, and we used the extracted phylogenies in our analyses.

Tree Comparison

We used the ete-compare module from the Environment for Tree Exploration toolkit (ETE3) (Huerta-Cepas et al., 2010; Huerta-Cepas et al., 2016) to quantify the similarity between the tree constructed using codon aversion and the corresponding reference trees from the OTL and the NCBI taxonomy. The following command calculates edge similarity of an unrooted tree, where `${INPUT}` is the path to the recovered tree and `${REF}` is the path to the reference tree from the OTL or the NCBI taxonomy:

```
ete3 compare -t {INPUT} -r {REF} --unrooted
```

We selected the percentage of edge similarity (i.e., the number of branches in one tree that are present in the other tree) to compute the topological distance between both trees. This metric was selected based on the following criteria: capability to efficiently compare very large trees, capability to compare unrooted trees (neighbor-joining is unrooted by definition (Saitou and Nei, 1987) and we wanted to account for potential variations at the root node in the reference tree), and capability to compare trees with polytomies. Although several tree-comparison metrics exist, many suffer from problems ranging from high computational cost to lack of robustness (Lin et al., 2012). Advantages for using the percentage of edge similarity metric from the compare method in ETE3 include: clarity in comparing the output as a percentage of congruent branches between trees, optimization for large datasets, capability to compare unrooted trees, and robustness to polytomies (Huerta-Cepas et al., 2016). The advantages and disadvantages of several common tree comparison techniques are listed in Supplementary Table 1.

Validation Using Maximum Likelihood

Since maximum likelihood (Felsenstein, 1981) has been widely used to construct the current version of the OTL, there is a potential confirmation bias when comparing it to the OTL (i.e., it is likely to have an artificially high percent overlap with the species relationships found in the OTL since it was used to create the OTL). However, it is still widely used and should be evaluated against our alignment-free technique. Using ortholog annotations approved by the HUGO Gene Nomenclature Committee (HGNC) (Gray et al., 2015), we extracted the most commonly used orthologs in each taxonomic group. Although we performed no formal tests for orthology, in cases where duplicated genes with the same gene names existed (e.g., RPS4 in the mitochondrion and rps4 in the chloroplast are both listed in *Arabidopsis thaliana*), both genes were removed. After this filtering, we performed a multiple sequence alignment (MSA) on the DNA sequences of each ortholog using the following CLUSTAL OMEGA (Sievers and Higgins, 2018) command:

```
clustalo -i ${INPUT} > ${OUTPUT}
```

We used CLUSTAL OMEGA because it performed very well in full-length sequence comparisons presented by Pais et al. (2014a), and we used full-length gene sequences in our analyses. After each MSA was completed, we created a super-matrix by concatenating the alignments from all orthologs for each species (if an ortholog was not annotated for a species, all nucleotide characters for that ortholog were expressed as "-" for that species). After the super-matrix was created, we used the following IQ-TREE (Nguyen et al., 2015) command to automatically choose the correct model (Posada and Crandall, 1998) and perform maximum likelihood to recover the phylogeny:

```
iqtree -s ${INPUT} -m TEST -pre ${OUTPUT}
```

The recovered phylogeny was then compared to the OTL and the NCBI Taxonomy using the unrooted compare method from ETE3 to identify branch similarities.

Comparison with Traditional k-mer Approach

One alignment-free technique to recover phylogenies is to create a feature frequency profile (FFP) which consists of counting the occurrences of different k-mers and comparing those profiles between species (Jun et al., 2010b; Sims et al., 2009). Although FFP is often used on the whole genome, it can also be used on the proteome (Jun et al., 2010b), which allowed us to do a direct comparison of this approach using our dataset, which consists of all CDS regions. All analyses were done using the step-by-step procedures outlined in the FFP software README. Since the FFP software requires uncompressed data, we uncompressed all FASTA files before conducting the analysis. Preprocessing time was not included in the comparison results.

We included all species FASTA files in a single directory `${DIR}`. If all species names are shorter than 10 characters, they can be included in a single file called `${SPECIES}`. However, if any species names are longer than 10 characters, then a list of numbers (IDs) can be substituted for the species names. We used unique IDs for this step and then converted them back to species names after the tree was recovered. We used the recommended command from the FFP README (https://sourceforge.net/projects/ffp-phylogeny/files/Documentation/) to create the distance matrix, `${MATRIX}`:

```
ffpry -l 5 ${DIR}/* | ffpcol | ffprwn | ffpjsd -p ${SPECIES} >
${MATRIX}
```

After the distance matrix was created in PHYLIP format, we used the same Biopython implementation of the neighbor-joining algorithm that CAM used by specifying the Phlyip input format option (`-p`) of makeNewick.py (provided in the GitHub repository for CAM):

```
python makeNewick.py -p -i ${MATRIX} -o ${OUTPUT}
```

After the Newick tree was recovered and the species IDs were converted back to species names, we compared the recovered tree with the OTL and the NCBI taxonomy using the unrooted compare method in ETE3.

Results

Since 64 codons exist, and each species typically uses only one of three possible stop codons and the one start codon per gene, there are 61 degrees of freedom (64 – 2 unused stop codons – 1 start codon), allowing for 2^{61} possible motifs. We observed 54 336 494 (~2^{26}) motifs across all genomes, with significant overlap between species (see Table 1). When including counts for multiple occurrences of a motif within the same species, there are still more than 5x as many completely unique motifs within that species as overlapping motifs in the same species (See Supplementary Figures 2-11). We also note that not all codons have equal probabilities of being present in a gene, and we show the frequency of codon aversion per codon within each taxonomic group in Supplementary Figures 12-21. Although most genes use most codons, some genes exclude significantly more codons than others. Across all species, the mean number of codons not used within a sequence is 14.4819, with a standard deviation of 8.6881 codons. The number of codons included in each codon motif is depicted in Supplementary Figures 22-31. In Supplementary Figures 32-41, we also show that relatively few motifs are present in more than a few genes.

We show the number of species included in the phylogenies recovered by each algorithm in Table 2. The alignment-free approaches (CAM, amino acid motifs, and FFP) each recovered a tree for all 23 428 species. Insufficient ortholog annotations were available in bacterial species for maximum likelihood to recover a tree for bacteria or all species. Maximum likelihood recovered trees for relatively few fungi (25%), protozoa (32%), invertebrates (38%), and plants (67%) because many of the species did not have ortholog annotations. The NCBI taxonomy included the most species, only missing 2 archaea, 456 bacteria, and 188 viruses. Since the OTL does not include viruses, it contains significantly fewer species, with the inferred phylogeny containing only 12 337 species out of the possible 23 428 species.

We compared the recovered phylogeny from CAM with the reference phylogenies from the OTL (Table 3) and the NCBI taxonomy (Table 4). Bacteria and viruses have the highest similarity with these phylogenies (84-91%), and invertebrates have the lowest similarity (60-70%). However, FFP had a lower similarity with the reference phylogenies than CAM in all taxonomic groups except bacteria (1% more similar). Maximum likelihood recovered trees that were more similar to the reference phylogenies in most taxonomic groups; however, CAM recovered more congruent phylogenies in fungi (4% more similar) and protozoa (1% more similar). Our method lends support to the NCBI taxonomy in every taxonomic group, with reported phylogenies being 3-13% more similar to the NCBI taxonomy than the OTL. We also ran the entire analysis excluding partial protein sequences. Excluding partial genes had a minimal effect on the overall percent overlap with the OTL (minus 2% to plus 5% similarity) and the NCBI taxonomy (minus 2% to plus 3% similarity).

Tables 3 and 4 also show how well each of the other approaches compares with the OTL and the NCBI taxonomy, respectively. In most instances, amino acid aversion motifs performed comparable to codon aversion motifs when compared against the OTL and the NCBI taxonomy. However, the percent overlap between the NCBI taxonomy and amino acid aversion motifs in mammals, other vertebrates, and viruses was much lower than the percent overlap with CAM (9-25% lower). The same trend exists when comparing the recovered trees with the OTL, with amino acid motifs recovering 10-14% fewer species relationships than CAM. The other taxonomic groups did not appear to vary significantly between the recovered trees using amino acids or codons, with the difference between the two methods being -3% to +3% for the NCBI taxonomy and -5% to +2% different for the OTL.

As expected, the NCBI taxonomy and the OTL are highly similar (Table 3), although 6-9% of species relationships disagree outside of invertebrates, plants, and mammals. Although maximum likelihood has been widely used to create the OTL, the alignment-free methods recovered trees that were more congruent with the OTL and the NCBI Taxonomy than maximum likelihood in fungi and protozoa. Feature frequency profiles of k-mers recovered more similar trees to the reference phylogenies than CAM in all species (+1%) and bacteria (+1%). In all other taxonomic groups, CAM recovers trees that are 1-25% more similar to the reference phylogenies than FFP.

Table 5 shows the CPU runtime of each algorithm in hours. The alignment-free techniques had significantly faster runtimes than the maximum likelihood approach. FFP always had the fastest runtime. Runtime was always longer for amino acid motifs than CAM because the DNA

sequences were translated into protein sequences before being evaluated for amino acid usage. In the smaller taxonomic groups outside of bacteria, each of the alignment-free methods computes the phylogeny within minutes. Maximum likelihood required at least 2.5 hours of CPU time to compute a tree for each taxonomic group.

Although the maximum likelihood analysis was not possible on bacteria or all species because insufficient ortholog gene annotations exist to accurately compare the majority of the bacterial species, it would have also been infeasible based on CPU runtime. As more species and orthologs are included in the maximum likelihood analysis, the runtime increases exponentially. The fastest iteration of maximum likelihood finished in 2.5 hours on 100 mammals, using 18 orthologous genes which were each present in at least 97 species. In contrast, CAM used all genes in 107 mammals and finished in 0.2101 hours (12 minutes, 36 seconds). The slowest iteration of maximum likelihood finished in 199.75 hours on 58 fungi using 648 orthologs which were each annotated in at least five species. CAM again analyzed all genes, both annotated and unannotated, across 234 fungi, finishing in 0.2167 hours (13 minutes).

In Table 6, we report the minimum number of species with an ortholog annotation, the number of orthologs used, and the total number of characters in the super-matrix for each taxonomic group. All orthologous genes with gene annotations spanning at least the number of species noted in column 2 (minimum number of species with orthologs) were included in the analysis. Differences in the minimum number of species with an ortholog are due to differences in the breadth of gene annotations within a taxonomic group. For instance, few orthologous gene annotations spanned more than five species in fungi, invertebrates, and protozoa; however, many

orthologs were annotated in 100 vertebrate species. We did not filter the orthologs on any metric besides the number of species with that gene annotation.

Discussion

The advent of Next Generation Sequencing (NGS) and RNA-seq enables researchers to quickly and inexpensively sequence genomes faster than orthologous relationships and species phylogenies can be annotated and examined. Although CAM requires genomes to be assembled with CDS regions annotated, it does not require an alignment of the genes against other species, nor does it require the time-consuming approaches of traditional methods such as maximum likelihood. As opposed to k-mer based frequency profiles, CAM recovers more similar phylogenies to the OTL and the NCBI Taxonomy in almost all instances. Furthermore, since we combine individual codon aversion motifs from each gene to a set of motifs across the whole species, we are able to allow for greater genetic diversity between species than multiple sequence alignments, which are limited to sequence identity. Since CAM analyzes only the codons that are not present in a gene, sequences that are very different could overlap if they are under the same pressure to maintain the aversion to certain codons.

This technique is also robust to partial gene annotations. Including or excluding partial gene sequences in the analysis had a minimal effect on the overall species relationships. Furthermore, CAM appears to consistently recover accurate phylogenies for all domains of life. This characteristic allows phylogenetic analyses to limit *ad hoc* hypotheses by using a character state that spans all domains of life, instead of piecing together the phylogenetic signal from different genes. Additionally, codon motifs can be used to examine coevolutionary forces between

different domains, such as viruses and hosts. Since similarities in codon usages have previously been identified between some viruses and their respective hosts (Chantawannakul and Cutler, 2008; Miller et al., 2017b), this technique could facilitate coevolutionary analyses by identifying overlapping motifs in distantly related species, which can then be analyzed using traditional techniques.

Although CAM does not recover the same phylogeny as the OTL or the NCBI taxonomy, the recovered phylogenies have comparable percent branch similarities as phylogenies recovered using traditional ortholog-based maximum likelihood estimates. For protozoa, the percent similarity with the OTL and the NCBI taxonomy was only 1% different between maximum likelihood and CAM. However, ortholog annotations were available for only 24 species, whereas CAM recovered 75 species relationships. Species relationships recovered for archaea, mammals, and other vertebrates were more similar to established phylogenies using maximum likelihood; however, since traditional ortholog-based techniques were used to construct the current representation of the OTL, it is expected that clades with well-documented orthologs should recover very similar trees to the reference. Codon aversion motifs display a strong phylogenetic signal in all domains of life, and the signal is similar to ortholog-based maximum likelihood in fungi and protozoa.

The recovered phylogenies from CAM were more similar to the OTL and the NCBI taxonomy than phylogenies recovered using feature frequency profiles in the following taxonomic groups: fungi, invertebrates, plants, mammals, and other vertebrates. FFP and CAM recovered

comparable trees from all other taxonomic groups. Although FFP CPU runtime was faster than CAM, it was not as accurate in most instances.

Codon aversion motifs provide a basis for alignment-free methods to recover robust phylogenies quickly and with sufficient resolution to account for future species discovery. In contrast to maximum likelihood, most cladal relationships were recovered using CAM within minutes. Furthermore, without relying on gene alignments, the recovered phylogeny is not dependent on the accuracy of the aligner or ortholog annotations, which allows for a more universal technique to compare distantly related species that might have incorrectly labeled genes or very divergent gene sequences.

We understand that certain limitations to our study exist. For instance, while we have shown that CAM successfully recovers most known species relationships and can be used in future alignment-free analyses to recover comparable phylogenies to maximum likelihood, we do not fully understand the biological mechanisms that govern the phylogenetic signal we identified. Future research will examine the processes involved in maintaining this phylogenetic signal, including the mechanism that maintains complete codon aversion within a gene. We also note that alignment-free methods often appear as a "black box" to researchers who are accustomed to homologous character analyses that allow for directly identifying nucleotide differences in sequences. While CAM presents a paradigm shift, it has the potential to be as informative as analyses of homologous character states. Since CAM is based in codon usages within each gene, we propose that percent similarities in codon aversions between species represents similarities in the mechanisms that maintain these codon usages. Although these mechanisms are presently not

fully understood, we show that they are phylogenetically conserved and can be utilized using our method.

Acknowledgements

We appreciate the contributions Brigham Young University for sponsoring our research. We also appreciate the Fulton Supercomputing Laboratory staff for their continued support.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest: None declared.

References

Chantawannakul, P., Cutler, R.W., 2008. Convergent host-parasite codon usage between honeybee and bee associated viral genomes. J Invertebr Pathol 98, 206-210.

Crick, F., 1970. Central dogma of molecular biology. Nature 227, 561-563.

Crick, F.H., Barnett, L., Brenner, S., Watts-Tobin, R.J., 1961. General nature of the genetic code for proteins. Nature 192, 1227-1232.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17, 368-376.

Felsenstein, J., 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5, 164-166.

Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., Bruford, E.A., 2015. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res 43, D1079-1085.

Haszprunar, G., 1992. The types of homology and their significance for evolutionary biology and phylogenetics. Journal of Evolutionary Biology 5, 13-24.

Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., Crandall, K.A., Deng, J., Drew, B.T., Gazis, R., Gude, K., Hibbett, D.S., Katz, L.A., Laughinghouse, H.D.t., McTavish, E.J., Midford, P.E., Owen, C.L., Ree, R.H., Rees, J.A., Soltis, D.E., Williams, T., Cranston, K.A., 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci U S A 112, 12764-12769.

Huerta-Cepas, J., Dopazo, J., Gabaldon, T., 2010. ETE: a python Environment for Tree Exploration. BMC Bioinformatics 11, 24.

Huerta-Cepas, J., Serra, F., Bork, P., 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol 33, 1635-1638.

Jun, S.-R., Sims, G.E., Wu, G.A., Kim, S.-H., 2010a. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proceedings of the National Academy of Sciences 107, 133-138.

Jun, S.R., Sims, G.E., Wu, G.A., Kim, S.H., 2010b. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proc Natl Acad Sci U S A 107, 133-138.

Lin, Y., Rajan, V., Moret, B.M., 2012. A metric for phylogenetic trees based on matching. IEEE/ACM Trans Comput Biol Bioinform 9, 1014-1022.

Michonneau, F., Brown, J., Winter, D., 2015. rotl , an R package to interact with the Open Tree of Life data rotl an R package to interact with the Open Tree of Life Data.

Miller, J.B., Hippen, A.A., Belyeu, J.R., Whiting, M.F., Ridge, P.G., 2017a. Missing something? Codon aversion as a new character system in phylogenetics. Cladistics, n/a-n/a.

Miller, J.B., Hippen, A.A., Wright, S.M., Morris, C., Ridge, P.G., 2017b. Human viruses have codon usage biases that match highly expressed proteins in the tissues they infect. Biomedical Genetics and Genomics 2.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32, 268-274.

Pais, F.S., Ruy, P.C., Oliveira, G., Coimbra, R.S., 2014a. Assessing the efficiency of multiple sequence alignment programs. Algorithms Mol Biol 9, 4.

Pais, F.S., Ruy Pde, C., Oliveira, G., Coimbra, R.S., 2014b. Assessing the efficiency of multiple sequence alignment programs. Algorithms Mol Biol 9, 4.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., Baurain, D., 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLOS Biology 9, e1000602.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817-818.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., Murphy, M.R., O'Leary, N.A., Pujar, S., Rajput, B., Rangwala, S.H., Riddick, L.D., Shkeda, A., Sun, H., Tamez, P., Tully, R.E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D.R., Murphy, T.D., Ostell, J.M., 2014. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42, D756-763.

Pruitt, K.D., Katz, K.S., Sicotte, H., Maglott, D.R., 2000. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet 16, 44-47.

Quax, T.E., Claassens, N.J., Soll, D., van der Oost, J., 2015. Codon Bias as a Means to Fine-Tune Gene Expression. Mol Cell 59, 149-161.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4, 406-425.

Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S., Feolo, M., Fingerman, I.M., Geer, L.Y., Helmberg, W., Kapustin, Y., Krasnov, S., Landsman, D., Lipman, D.J., Lu, Z., Madden, T.L., Madej, T., Maglott, D.R., Marchler-Bauer, A., Miller, V., Karsch-Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Wang, Y., Wilbur, W.J., Yaschenko,

E., Ye, J., 2012. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 40, D13-25.

Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S., Feolo, M., Fingerman, I.M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Lu, Z., Madden, T.L., Madej, T., Maglott, D.R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Wang, Y., Wilbur, W.J., Yaschenko, E., Ye, J., 2011. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 39, D38-51.

Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Lu, Z., Madden, T.L., Madej, T., Maglott, D.R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Wang, Y., John Wilbur, W., Yaschenko, E., Ye, J., 2010. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 38, D5-16.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Yaschenko, E., Ye, J., 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 37, D5-15.

Sievers, F., Higgins, D.G., 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci 27, 135-145.

Sims, G.E., Jun, S.R., Wu, G.A., Kim, S.H., 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci U S A 106, 2677-2682.

Soltis, D.E., Soltis, P.S., 2003. The Role of Phylogenetics in Comparative Genetics. Plant Physiology 132, 1790-1800.

Talevich, E., Invergo, B.M., Cock, P.J., Chapman, B.A., 2012. Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. BMC Bioinformatics 13, 209.

Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison—a review. Bioinformatics 19, 513-523.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Miller, V., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L., Yaschenko, E., 2007. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35, D5-12.

Yandell, M., Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics 13, 329.

Tables and Figures

Chapter 4 Tables

Table 4.1. Unique Tuples in Each Taxonomic Group

Taxonomic Group	Number of Unique Motifs	Number of Genes	Average Number of Genes with a Given Motif
All	54 336 494	229 742 339	4.228
Archaea	1 057 898	1 903 114	1.799
Bacteria	49 177 047	215 581 296	4.384
Fungi	904 513	2 194 206	2.426
Invertebrates	951 901	2 153 164	2.262
Plants	1 009 268	2 510 219	2.487
Protozoa	510 582	841 682	1.648
Mammals	732 868	2 004 675	2.735
Other Vertebrates	806 510	2 274 837	2.821
Viruses	234 768	303 129	1.291
Total (without all)	55 385 355	229 766 322	4.149
Total (without all and without bacteria)	5 159 447	14 161 043	2.745

Unique tuples were calculated by adding all tuples of unused codons from all genes within each species from a taxonomic group to a set, and then counting the number of elements in that set. The All group includes all species in the same analysis. Total (without all) sums the number of motifs and genes from each taxonomic group, calculated individually. Since most species in this analysis are bacteria, Total (without all and without bacteria) summed the values from each taxonomic group without including bacteria or all species combined. Note: 23 983 viral and bacterial genes overlap and 1 048 861 motifs span different taxonomic groups (difference between values in All and Total (without all).

Table 4.2. Number of Species Included in Phylogenies

Taxonomic Group	CAM	Amino Acid Motifs	FFP	Maximum Likelihood	NCBI Taxonomy	OTL
All	23 428	23 428	23 428	N/A	22 794	12 337
Archaea	418	418	418	418	416	362
Bacteria*	15 068	15 068	15 068	N/A	14 612	11 227
Fungi	234	234	234	58	234	214
Invertebrates	149	149	149	57	149	147
Plants	89	89	89	60	89	87
Protozoa	75	75	75	24	75	75
Mammals	107	107	107	100	107	105
Other vertebrates	123	123	123	118	123	120
Viruses*	7 233	7 233	7 233	N/A	7 045	N/A

For each algorithm, we report the number of species used to recover the phylogeny. *Note: Some species are included in both bacteria and viruses.

Table 4.3. Comparison to the OTL

Taxonomic Group	CAM	Amino Acid Motifs	FFP	Maximum Likelihood	NCBI Taxonomy
All	82	84	83	N/A	95
Archaea	75	77	74	89	94
Bacteria	84	84	85	N/A	95
Fungi	69	67	67	65	91
Invertebrates	60	57	55	73	98
Plants	64	63	54	73	98
Protozoa	65	65	64	64	93
Mammals	77	63	52	93	99
Other Vertebrates	66	56	54	81	94

Percent edge overlap of an unrooted tree comparison of each algorithm versus the established phylogeny from the OTL for each taxonomic group. Maximum likelihood could not compute a tree for bacteria or all species because insufficient ortholog annotations were available for the majority of these species.

Table 4.4. Comparison to the NCBI Taxonomy

Taxonomic Group	CAM	Amino Acid Motifs	FFP	Maximum Likelihood
All	89	90	90	N/A
Archaea	81	84	80	92
Bacteria	91	90	91	N/A
Fungi	73	69	69	70
Invertebrates	70	68	65	78
Plants	71	70	61	79
Protozoa	72	71	72	73
Mammals	87	73	63	98
Other Vertebrates	79	70	67	95
Viruses	90	65	91	N/A

Percent edge overlap of an unrooted tree comparison of each algorithm versus the established phylogeny from the NCBI taxonomy for each taxonomic group. Maximum likelihood could not compute a tree for bacteria, viruses, or all species because insufficient ortholog annotations were available for the majority of these species.

Table 4.5. CPU Runtime of Each Algorithm in Hours

Taxonomic Group	CAM	Amino Acid Motifs	FFP	Maximum Likelihood
All	17.2794	20.2692	3.9072	N/A
Archaea	0.0667	0.1436	0.0408	161.5
Bacteria	14.6994	17.4458	3.7442	N/A
Fungi	0.0783	0.2167	0.0294	199.75
Invertebrates	0.0763	0.2126	0.0447	2.5
Plants	0.0781	0.2211	0.0383	6.0
Protozoa	0.0287	0.0833	0.0183	4.0
Mammals	0.0718	0.2101	0.0294	2.5
Other vertebrates	0.0872	0.2356	0.0322	6.75
Viruses	0.1028	0.1161	0.1019	N/A

Table 4.6. Matrix Statistics for Maximum Likelihood Analysis

Taxonomic Group	Minimum number of species with ortholog	Number of orthologs in super-matrix	Characters in super-matrix
Archaea	95	45	62 442
Fungi	5	648	1 403 618
Invertebrates	5	20	17 665
Plants	40	75	87 764
Protozoa	5	200	411 028
Mammals	97	18	24 767
Other vertebrates	108	28	30 900

The first column is the taxonomic group. The second column is the minimum number of species which must include an ortholog annotation for it to be included in the matrix. The third column is the number of orthologs with the minimum number of species annotations. The fourth column is the number of nucleotide characters in the combined alignment of all orthologs included in the analysis.

Chapter 4 Figures

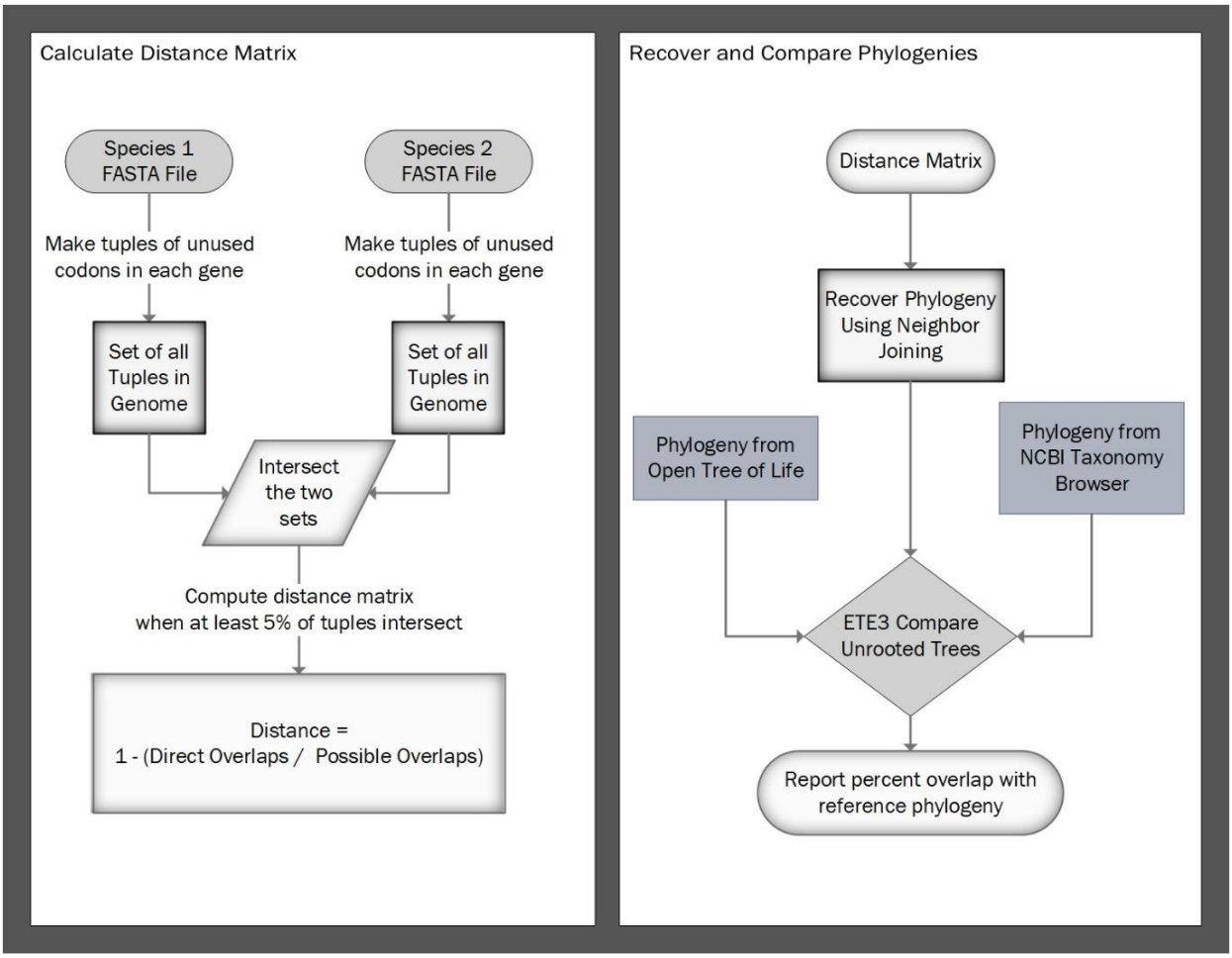


Figure 4.1. Flow Charts for Calculating the Distance Matrix and Comparing the Recovered Phylogenies Calculate Distance Matrix: Start with two FASTA files of the DNA coding sequences of two species. For each species, find the unused codons within each gene, alphabetize them, and make those codons into a tuple. Add the tuple to an unordered set for that species. The distance is calculated by dividing the number of tuples in the intersection of the two sets by the minimum number of tuples in the two original sets.
Recover and Compare Phylogenies: From the distance matrix, use neighbor-joining to recover a phylogeny. We do not use a model of evolution to compute distances because distance is a function of the number of shared codon aversion motifs within a species. This technique allows a fair comparison of diverse or unknown species. Using the compare method within the Environment for Tree Exploration (ETE3), we then compare the unrooted tree with the OTL and the NCBI taxonomy. Finally, we report the percentage of the phylogenies that overlap.

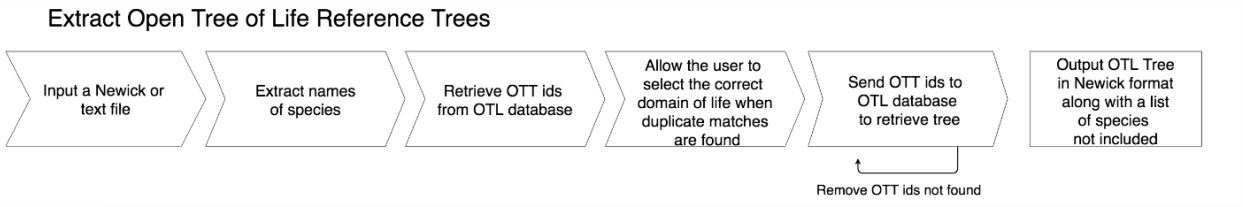


Figure 4.2. Flow Chart Depicting the Process getOTLtree Takes to Infer a Subtree Phylogeny from the OTL All steps are done with a single command at runtime.

Chapter 5

Codon Pairs are Phylogenetically Conserved: Codon pairing as a novel phylogenetic character state for parsimony and alignment-free methods

Justin B. Miller[1,+], Lauren M. McKinnon[1,+], Michael F. Whiting[1,2], and 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*

[+]*Contributed equally to this work*

Abstract

Identical codon pairing and co-tRNA codon pairing increase translational efficiency within genes when two codons that encode the same amino acid are located within a ribosomal window. By examining identical and co-tRNA codon pairing independently and combined across 23 423 species, we determined that both pairing techniques are phylogenetically informative using either an alignment-free or parsimony framework in all domains of life. We also determined that the minimum optimal window size for conserved codon pairs is typically smaller than the length of a ribosome. We thoroughly analyze codon pairing across various taxonomic groups. We determined which codons are more likely to pair and we analyzed the frequencies of codon pairings between species. The alignment-free method does not require orthologous gene annotations and recovers species relationships that are more congruent with established phylogenies than other alignment-free techniques in all instances. Parsimony recovers trees that are more congruent with the established phylogenies than the alignment-free method in four out of six taxonomic groups. Four taxonomic groups do not have sufficient ortholog annotations and are excluded from the parsimony and/or maximum likelihood analyses. Using only codon pairing, the alignment-free or parsimony-based approaches recover the most congruent trees compared with the established phylogenies in six out of ten taxonomic groups. Since the recovered phylogenies using only codon pairing largely match established phylogenies, we propose that codon pairing biases are phylogenetically conserved and should be considered in conjunction with current techniques in future phylogenomic studies.

Availability: All scripts used to recover and compare phylogenies, including documentation and test files, are freely available on GitHub at https://github.com/ridgelab/codon_pairing.

Key Words: [Codon usage bias, identical codon pairing, co-tRNA codon pairing, phylogeny, taxonomy, phylogenetically informative character, alignment-free, parsimony]

Introduction

Phylogenies allow biologists to infer similar characteristics of closely related species and provide an evolutionary framework for analyzing biological patterns (Soltis and Soltis, 2003). Phylogenies are statements of homology, and represent a continuity of biological information (Haszprunar, 1992). Although genetic data allow researchers to analyze more species cheaper and faster than morphological features, molecular data typically require data cleaning (e.g., alignment, annotation, and ortholog identification) before they become useful (Philippe et al., 2011). After orthologs are identified, phylogenies can be recovered through parsimony (Farris, 1983; Wilgenbusch and Swofford, 2003), maximum likelihood (Felsenstein, 1981), Bayesian inference (Yang and Rannala, 2012), or distance-based techniques such as neighbor-joining (Saitou and Nei, 1987). Alignment-free techniques typically use Chaos Theory to calculate distances of basic genomic features (e.g., GC content, oligomer frequency, etc.) that are then used to recover the phylogeny (Vinga and Almeida, 2003; Chan et al., 2014). More recently, other techniques limit the alignment-free search space to genic regions, constructing profiles of amino acids or codon usages (Jun et al., 2010; Chapter 4) .

Codons are sequences of three consecutive nucleotides of coding DNA that are transcribed into mRNA, mRNA is translated into amino acids, and amino acids form proteins (Crick, 1970). The 20 canonical amino acids are formed from 61 codons, with the other three codons encoding the stop signal (Crick et al., 1961). Although multiple codons encode the same amino acid, an unequal distribution of synonymous codons occurs within species, suggesting that synonymous codons might play different roles in species fitness (Sharp and Li, 1986). An unequal distribution

of tRNA anticodons directly coupling codons led to the wobble hypothesis: tRNA anticodons do not need to latch onto all three codon nucleotides for translation (Crick, 1966). Codon usage is also 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).

Recharging a tRNA while the tRNA is still attached to the ribosome is used to increase translational efficiency and decrease overall resource utilization. This process occurs when codons encoding the same amino acid are located in close proximity to each other on the mRNA strand (Cannarozzi et al., 2010). Co-tRNA codon pairing is when two non-identical codons that use the same tRNA are near each other in a gene and the tRNA is recharged to translate both codons before the ribosome diffuses. Similarly, identical codon pairing occurs when identical codons are near each other in a gene and the tRNA is recharged to translate both codons before the ribosome diffuses. Co-tRNA and identical codon pairing conserve resources and increase translational speed by approximately 30% (Cannarozzi et al., 2010). Co-tRNA codon pairing has previously been reported as more prominent in eukaryotes, while identical codon pairing has been reported in eukaryotes, bacteria (Shao et al., 2012), and archaea (Zhang et al., 2013).

We report codon pairing as a phylogenetic character state using both parsimony and alignment-free techniques. Our results suggest that both identical codon pairing and co-tRNA codon pairing are phylogenetically conserved and prominent in all domains of life. We further show that combining the two techniques generally recovers more congruent phylogenies compared with established phylogenies. Codon pairing recovers trees that

are more congruent with the Open Tree of Life (OTL) (Hinchliff et al., 2015) and the NCBI Taxonomy Browser (Sayers et al., 2009; Sayers et al., 2010; Sayers et al., 2011; Sayers et al., 2012) than maximum likelihood trees recovered using IQ-TREE (Nguyen et al., 2015) and other alignment-free methods in six out of ten taxonomic groups.

Materials and Methods

Data Collection and Processing

We downloaded all reference genomes and annotations from the National Center for Biotechnology Information (NCBI) (Pruitt et al., 2000; Wheeler et al., 2007; Pruitt et al., 2014) in September, 2017. Reference genomes were used because they represent the most commonly accepted nucleotides in each species (Pruitt et al., 2000; Wheeler et al., 2007). We used the coding sequences (CDS) from the longest isoform of each gene and we removed annotated exceptions (i.e., translational exception, unclassified transcription discrepancy, suspected errors, partial genes, etc.). A total of 23 423 species were divided into the following taxonomic groups based on NCBI annotations, with some overlap between bacteria and viruses: 418 archaea, 15 063 bacteria, 234 fungi, 149 invertebrates, 89 plants, 75 protozoa, 107 mammalian vertebrates, 123 other vertebrates, and 7 233 viruses. While some of these taxonomic groups do not represent monophyletic clades, we opted to maintain these species classifications to facilitate analyses between different studies that use the NCBI annotations.

Accounting for Differences in Ribosomal Footprint

Estimates of the ribosome footprint vary drastically and can range from 15 nucleotides (5 codons) to about 45 nucleotides (15 codons) with a commonly accepted length of 28

nucleotides (about nine codons) (Martens et al., 2015) . Since codon pairing requires at least two codons, we examined pairing lengths (i.e., a sliding window) of 2-11 codons. This technique allows for variations in the ribosomal footprint among different taxonomic groups and can determine if codon pairing is dispersed throughout the ribosomal footprint or is more phylogenetically conserved at a smaller window size.

Calculating Identical and co-tRNA Codon Pairing

For both the parsimony and alignment-free methods, we used a binary representation of codon pairings, co-tRNA codon pairings, and combined identical and co-tRNA codon pairings (i.e., if a codon paired within a gene, it was given a value of "1" regardless of the number of times the pairing occurred). We determined which codons used identical codon pairing for each gene by adding each codon that occurred multiple times within the sliding window to an ordered set of codons for that gene. Similarly, we created an ordered sets of amino acids for co-tRNA codon pairings for each gene by adding the amino acid product of the paired non-identical codons that encode that amino acid to the ordered set. Since the combined approach uses both identical and co-tRNA codon pairing, we calculated combined pairing by translating the gene sequence and identifying amino acids that paired within the ribosome window, adding each paired amino acid to the ordered set.

Alignment-free Codon Pairing Calculation

We present three alignment-free methods to calculate a distance matrix: 1) based on identical codon pairing, 2) based on co-tRNA codon pairing, and 3) based on a combination of identical and co-tRNA codon pairing. Although genes must be assembled, orthologous relationships are

not required or used in the distance matrix calculation. Both methods use a binary (occurs or does not occur) representation of codon pairing within a gene. First, if identical codon pairing occurs anywhere within a gene, the codons are added to an ordered set. If co-tRNA codon pairing or the combined approach is selected, then amino acids are added to an ordered set if they occur within the ribosomal footprint anywhere in the gene. Next, the sets are converted to a tuple (immutable list) so they can be added to a set for the entire species. This process is repeated for each gene within a species until all gene pairings have been made into tuples and added to a set for the species. We repeat this process for each species until all species have a set of tuples representing the codons (or amino acids) that are pairing within a gene. Finally, we calculate the distances between each species in a pairwise manner. This process is depicted in Figure 1. Similar to the method used in Chapter 4, the pairwise distance between two species, A and B, is calculated as one minus the relative similarity of the species. The relative similarity of the species is the number of overlapping tuples between the sets of tuples in both species divided by the total number of tuples in the species, A or B, with the fewest number of tuples. This ratio must exceed 5% or else the species are assigned the maximum distance of 1.0. This filter limits small genome bias (e.g., without this cutoff, if one gene from a virus with two genes has the same codon pairing profile as a gene in a vertebrate with 20 000 genes, then the distance between the virus and the vertebrate would be 0.5). This process allows us to calculate a distance, with a maximum of 1.0, where more closely related species have a smaller distance to each other because their genes utilize more similar codon pairings. We implemented `pairing_distance.py` in Python 3.5 to calculate the distance matrix based on the alignment-free comparison of identical codon pairing or co-tRNA codon pairing outlined above.

Summary of Alignment-free Options

We provide several additional options for `pairing_distance.py` to give users greater flexibility in their research. Input FASTA files can be provided either as a list (standard bash expansion) with the `-i` option, or included in a single directory with the `-id` option. The program automatically handles gzipped compressed files with the `.gz` or `.gzip` file extension or uncompressed data with any other file extension. The output distance matrix by default is written to standard out, although an output file can be provided through the `-o` option. Although all available processing cores are used by default to calculate the distance, this can be modified with the `-t` option. RNA sequences can also be provided using the `-rna` flag. The `-l` option allows the user to specify an alternative codon table, with the standard codon table being used by default. By default, the ribosome footprint is set to nine codons, although this option can be modified using `-f`. In the same program, we also provide a flag, `-c`, to allow users to use co-tRNA codon pairing instead of identical codon pairing and the `-b` flag to signify both identical and co-tRNA codon pairing. These options are explained in more detail in the accompanying README file found in the GitHub repository: https://github.com/ridgelab/codon_pairing/tree/master/alignment_free.

Parsimony Analysis

We used ortholog annotations from the HUGO Gene Nomenclature Committee (HGNC) (Gray et al., 2015), which unifies gene annotations across species and derives most gene annotations from UniProt (UniProt Consortium, 2018) to determine homologous codon pairing characters. We use Python 3.5 to implement `parsimony_pairing.py` to create a character matrix of parsimony-informative codon pairing usages from a multiple FASTA files containing gene

sequences for each species. Each row in the matrix contains a record for a different species. Each column in the matrix represents a parsimony-informative codon (or amino acid for co-tRNA codon pairing) within a specific ortholog. For each species, each codon (or amino acid) in each ortholog is labelled '0' if it does not pair within a ribosomal window, '1' if it does pair, or '?' if the ortholog annotation is not available for that species. To be parsimony informative, each included ortholog was present in at least four species, each codon (or amino acid) paired in at least one species, and each codon (or amino acid) did not pair in at least one species. We further required all species to contain at least 5% of all the parsimony-informative codons (or amino acids) to limit the effect of missing data. We create this character matrix and as a key file containing an ordered list of each parsimony-informative codon (or amino acid) that was included in the matrix in a single step at runtime (see Figure 2). The following command demonstrates typical usage for identical codon pairing, where `${DIR}` is the path to a directory containing one FASTA file per species, `${MATRIX}` is the path to the output matrix, and `${KEYS}` is the path to the output key file containing the ordered list of parsimony-informative codons.

```
python getPairingMatrix.py -id ${DIR} -o ${MATRIX} -oc ${KEYS}
```

Summary of Parsimony Options

We provide the same options in `parsimony_pairing.py` as the alignment-free method, with a few notable exceptions. In addition to the options described in the alignment-free section, `-oc` optionally indicates the path to an output file containing the ordered parsimony-informative codons included in the character matrix. Optionally, the `-on` option will use a numbering system to create names for the species instead of using the names of the input files. This option is most useful when file names are very long or do not correlate to the species names.

Constructing Phylogenetic Trees Using Parsimony

We used Tree Analysis Using New Technology (TNT) (Goloboff et al., 2005) to recover phylogenetic trees using parsimony. We selected TNT based on its ability to handle large datasets and its fast tree-searching algorithms. We found up to the 100 most parsimonious trees, saving multiple trees recovered using tree bisection reconnection (tbr) branch swapping (Kumar et al., 2018).

Reference Phylogenies

We inferred subtrees from both the OTL and the NCBI Taxonomy Browser for each taxonomic group. The OTL combines phylogenetic relationships reported in primary literature and contains a web application programming interface (API) that allows for querying the OTL database. Although the NCBI Taxonomy Browser gathers information from a variety of sources and is therefore not considered a primary source for taxonomic relationships, it contains more species than the OTL, and provides added insights into our analyses. We use both phylogenies as reference trees to compare the alignment-free and parsimony trees obtained from codon pairing.

Open Tree of Life

We used `getOTLtree.py` from Chapter 4 to obtain reference trees for each taxonomic group from the OTL in a single step at runtime. This program utilizes the OTL API to programmatically query the OTL database to first obtain OTL taxonomy identifiers (OTT ids) for each species, and then query the OTL database to retrieve the reference tree for the species found. The program also allows users to select the correct domain of life when multiple OTT ids are found for a species (e.g., *Nannospalax galili* is currently

listed in the OTL database as both a eukaryote and a bacterium). The output file contains the inferred reference tree from the OTL and a list of any species that the OTL did not include in the tree. We ran this program using the following command, where `${INPUT}` is a list of species, and `${OUTPUT}` is the output file:

```
python getOTLtree.py -i ${INPUT} -o ${OUTPUT}
```

NCBI Taxonomy Browser

We used the NCBI taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) to download the taxonomical relationships in PHYLIP (Felsenstein, 1989) format. We included unranked taxa to maximize the number of included species for each taxonomic group.

Tree Comparisons

We assessed the accuracy of our identical, co-tRNA, and combined codon pairing algorithms by comparing the trees we recovered to the reference trees from the OTL and the NCBI taxonomy. We determined the similarity between trees by using the ete-compare module from the Environment for Tree Exploration toolkit (ETE3) (Huerta-Cepas et al., 2016) , which computes the percentage of branch similarity between two trees. A higher percentage of branch similarity indicates higher congruence between trees. The branch similarity method has a relatively low computational cost for large datasets and it allows for unrooted tree comparisons and comparisons of trees with polytomies. For the parsimony analysis, if any taxonomic comparison produced more than one equally parsimonious tree, we computed the percentage of edge similarity

between each generated tree and the reference tree. We then reported the average percent overlap of all comparisons.

Comparison with Maximum Likelihood

We used the same maximum likelihood validation results as previously reported in Chapter 4. The ortholog-based maximum likelihood technique first compiled all ortholog annotations from the HUGO Gene Nomenclature Committee (HGNC) (Gray et al., 2015) and subsampled the most commonly used orthologs in each taxonomic group, where all gene annotations must be unique within a given species. Next, they used CLUSTAL OMEGA (Sievers and Higgins, 2018) to perform a multiple sequence alignment (MSA) on each orthologous gene cluster. Finally, IQ-TREE (Nguyen et al., 2015) was used to perform a maximum likelihood analysis on the combined MSA super-matrix from all orthologs. ETE3 was used to compare the recovered phylogenies to the OTL and the NCBI taxonomy. The methods presented in Chapter 4 excluded bacteria and viruses from their analysis because of the lack of orthologs spanning a sufficient number of species.

Comparison with Feature Frequency Profiles

Comparisons were also done with a k-mer based alignment-free phylogenomic approach, Feature Frequency Profiles (FFP) (Sims et al., 2009; Jun et al., 2010). The FFP method works by counting shared k-mers between species, with more directly overlapping k-mer counts being associated with closer species relatedness. Since we use the same dataset as previously reported in Chapter 4, we also use their FFP validation set to compare the congruence of FFP with the OTL and the NCBI taxonomy.

Comparison with Codon Aversion Motifs

Codon aversion motifs (CAM) are sets of codons that are not used within genes (see Chapter 4). They have also been used to recover phylogenies using alignment-free techniques. Since our method using codon pairing is also a codon-based method, we included CAM in our comparisons to determine if the phylogenetic signal is stronger in codon use/aversion or codon pairing.

Results

We aimed to determine how phylogenies recovered using identical codon pairing and/or co-tRNA pairing compare to classic methods (i.e., parsimony and maximum likelihood) and other alignment-free methods. First, we determined the theoretical maximum number of character states for each gene using codon pairing in order to determine the maximum number of species we can differentiate using this technique. For identical codon pairing, there are 61 possible pairing combinations (64 codons – 3 stop codons), meaning each gene can separate a maximum of $2^{61} = 2.306 \times 10^{18}$ species. For co-tRNA codon pairing, there are 18 amino acids that use more than one codon, meaning there are 18 possible pairing combinations. Using co-tRNA codon paring, each gene can separate a maximum of $2^{18} = 262{,}144$ species. Using the combined approach, there are 20 possible pairing combinations, one for each of the 20 amino acids. This approach allows each gene to separate a maximum of $2^{20} = 1{,}048{,}576$ species. However, since genes are conserved between species, closely related species share a higher number of codon pairings than more distantly related species. We observed this overlap in codon pairings, with closely related species often having smaller observed distances than distantly related species.

Table 1 shows the number of species that were included in each analysis after the preprocessing filters were applied (e.g., each species in the parsimony analysis included at least 5% of the parsimony-informative characters). In total, we included 23 428 species, with each species generally containing thousands of genes. Supplementary Tables 1-3 show the number of species that were included for each ribosomal window size in the three parsimony analyses. The alignment-free methods included all species because the method is not affected by missing orthologous gene annotations. The NCBI taxonomy contains more taxonomic relationships than the OTL and the OTL does not contain any viruses. Furthermore, the species trees vary between the OTL and the NCBI taxonomy by 1-9%, with the mammal phylogenies being the most similar and the fungi phylogenies being the least similar. Parsimony and maximum likelihood used similar numbers of species in each analysis. A stricter filter was applied to the parsimony analysis than the maximum likelihood analysis, which required the parsimony character matrix to include at least 5% of the parsimony-informative characters. After that filter was applied, we required that at least 5% of the total number of species be included in the analysis (e.g., if 100 species were analyzed, at least 5 species must pass the preprocessing step for the taxonomic group to be included). Applying this filter removed the results from all species, bacteria, and viruses from both the parsimony and maximum likelihood analyses. This filter also removed fungi from the parsimony analysis.

After filtering for parsimony-informative codons, we used parsimony to recover phylogenies with the highest percent overlap based on codon pairings. The identical codon pairing parsimony analysis was based on 794 (invertebrates) to 197 074 (mammals) parsimony-informative codons.

The co-tRNA codon pairing analysis used 382 (invertebrates) to 94 018 (mammals) parsimony-informative codons. The combined codon pairing analysis used 272 (invertebrates) to 72 029 (mammals) parsimony-informative codons. Supplementary Tables 4-6 show the number of informative codons used for each parsimony analysis.

Figure 3 shows the percent overlap of the unrooted trees recovered using the six codon pairing methods (three for parsimony and three for alignment-free) compared to the OTL. For comparison, trees recovered from other alignment-free techniques (CAM and FFP) and maximum likelihood are also compared to the OTL in Figure 3. Figure 4 shows unrooted tree comparisons of the same algorithms compared to the NCBI taxonomy.

In four of the six taxonomic groups where enough species passed the parsimony filters, the parsimony approach for codon pairing recovered phylogenies that were more congruent with the OTL and the NCBI than the alignment-free approach. Parsimony also tied the alignment-free codon pairing approach in protozoa. The only taxonomic group in which the alignment-free method outperformed parsimony was for invertebrates, which also had the fewest parsimony informative characters (See Supplementary Tables 4-6). However, in the three taxonomic groups that were not recovered using maximum likelihood or parsimony, the codon pairing alignment-free approach was more congruent with the established phylogenies than FFP or CAM. The codon pairing alignment-free approach was the most congruent with established phylogenies in all species, bacteria, fungi, protozoa, and viruses. Maximum likelihood was the most congruent with the established phylogenies in four taxonomic groups: archaea, invertebrates, mammals, and

other vertebrates. In plants and protozoa, the codon pairing parsimony approach recovered the most congruent phylogeny with the OTL and the NCBI taxonomy. In archaea, maximum likelihood was the most congruent with the NCBI taxonomy and the parsimony-based approach was the most congruent with the OTL.

Using at least one of the codon pairing techniques, recovered phylogenies were at least 80% congruent with species relationships proposed in the OTL and the NCBI Taxonomy in each of the following taxonomic groups: all species, archaea, bacteria, mammals, and viruses. The alignment-free codon pairing tree recovered over 80% of the unrooted species relationships proposed in the NCBI Taxonomy for plants and other vertebrates. However, the recovered trees were 70-80% congruent with the OTL for plants, and other vertebrates. Recovered unrooted species relationships in fungi and protozoa were greater than 69.2-77.7% identical to both the OTL and the NCBI Taxonomy. Invertebrates had the lowest percent identity, with 65.6% of unrooted edges agreeing with the OTL and 74.8% of species relationships agreeing with the NCBI Taxonomy.

Supplementary Table 7 shows the optimal window sizes and the method (identical or co-tRNA codon pairing) that recovered the most congruent tree with the established phylogenies. We define the minimum optimal window size as the smallest window size to recover the most congruent phylogeny when compared to the reference. Across all taxonomic groups, the minimum optimal ribosome window size was relatively small. Averaged for all minimum optimal window sizes that produced the highest congruence with the OTL, parsimony had a mean minimum optimal window size of 4.000 with a

sample standard deviation of 3.033. The alignment-free method had a mean minimum optimal window size of 3.500 with a sample standard deviation of 1.509. Supplementary Tables 8-19 show the percent edge overlap for identical, co-tRNA, and combined codon pairing compared to the OTL and the NCBI taxonomy for both the alignment-free and parsimony approaches at each ribosome window size from 2-11. For both the alignment-free and parsimony approaches, combining co-tRNA codon pairing with identical codon pairing produced the most congruent tree with the OTL and NCBI Taxonomy in the following taxonomic groups: all species, archaea, bacteria, fungi, invertebrates, protozoa, and viruses. For both methods, identical codon pairing was more congruent with the reference phylogenies in mammals. Parsimony produced a more congruent tree for plants using co-tRNA codon paring, while the alignment-free method preferred identical codon pairing. Furthermore, the non-mammalian vertebrate trees were most congruent with the reference phylogenies using identical codon pairing for the alignment-free method and the combined method using parsimony.

We also compared the codon pairing motifs (i.e., an ordered set of codons that paired within a gene) in each taxonomic group. For example, a gene that has identical codon pairing for AAA and AAT would have a motif of {AAA, AAT}. We found that fewer than 10% of codon pairing motifs were identified in multiple species in most taxonomic groups (see Supplementary Figures 1-10). Bacteria had the most repeated codon pairing motifs (13.7%) and fungi had the fewest repeated motifs (0.7%).

We also determined the frequency of identical codon pairing in genes. We counted the number of genes in a species that used identical codon pairing for each codon. We then calculated the frequency of codon pairing for each codon by dividing the number of genes using codon pairing for that codon by the total number of genes in each species. We repeated this process for each codon, creating boxplots of codon pairing frequencies across each taxonomic group (see Supplementary Figures 11-20). Bacteria, archaea, protozoa, and viruses had very wide distributions of codon pairing frequencies. Fungi and invertebrates had narrower distributions of codon pairing frequencies. Mammals, plants, and other vertebrates had very narrow distributions of codon pairing frequencies. Narrow distributions indicate less variability in codon pairing between species among those taxonomic groups. Each taxonomic group has the same pattern of pairing usage (i.e., if a codon pairs frequently in one taxonomic group, it pairs frequently in other taxonomic groups as well), although mammals have the least variation between species. Excluding stop codons, codons encoding arginine are the least likely to pair (occurs in ~20-25% of genes) and codons encoding asparagine and leucine are the most likely to pair (occurs in ~60-75% of genes), except leucine-encoding CTA, which pairs in ~20-25% of genes.

We further analyzed the number of codons that paired within each gene. We counted the number of codon pairing motifs that included 1, 2, 3,…, 61 codons and report the distribution for each taxonomic group in Supplementary Figures 21-30. In most taxonomic groups, each motif contains ~10-40 codons. However, bacteria, archaea, and viruses are more likely to have fewer codons in each motif, while vertebrates typically have more codons in each motif.

Finally, we quantified the frequency of repeated motifs. We counted the number of times each motif was used in each taxonomic group. Supplementary Figures 31-40 show the distribution of repeated motif frequencies in each taxonomic group. In most taxonomic groups, most repeated motifs are repeated 1-20 times with a steep decreasing slope as the motif is repeated more frequently. However, in archaea, the number of times a motif repeats quickly decreases between 1-30 and then the slope increases until 61 before sharply dropping to near zero. The scripts we used to create each supplementary figure can be found at https://github.com/ridgelab/codon_pairing/supplementary_graphs.

Discussion

Through our analyses, we show that both identical and co-tRNA codon pairing are phylogenetically conserved across all domains of life. We further illustrate that combining identical and co-tRNA codon pairing improves the concordance of recovered phylogenies with the NCBI taxonomy and the OTL in the following taxonomic groups: all species, archaea, bacteria, fungi, invertebrates, protozoa, and viruses. Using parsimony, combining identical and co-tRNA codon pairing also improved the overall concordance of the tree containing non-mammalian vertebrates. In mammals, identical codon pairing had the strongest phylogenetic signal. The most congruent recovered phylogenies for plants were split between only identical codon pairing using the alignment-free method and only co-tRNA codon pairing for the parsimony approach. This comprehensive analysis shows that codon pairing is a novel phylogenetic character state and should be used in conjunction with other phylogenetic techniques in the future.

We also provide tools for quickly analyzing thousands of species using our provided framework. As opposed to common ortholog-based techniques that use sequence divergence to recover phylogenies, identical and co-tRNA codon pairing analyze sequence features that govern gene expression. Since gene expression plays a crucial role in adaptive divergence and ecological speciation (Pavey et al., 2010) and codon pairing affects gene expression, we propose that patterns in codon pairing originated from past speciation events. In our analysis, we show that codon pairing alone can recover phylogenies that are more congruent with the OTL and the NCBI Taxonomy than other alignment-free or maximum likelihood approaches in many instances.

Our analysis of identical codon pairing found several instances of increased (or decreased) codon pairing within certain codons and amino acids. In some instances, codon pairing (or lack of codon pairing) might be due to protein structure instead of translational efficiency. Arginine (Arg) is very positively charged and highly repulsive to other like-charged amino acids. Although rarely pairing compared with other amino acid residues, Arg pairing is essential to some protein-protein interactions and occurs more frequently than expected by random chance (Lee et al., 2013). In protein folding, coiled-coil interfaces often make asparagine (Asn)-Asn conformations that face away from the hydrophobic core (Thomas et al., 2017). Our analysis of codon pairing confirms that Asn pairing occurs much more frequently than Arg pairing. These interactions suggest that Asn and Arg pairing conservation might be based on structure instead of codon translational efficiency. In contrast, leucine zipper T cell receptors have the highest expression values (Foley et al., 2017). Furthermore, the leucine zipper is a 60-80 amino

acid protein domain that allows for faster gene expression, sequence-specific DNA-binding, and dimerization (Ellenberger, 1994). Our results show that leucin-encoding codons are among the most commonly paired codons. However, leucine-encoding CTA pairs significantly less frequently than other leucine-encoding codons. Further exploration into CTA interactions with other leucine-encoding codons may help determine why CTA pairs much less frequently.

Although co-tRNA codon pairing is less prominent in prokaryotes than in eukaryotes (Shao et al., 2012; Zhang et al., 2013; Quax et al., 2015), we show that identical codon pairing and co-tRNA codon pairing are both phylogenetically conserved in all domains of life. However, we also show that the most congruent vertebrate and plant phylogenies are generally recovered using only identical codon pairing using the alignment-free method. Similarly, the parsimony method recovered the most congruent mammal phylogeny using only identical codon pairing. However, parsimony used only co-tRNA codon pairing in plants and used the combined approach in non-mammalian vertebrates. We show that although identical and co-tRNA codon pairing do not occur in equal frequencies, they are both phylogenetically conserved. We also show that combining identical and co-tRNA codon pairing recovers phylogenies that most support established phylogenies in seven out of ten taxonomic groups.

Since many orthologous genes are not currently annotated, our alignment-free approach allows researchers to quickly determine where new genomes fit on the OTL without first verifying orthology. In taxonomic groups that include many recently sequenced genomes,

such as bacteria, fungi, and viruses, the alignment-free approach can provide an accurate method to quickly determine the taxonomic relationships of those species without first annotating orthologs. Furthermore, vastly divergent species can be analyzed with a single command at runtime, facilitating the analysis of thousands of species across various taxonomic groups.

In taxonomic groups that have well-documented orthologous relationships, we show that codon pairing recovers parsimony trees that are largely congruent with the OTL and the NCBI taxonomy. Since maximum likelihood has been widely used to establish the reference phylogenies that we used, it is unsurprising that in the most established taxonomic groups, such as mammals and other vertebrates, maximum likelihood recovers trees that are most congruent with the references. However, in plants and protozoa, the parsimony analysis elucidates a phylogenetic signal in only codon pairing that is sufficient to recover the most congruent trees with the OTL and the NCBI taxonomy. Given the high degree of congruence between the reference phylogenies and the trees recovered using only codon pairing, we propose that codon pairing should be considered in future phylogenomic analyses.

Acknowledgements

We appreciate the financial support of Brigham Young University and the technical assistance of the Fulton Supercomputing Laboratory staff in supporting our research.

References

Cannarozzi, G., Schraudolph, N. N., Faty, M., von Rohr, P., Friberg, M. T., Roth, A. C., Gonnet, P., Gonnet, G., Barral, Y. 2010. A role for codon order in translation dynamics. Cell 141, 355-367.

Chan, C. X., Bernard, G., Poirion, O., Hogan, J. M., Ragan, M. A. 2014. Inferring phylogenies of evolving sequences without multiple sequence alignment. Sci Rep 4, 6504.

Crick, F. 1970. Central dogma of molecular biology. Nature 227, 561-563.

Crick, F. H. 1966. Codon--anticodon pairing: the wobble hypothesis. J Mol Biol 19, 548-555.

Crick, F. H., Barnett, L., Brenner, S., Watts-Tobin, R. J. 1961. General nature of the genetic code for proteins. Nature 192, 1227-1232.

Ellenberger, T. 1994. Getting a grip on DNA recognition: structures of the basic region leucine zipper, and the basic region helix-loop-helix DNA-binding domains. Current Opinion in Structural Biology 4, 12-21.

Farris, J. 1983. The logical basis of phylogenetic analysis.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17, 368-376.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5, 164-166.

Foley, K. C., Spear, T. T., Murray, D. C., Nagato, K., Garrett-Mayer, E., Nishimura, M. I. 2017. HCV T Cell Receptor Chain Modifications to Enhance Expression, Pairing, and Antigen Recognition in T Cells for Adoptive Transfer. Molecular Therapy Oncolytics 5, 105-115.

Goloboff, P. A., Farris, J. S., Nixon, K. C. 2005. TNT: Tree Analysis Using New Technology. 54, 176-178.

Gray, K. A., Yates, B., Seal, R. L., Wright, M. W., Bruford, E. A. 2015. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res 43, D1079-1085.

Gutman, G. A., Hatfield, G. W. 1989. Nonrandom utilization of codon pairs in Escherichia coli. Proc Natl Acad Sci U S A 86, 3699-3703.

Haszprunar, G. 1992. The types of homology and their significance for evolutionary biology and phylogenetics. Journal of Evolutionary Biology 5, 13-24.

Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K. A., Deng, J., Drew, B. T., Gazis, R., Gude, K., Hibbett, D. S., Katz, L. A., Laughinghouse, H. D. t., McTavish, E. J., Midford, P. E., Owen, C. L., Ree, R. H., Rees, J. A., Soltis, D. E., Williams, T., Cranston, K. A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci U S A 112, 12764-12769.

Huerta-Cepas, J., Serra, F., Bork, P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol 33, 1635-1638.

Jun, S.-R., Sims, G. E., Wu, G. A., Kim, S.-H. 2010. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proceedings of the National Academy of Sciences 107, 133-138.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35, 1547-1549.

Lee, D., Lee, J., Seok, C. 2013. What stabilizes close arginine pairing in proteins? Phys Chem Chem Phys 15, 5844-5853.

Martens, A. T., Taylor, J., Hilser, V. J. 2015. Ribosome A and P sites revealed by length analysis of ribosome profiling data. Nucleic Acids Res 43, 3680-3687.

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., Minh, B. Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32, 268-274.

Pavey, S. A., Collin, H., Nosil, P., Rogers, S. M. 2010. The role of gene expression in ecological speciation. Ann N Y Acad Sci 1206, 110-129.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., Baurain, D. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLOS Biology 9, e1000602.

Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H., Dennis, P. P. 1979. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. Proc Natl Acad Sci U S A 76, 1697-1701.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., Tully, R. E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D. R., Murphy, T. D., Ostell, J. M. 2014. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42, D756-763.

Pruitt, K. D., Katz, K. S., Sicotte, H., Maglott, D. R. 2000. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet 16, 44-47.

Quax, T. E., Claassens, N. J., Soll, D., van der Oost, J. 2015. Codon Bias as a Means to Fine-Tune Gene Expression. Mol Cell 59, 149-161.

Saitou, N., Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4, 406-425.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Fingerman, I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Krasnov, S., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Karsch-Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E., Ye, J. 2012. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 40, D13-25.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Federhen, S., Feolo, M., Fingerman, I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E., Ye, J. 2011. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 39, D38-51.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., John Wilbur, W.,

Yaschenko, E., Ye, J. 2010. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 38, D5-16.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E., Ye, J. 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 37, D5-15.

Shao, Z. Q., Zhang, Y. M., Feng, X. Y., Wang, B., Chen, J. Q. 2012. Synonymous codon ordering: a subtle but prevalent strategy of bacteria to improve translational efficiency. PLoS One 7, e33547.

Sharp, P. M., Li, W. H. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24, 28-38.

Sievers, F., Higgins, D. G. 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci 27, 135-145.

Sims, G. E., Jun, S. R., Wu, G. A., Kim, S. H. 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci U S A 106, 2677-2682.

Soltis, D. E., Soltis, P. S. 2003. The Role of Phylogenetics in Comparative Genetics. Plant Physiology 132, 1790-1800.

Thomas, F., Niitsu, A., Oregioni, A., Bartlett, G. J., Woolfson, D. N. 2017. Conformational Dynamics of Asparagine at Coiled-Coil Interfaces. Biochemistry 56, 6544-6554.

UniProt Consortium, T. 2018. UniProt: the universal protein knowledgebase. Nucleic Acids Res 46, 2699.

Vinga, S., Almeida, J. 2003. Alignment-free sequence comparison—a review. Bioinformatics 19, 513-523.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., Yaschenko, E. 2007. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35, D5-12.

Wilgenbusch, J. C., Swofford, D. 2003. Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics Chapter 6, Unit 6 4.

Yang, Z., Rannala, B. 2012. Molecular phylogenetics: principles and practice. Nat Rev Genet 13, 303-314.

Zhang, Y. M., Shao, Z. Q., Yang, L. T., Sun, X. Q., Mao, Y. F., Chen, J. Q., Wang, B. 2013. Non-random arrangement of synonymous codons in archaea coding sequences. Genomics 101, 362-367.

Tables and Figures

Chapter 5 Tables

Table 5.1. Number of Species Passing Preprocessing Filters and Analyzed by Each Algorithm

Taxonomic Group	Alignment-free	Parsimony	Maximum Likelihood	NCBI Taxonomy	OTL
All	23 428	0	0	22 794	12 337
Archaea	418	100	418	416	362
Bacteria*	15 068	0	0	14 612	11 227
Fungi	234	0	58	234	214
Invertebrates	149	57	57	149	147
Plants	89	61	60	89	87
Protozoa	75	15	24	75	75
Mammals	107	97	100	107	105
Other vertebrates	123	114	118	123	120
Viruses*	7 233	0	0	7 045	0

The alignment-free methods include codon pairing, CAM, and FFP, and did not require any preprocessing. Parsimony used a stricter preprocessing cutoff than maximum likelihood, and therefore used fewer species. The NCBI taxonomy includes viruses and more species than the OTL. Zero species passed the filters when fewer than 5% of the total species had sufficient ortholog annotations to run the analysis.

Chapter 5 Figures

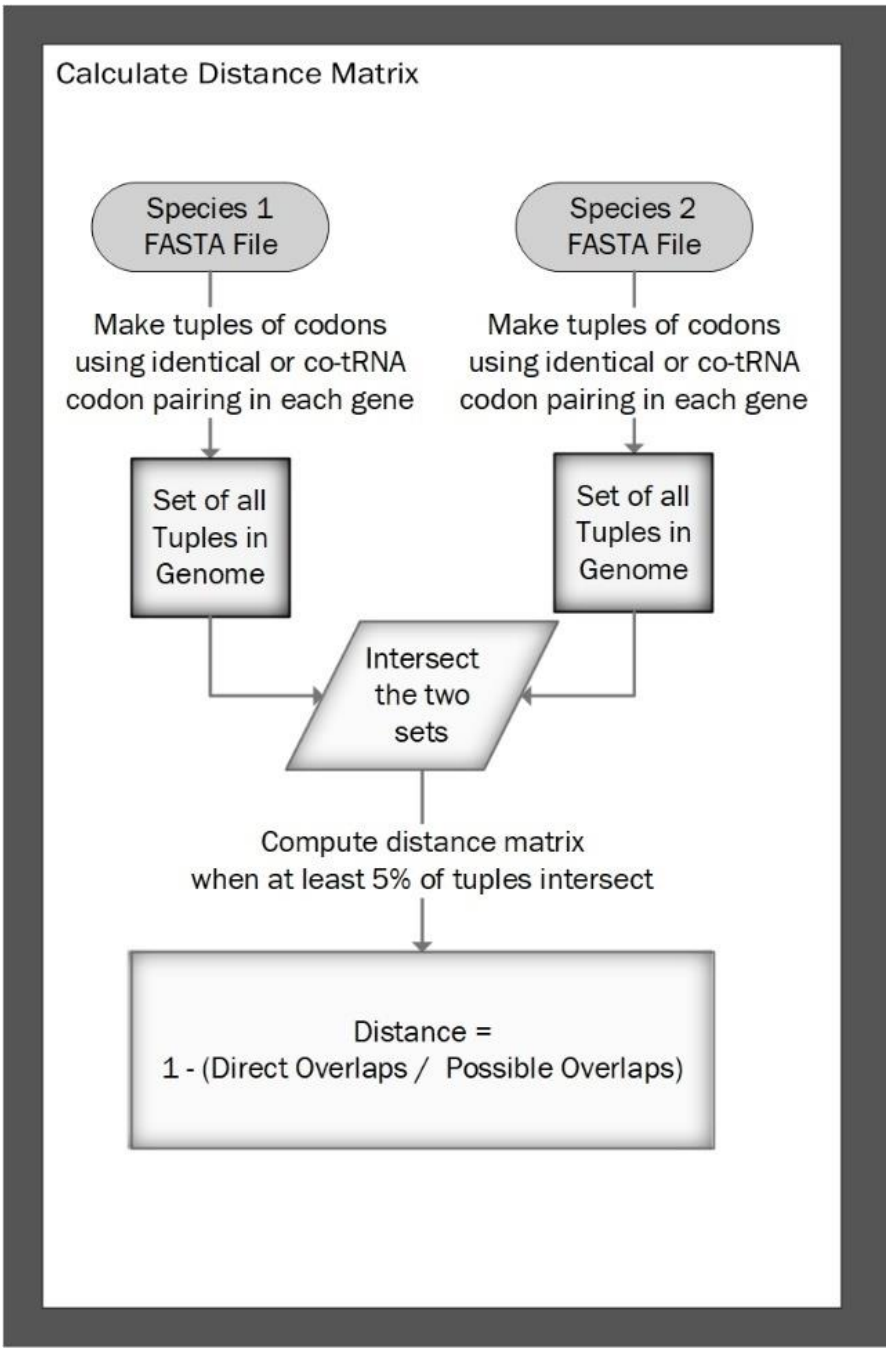


Figure 5.1. Process to Calculate the Distance Matrix Based on Identical Codon Pairing
Starting with the coding sequences of each gene in a species (FASTA file), codons that use codon pairing within the ribosomal footprint are included in a tuple that is then added to a set for that species. Sets of tuples are intersected to calculate the distance between species. These distances are then added to a distance matrix that can be used to recover phylogenies. Similarly, co-tRNA codon pairing and the combined methods are calculated by using sets of amino acid tuples instead of codon tuples.

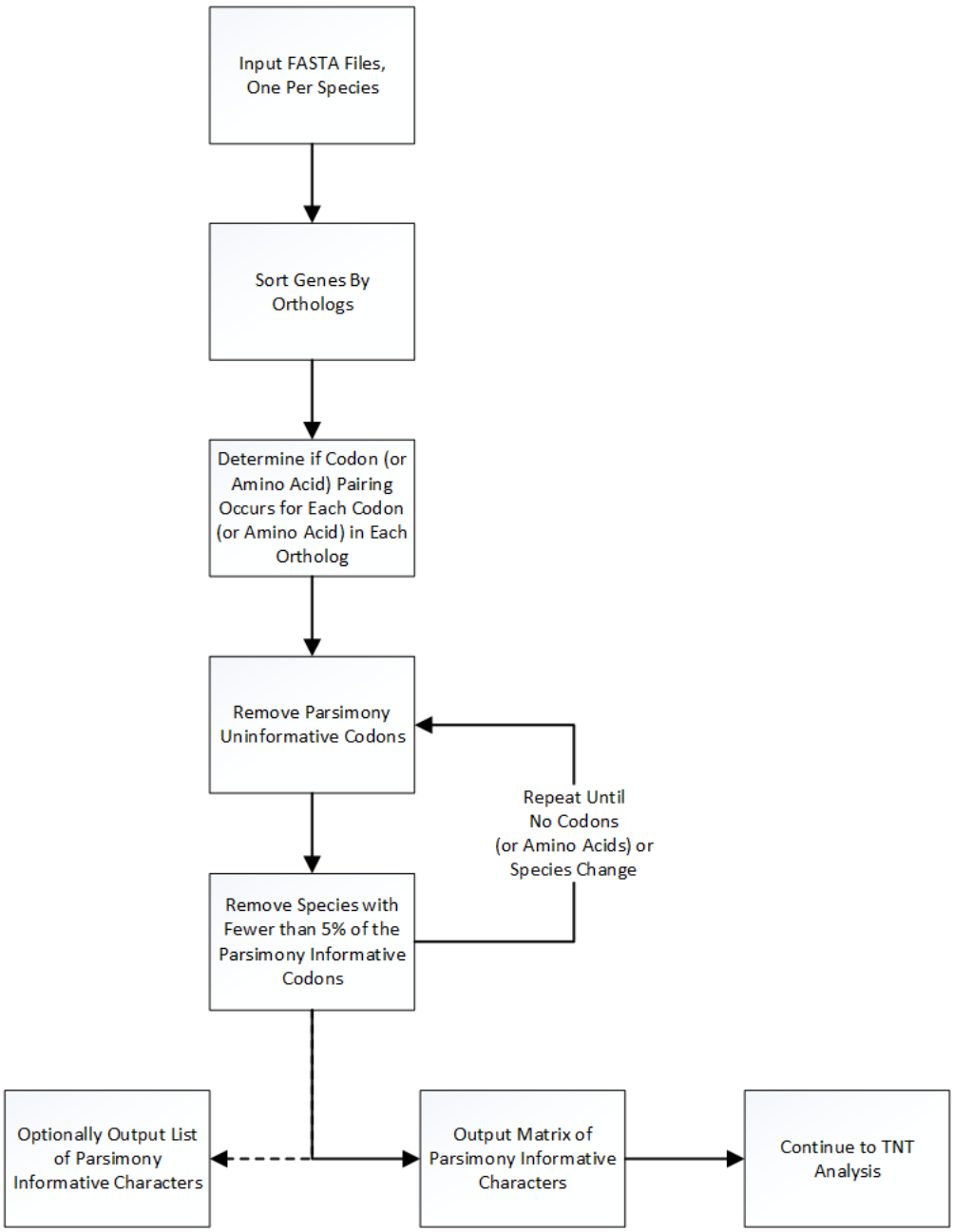
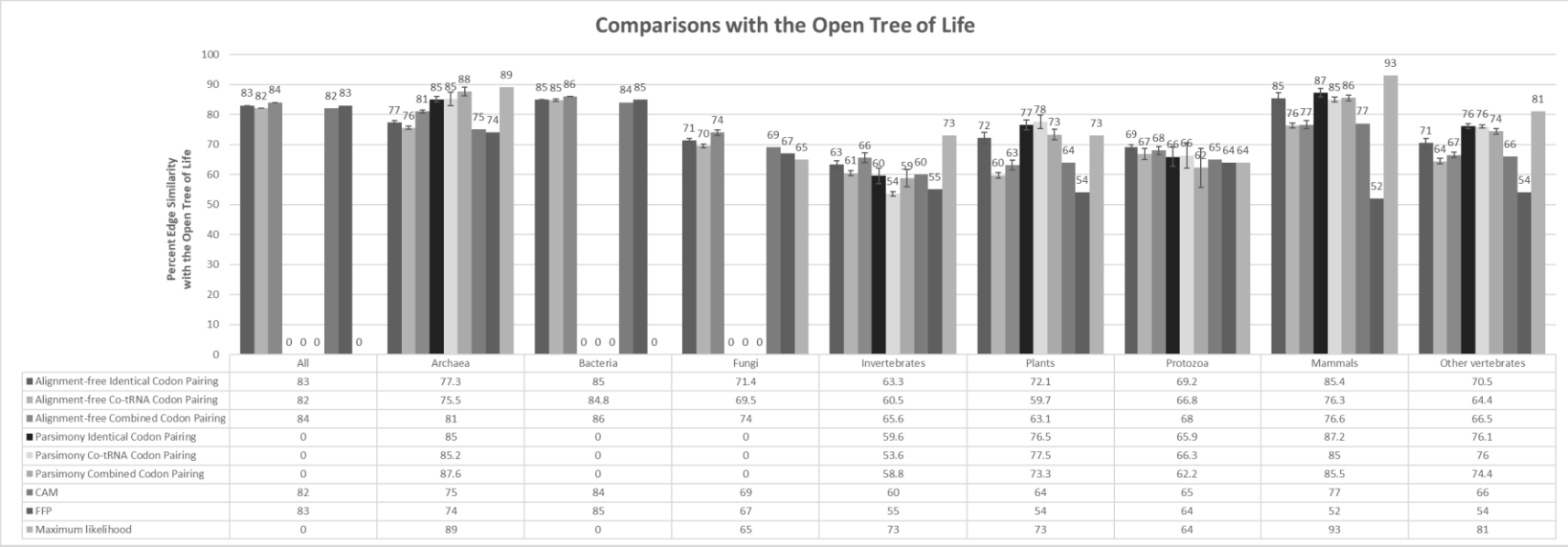
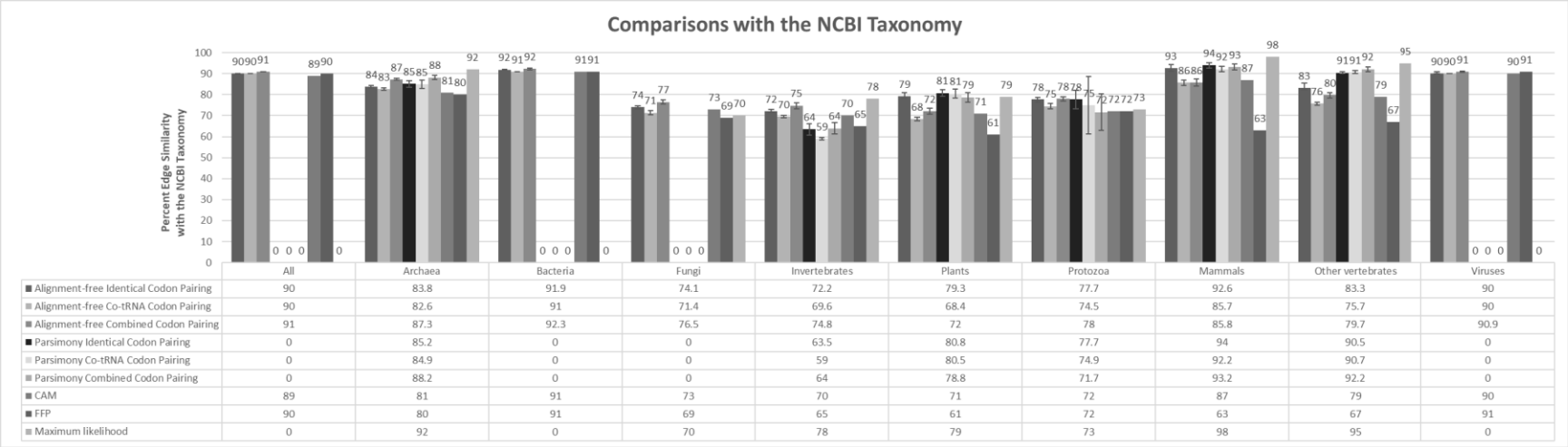


Figure 5.2. Flow Chart for the Parsimony Analysis We start with input FASTA files, one for each species. For each codon (or amino acid) within each ortholog, we assign a binary value of '0', '1', or '?' depending on if codon pairing for that codon (or amino acid) occurs. We then remove parsimony-uninformative characters. We then remove any species that do not contain at least 5% of the parsimony informative codons and we conduct the analysis only if at least 5% of the species pass the filter. Finally, we output the parsimony-informative character matrix for each codon (or amino acid) pairing to be used in a TNT analysis and an optional list of parsimony informative characters.



	All	Archaea	Bacteria	Fungi	Invertebrates	Plants	Protozoa	Mammals	Other vertebrates
Alignment-free Identical Codon Pairing	83	77.3	85	71.4	63.3	72.1	69.2	85.4	70.5
Alignment-free Co-tRNA Codon Pairing	82	75.5	84.8	69.5	60.5	59.7	66.8	76.3	64.4
Alignment-free Combined Codon Pairing	84	81	86	74	65.6	63.1	68	76.6	66.5
Parsimony Identical Codon Pairing	0	85	0	0	59.6	76.5	65.9	87.2	76.1
Parsimony Co-tRNA Codon Pairing	0	85.2	0	0	53.6	77.5	66.3	85	76
Parsimony Combined Codon Pairing	0	87.6	0	0	58.8	73.3	62.2	85.5	74.4
CAM	82	75	84	69	60	64	65	77	66
FFP	83	74	85	67	55	54	64	52	54
Maximum likelihood	0	89	0	65	73	73	64	93	81

Figure 5.3. Percent Edge Overlap for Comparisons of Each Algorithm Against the OTL The alignment-free and parsimony codon pairing methods report the mean percent edge overlap with the OTL based on using different ribosome windows from 2-11. Error bars are reported for the codon pairing methods, signifying one standard deviation from the mean. The other methods were previously reported in Chapter 4 and are used for comparison.



	All	Archaea	Bacteria	Fungi	Invertebrates	Plants	Protozoa	Mammals	Other vertebrates	Viruses
Alignment-free Identical Codon Pairing	90	83.8	91.9	74.1	72.2	79.3	77.7	92.6	83.3	90
Alignment-free Co-tRNA Codon Pairing	90	82.6	91	71.4	69.6	68.4	74.5	85.7	75.7	90
Alignment-free Combined Codon Pairing	91	87.3	92.3	76.5	74.8	72	78	85.8	79.7	90.9
Parsimony Identical Codon Pairing	0	85.2	0	0	63.5	80.8	77.7	94	90.5	0
Parsimony Co-tRNA Codon Pairing	0	84.9	0	0	59	80.5	74.9	92.2	90.7	0
Parsimony Combined Codon Pairing	0	88.2	0	0	64	78.8	71.7	93.2	92.2	0
CAM	89	81	91	73	70	71	72	87	79	90
FFP	90	80	91	69	65	61	72	63	67	91
Maximum likelihood	0	92	0	70	78	79	73	98	95	0

Figure 5.4. Percent Edge Overlap for Comparisons of Each Algorithm Against the NCBI Taxonomy The alignment-free and parsimony codon pairing methods report the mean percent edge overlap with the NCBI taxonomy based on using different ribosome windows from 2-11. Error bars are reported for the codon pairing methods, signifying one standard deviation from the mean. The other methods were previously reported in Chapter 4 and are used for comparison.

Manuscript published in *Biomedical Genetics and Genomics*. 2017; DOI:

10.15761/BGG.1000134

Chapter 6

Human Viruses Have Codon Usage Biases That Match Highly Expressed Proteins in the Tissues They Infect

Justin B. Miller[1], Ariel A. Hippen[1], Sage M. Wright[1], Caroline Morris[1], and Perry G. Ridge[1]

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

Abstract

It is well-documented that codon usage biases affect gene translational efficiency; however, it is less known if viruses share their host's codon usage motifs. We determined that human-infecting viruses share similar codon usage biases as proteins that are expressed in tissues the viruses infect. By performing 7,052,621 pairwise comparisons of genes from humans versus genes from 113 viruses that infect humans, we determined which codon usage motifs were most highly correlated. We found that 16 viruses averaged a significant correlation in codon usage with over 500 human genes per viral gene, 58 viruses were highly correlated with an average of at least 100 human genes per viral gene, and 37 viruses were significantly correlated with an average of at least one human gene per viral gene at an alpha level of 7.09 x (0.05 alpha / 7,052,621 comparisons). Only two viruses were not highly correlated with an average of one human gene per viral gene. While relatively few of the interactions were previously documented, the high statistical correlations suggest that researchers may be able to determine which tissues a virus is most likely to infect by analyzing codon usage biases.

Key Words: [codon usage bias, host, human, virus, virus-host interactions]

Introduction

Amino acids are encoded by DNA triplets known as codons; however, since there are only 20 canonical amino acids and 64 possible codons, multiple codons encode a single amino acid (Crick, 1968). The majority of amino acids are encoded by 2-6 different codons. Despite multiple codons encoding a single amino acid, codon usage is not random in most species (Ikemura, 1985, Sharp and Li, 1986, Gutman and Hatfield, 1989, Zhang et al., 2013). Various species, including many plant species, *E. coli* and *Drosophila*, also maintain DNA triplet preferences, or codon usage biases, over time in both intronic and exonic regions (Akashi et al., 2007, Yang and Nielsen, 2008, Xu et al., 2015).

It is generally accepted that non-random mutations occur more frequently at the third position in the codon, and codon bias persists through selection (Hershberg and Petrov, 2008, Quax et al., 2015). Numerous biological factors create evolutionary pressure to use certain codons. First, an incomplete set of transfer RNAs (tRNAs) or unequal expression of tRNA anticodons within a tissue or species creates pressure for codons with complementary tRNAs available. Second, translational speed may either increase or decrease depending on the codon used, creating pressure to select codons for which translational efficiency matches the needs of the tissue/cell (i.e. suboptimal codons might be preferential to some species for increased translational efficiency, while in other instances suboptimal codons might decrease translational efficiency) (Quax et al., 2015, Xu et al., 2013). Finally, codon usage bias primarily affects the translation of a gene and is a main determinant of gene expression (Zhou et al., 2016).

Recently, significant correlations for codon usage preferences between RNA viruses (e.g. SBV and KV) and their host, the honeybee, were reported (Chantawannakul and Cutler, 2008). They proposed that such similarities resulted from co-evolution, which typically occurs in a leapfrog fashion (i.e. as the host evolves to combat the parasite, the parasite evolves to adapt to the new conditions).

We aimed to determine whether the same relationship exists between human and viral genes expressed in tissues targeted by the virus. We analyzed 19,482 human proteins, and compared their codon usage biases against 113 viruses that infect human hosts. We found significant correlations for many viral and human proteins, and where tissue information was available, the top correlated human protein was frequently highly expressed in the tissue type targeted by the virus.

Materials and Methods

Data Collection and Cleaning

We used gene annotations from the General Feature Format (GFF) and GFF3 files from the National Center for Biotechnology Information (NCBI) to extract the reference viral and human sequences (Pruitt et al., 2014, Tatusova et al., 2014, Wheeler et al., 2007). Since the reference genome is intended to most accurately represent an average individual in a species, we downloaded all reference sequence data, including the corresponding gene annotations, from NCBI. Similar to the methods used by (Camiolo et al., 2015), when multiple isoforms were annotated, the longest isoform was always chosen as the representative isoform for that gene, and we removed all genes with any annotated

translational exceptions (e.g., translational, unclassified transcription discrepancy, suspected errors, etc.). These filters had only a minor effect on our data because they eliminated less than 5% of the total sequences. All 19,482 sequence accession numbers can be found in the NCBI database by downloading the complete genome annotations for Homo sapiens; the accession numbers for each virus and their highest correlating genes are located in Supplementary Table 1.

Codon Usage Correlation Values

To determine if there was a correlation between human and viral codon usage biases, we performed a Pearson's r correlation test with discrete codon usage counts by comparing total codon usage counts in human and viral coding sequences (CDS). We used Pearson's r because it uses a product-moment correlation coefficient that is used to determine the correlation between two variables with different units or different magnitudes (Hane et al., 1993). Since gene lengths can vary greatly between genes, and genes do not contain all codons, the assumptions for most statistical tools would not be adequately met using the raw data. Furthermore, the high number of zero codon usage counts in some genes meant that a percentage comparison of codon usages using a traditional t-test was unfeasible, even with a transformation. We chose an implementation of Pearson's r from the package SciPy in Python version 2.7 because Pearson's r is robust to variations in sequence sizes as well as zero values. Using Pearson's r, we graphed a linear regression and calculated the R^2 coefficients of determination and p-values by plotting the discrete codon counts from each gene within each virus against each human gene. Next, we ranked the correlation of codon usage between viral and human genes from highest to lowest. We corrected for multiple

tests using a Bonferroni correction; the significance threshold used was 7.09×10^{-9} (0.05/7,052,621 total comparisons). We obtained the highest correlations when the viral and human protein codon usage motifs were most similar.

Human Tissue Comparisons

We determined which proteins were expressed in each human tissue by querying each highly correlated human protein against the Human Protein Atlas (Uhlen et al., 2005, Uhlen et al., 2015). We checked the top correlating human proteins for each virus (113 total proteins) to determine in which tissues they were most highly expressed. While many proteins were expressed in low levels throughout the body, we were most concerned with high expression areas, and only the high expression areas were compared in this study.

Results

Of the 113 viruses analyzed, we found that on average, each viral gene in 16 viruses was significantly correlated with more than 500 human proteins (see Supplementary Table 2). Of the remaining 97 viruses, 58 were significantly correlated with at least 100 human proteins per viral gene, and 37 were significantly correlated with at least one human gene per viral gene on average at a p-value $< 7.09 \times 10^{-9}$. Only two viruses, Human papillomavirus type 90 (NC_004104) and Human gyrovirus type 1 (NC_015630) were not significantly correlated with the codon usage of at least one human gene per viral gene, on average.

The viruses listed in Table 1 have the highest Pearson r correlation of all comparisons, with their codon usages strongly correlating to their host codon usages (p-value $< 10^{-25}$). Four of

the top 10 correlations in Table 1 belong to the group of 16 viruses that strongly correlate to over 500 human proteins per viral gene on average, and the rest of them belong to the group of 58 with significant correlations with at least 100 human genes significantly correlating to each viral gene, on average. Overall, the average correlation of the 113 viruses with the top hit from each virus was 83.1%, meaning about 83% of the codon usage bias in the virus also existed in the human host protein. Each viral protein strongly correlated to an average of 303 human genes.

To demonstrate the strong correlations in codon usage bias, we plotted codon usage for several representative viral proteins compared to the human protein with the strongest correlation (see Figure 1).

Finally, we analyzed the correlations of codon usage biases for human proteins expressed in tissues infected by a specific virus. With the exception of sexually transmitted diseases (STDs), tissue information was incomplete for many viruses, and further exacerbating this problem is that many human proteins expressed in a specific tissue were also expressed in many other tissues. We report all known tissue information in Supplementary Table 3, and in Table 2 list representative viruses with their highest correlating protein and affected tissues.

Discussion

The high number of proteins significantly correlated with each virus suggests that humans and human-host viruses share similar codon usage biases. For example, each of the 80 Human herpesvirus 4 (HHV-4, NC_009334) genes significantly correlated with 1 to 10,012 human genes with a median of 8,290 highly correlated human genes and an average of 1,036

highly correlated human genes. HHV-4 was previously identified as having a similar codon usage bias to its host cells (Roychoudhury and Mukherjee, 2010, Virgin et al., 2009), which may provide insights into the efficient proliferation of HHV-4, since it can more readily utilize host tRNA machinery in the tissue types it infects. Indeed, HHV-4 (commonly known as mononucleosis or "the kissing disease") is one of the most common viruses known to infect humans, with almost 90% of adults having antibodies suggesting previous HHV-4 infection (Virgin et al., 2009). Herpesviruses overtake host translational machinery through virion host shutoff (vhs), which limits the expression of host mRNA (Smiley, 2004), and through the degradation of host mitochondrial DNA (Saffran et al., 2007), although some herpesvirus strains act differently (Duguay et al., 2014). Our data suggest that herpesvirus is able to co-opt the translational apparatus of the infected cell by closely matching codon usage biases. The virus is able to use existing tRNAs in the cell, which are not being used by the cell due to vhs.

Furthermore, viruses such as HPV-90 (NC_004104) and Human gyrovirus 1 (NC_015630) with fewer correlating proteins typically occur less frequently in human populations. Although limited data exist for the prevalence of HPV-90 in the general population, in general it presents a very low risk to the general population (Schmitt et al., 2013, Quiroga-Garza et al., 2013). Human gyrovirus 1, which is identical to the Chicken Anemia Virus, is relatively rare and the effects of the virus still remain largely unknown, although it may affect the apoptosis pathway (Sauvage et al., 2011, Chaabane et al., 2014).

Human-host viruses appear to target tissues where the correlating human protein also has high expression. Although many viruses analyzed were not clearly annotated as infecting a particular human tissue, the viruses with documented tissue interactions were always highly correlated with a protein that was highly expressed in that tissue. For instance, HPV-128 correlates most with the human protein TIGD4, which is mainly expressed in the genitalia. In addition, other STDs were strongly correlated with proteins that were also mainly expressed in genitalia (see Table 2, Supplementary Table 3). We note that viruses tend to share the same codon usage biases as at least one protein that is highly expressed in the disease targeted area, further emphasizing our conclusion that viral and host codon usage biases are highly correlated.

Highly expressed genes have codon biases that utilize highly abundant tRNAs in order for optimal translational and transcriptional speed (Zhou et al., 2016, Chantawannakul and Cutler, 2008, Grosjean and Fiers, 1982, Morton, 1998, Morton and So, 2000, Merkl, 2003). The Human Adenovirus E (NP_009115.2), which causes respiratory illness, has an 89.9% codon usage correlation with the NISCH gene, which is mainly expressed in the bronchus. Since NISCH is highly expressed in the tissues that the adenovirus normally infects, the virus is able to take advantage of its codon usage bias similarities with the host proteins to rapidly proliferate and infect additional hosts.

There are other possibilities for the observed shared codon usage biases. For example, co-evolution may have contributed to the appearance of such strong codon bias correlations, in which the host and the virus evolve at similar rates in order to either combat or maintain

parasitic infection (Parrish et al., 2008). Since viruses have smaller genomes, they can selectively evolve more rapidly toward being similar to a preferred host.

While co-evolution and the abundance of optimal tRNAs are thought to allow greater viral spread, determining the exact cause of this correlation remains unexplored. Our extensive analysis of codon usage determined that a strong correlation in codon usage bias exists between human-host viruses and proteins expressed in the human tissues that they infect. Future research should focus on the causes of these correlations.

Authorship and Contributorship

JM and PR conceived the idea. JM oversaw all aspects of the project. AH developed the comparison algorithms and ran the comparisons. CM and SW conducted literature searches and wrote sections of the paper. JM and PR were primarily responsible for editing the manuscript. PR mentored the project.

Acknowledgements

We also appreciate Mark Ebbert and Samantha Jensen who provided expert suggestions for the project flow and design.

Funding Information

We appreciate the contributions of Brigham Young University and the Fulton Supercomputing Laboratory in supporting our research.

Competing Interests

The authors declare that they have no competing interests.

Availability of Data and Material

All data are freely available from the NCBI database at ftp://ftp.ncbi.nlm.nih.gov/

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.

CAMIOLO, S., MELITO, S. & PORCEDDU, A. 2015. New insights into the interplay between codon bias determinants in plants. *DNA Res,* 22**,** 461-70.

CHAABANE, W., CIESLAR-POBUDA, A., EL-GAZZAH, M., JAIN, M. V., RZESZOWSKA-WOLNY, J., RAFAT, M., STETEFELD, J., GHAVAMI, S. & LOS, M. J. 2014. Human-gyrovirus-Apoptin triggers mitochondrial death pathway--Nur77 is required for apoptosis triggering. *Neoplasia,* 16**,** 679-93.

CHANTAWANNAKUL, P. & CUTLER, R. W. 2008. Convergent host-parasite codon usage between honeybee and bee associated viral genomes. *J Invertebr Pathol,* 98**,** 206-10.

CRICK, F. H. 1968. The origin of the genetic code. *J Mol Biol,* 38**,** 367-79.

DUGUAY, B. A., SAFFRAN, H. A., PONOMAREV, A., DULEY, S. A., EATON, H. E. & SMILEY, J. R. 2014. Elimination of mitochondrial DNA is not required for herpes simplex virus 1 replication. *J Virol,* 88**,** 2967-76.

GROSJEAN, H. & FIERS, W. 1982. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene,* 18**,** 199-209.

GUTMAN, G. A. & HATFIELD, G. W. 1989. Nonrandom utilization of codon pairs in Escherichia coli. *Proc Natl Acad Sci U S A,* 86**,** 3699-703.

HANE, B. G., JAGER, K. & DREXLER, H. G. 1993. The Pearson product-moment correlation coefficient is better suited for identification of DNA fingerprint profiles than band matching algorithms. *Electrophoresis,* 14**,** 967-72.

HERSHBERG, R. & PETROV, D. A. 2008. Selection on codon bias. *Annu Rev Genet,* 42**,** 287-99.

IKEMURA, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol,* 2**,** 13-34.

MERKL, R. 2003. A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency. *J Mol Evol,* 57**,** 453-66.

MORTON, B. R. 1998. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *J Mol Evol,* 46**,** 449-59.

MORTON, B. R. & SO, B. G. 2000. Codon usage in plastid genes is correlated with context, position within the gene, and amino acid content. *J Mol Evol,* 50**,** 184-93.

PARRISH, C. R., HOLMES, E. C., MORENS, D. M., PARK, E. C., BURKE, D. S., CALISHER, C. H., LAUGHLIN, C. A., SAIF, L. J. & DASZAK, P. 2008. Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol Mol Biol Rev,* 72**,** 457-70.

PRUITT, K. D., BROWN, G. R., HIATT, S. M., THIBAUD-NISSEN, F., ASTASHYN, A., ERMOLAEVA, O., FARRELL, C. M., HART, J., LANDRUM, M. J., MCGARVEY, K. M., MURPHY, M. R., O'LEARY, N. A., PUJAR, S., RAJPUT, B., RANGWALA, S. H., RIDDICK, L. D., SHKEDA, A., SUN, H., TAMEZ, P., TULLY, R. E., WALLIN, C., WEBB, D., WEBER, J., WU, W., DICUCCIO, M., KITTS, P., MAGLOTT, D. R., MURPHY, T. D. & OSTELL, J. M. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res,* 42**,** D756-63.

QUAX, T. E., CLAASSENS, N. J., SOLL, D. & VAN DER OOST, J. 2015. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell,* 59**,** 149-61.

QUIROGA-GARZA, G., ZHOU, H., MODY, D. R., SCHWARTZ, M. R. & GE, Y. 2013. Unexpected high prevalence of HPV 90 infection in an underserved population: is it really a low-risk genotype? *Arch Pathol Lab Med,* 137**,** 1569-73.

ROYCHOUDHURY, S. & MUKHERJEE, D. 2010. A detailed comparative analysis on the overall codon usage pattern in herpesviruses. *Virus Res,* 148**,** 31-43.

SAFFRAN, H. A., PARE, J. M., CORCORAN, J. A., WELLER, S. K. & SMILEY, J. R. 2007. Herpes simplex virus eliminates host mitochondrial DNA. *EMBO Rep,* 8**,** 188-93.

SAUVAGE, V., CHEVAL, J., FOULONGNE, V., GOUILH, M. A., PARIENTE, K., MANUGUERRA, J. C., RICHARDSON, J., DEREURE, O., LECUIT, M., BURGUIERE, A., CARO, V. & ELOIT, M. 2011. Identification of the first human gyrovirus, a virus related to chicken anemia virus. *J Virol,* 85**,** 7948-50.

SCHMITT, M., DEPUYDT, C., BENOY, I., BOGERS, J., ANTOINE, J., ARBYN, M., PAWLITA, M. & GROUP, V. S. 2013. Prevalence and viral load of 51 genital human papillomavirus types and three subtypes. *Int J Cancer,* 132**,** 2395-403.

SHARP, P. M. & LI, W. H. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol,* 24**,** 28-38.

SMILEY, J. R. 2004. Herpes simplex virus virion host shutoff protein: immune evasion mediated by a viral RNase? *J Virol,* 78**,** 1063-8.

TATUSOVA, T., CIUFO, S., FEDOROV, B., O'NEILL, K. & TOLSTOY, I. 2014. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res,* 42**,** D553-9.

UHLEN, M., BJORLING, E., AGATON, C., SZIGYARTO, C. A., AMINI, B., ANDERSEN, E., ANDERSSON, A. C., ANGELIDOU, P., ASPLUND, A., ASPLUND, C., BERGLUND, L., BERGSTROM, K., BRUMER, H., CERJAN, D., EKSTROM, M., ELOBEID, A., ERIKSSON, C., FAGERBERG, L., FALK, R., FALL, J., FORSBERG, M., BJORKLUND, M. G., GUMBEL, K., HALIMI, A., HALLIN, I., HAMSTEN, C., HANSSON, M., HEDHAMMAR, M., HERCULES, G., KAMPF, C., LARSSON, K., LINDSKOG, M., LODEWYCKX, W., LUND, J., LUNDEBERG, J., MAGNUSSON, K., MALM, E., NILSSON, P., ODLING, J., OKSVOLD, P., OLSSON, I., OSTER, E., OTTOSSON, J., PAAVILAINEN, L., PERSSON, A., RIMINI, R., ROCKBERG, J., RUNESON, M., SIVERTSSON, A., SKOLLERMO, A., STEEN, J., STENVALL, M., STERKY, F., STROMBERG, S., SUNDBERG, M., TEGEL, H., TOURLE, S., WAHLUND, E., WALDEN, A., WAN, J., WERNERUS, H., WESTBERG, J., WESTER, K., WRETHAGEN, U., XU, L. L., HOBER, S. & PONTEN, F. 2005. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics,* 4**,** 1920-32.

UHLEN, M., FAGERBERG, L., HALLSTROM, B. M., LINDSKOG, C., OKSVOLD, P., MARDINOGLU, A., SIVERTSSON, A., KAMPF, C., SJOSTEDT, E., ASPLUND, A., OLSSON, I., EDLUND, K., LUNDBERG, E., NAVANI, S., SZIGYARTO, C. A., ODEBERG, J., DJUREINOVIC, D., TAKANEN, J. O., HOBER, S., ALM, T., EDQVIST, P. H., BERLING, H., TEGEL, H., MULDER, J., ROCKBERG, J., NILSSON, P., SCHWENK, J. M., HAMSTEN, M., VON FEILITZEN, K., FORSBERG, M., PERSSON, L., JOHANSSON, F., ZWAHLEN, M., VON HEIJNE, G., NIELSEN, J. & PONTEN, F. 2015. Proteomics. Tissue-based map of the human proteome. *Science,* 347**,** 1260419.

VIRGIN, H. W., WHERRY, E. J. & AHMED, R. 2009. Redefining chronic viral infection. *Cell,* 138**,** 30-50.

WHEELER, D. L., BARRETT, T., BENSON, D. A., BRYANT, S. H., CANESE, K., CHETVERNIN, V., CHURCH, D. M., DICUCCIO, M., EDGAR, R., FEDERHEN, S., GEER, L. Y., KAPUSTIN, Y., KHOVAYKO, O., LANDSMAN, D., LIPMAN, D. J., MADDEN, T. L., MAGLOTT, D. R., OSTELL, J., MILLER, V., PRUITT, K. D., SCHULER, G. D., SEQUEIRA, E., SHERRY, S. T., SIROTKIN, K., SOUVOROV, A., STARCHENKO, G., TATUSOV, R. L., TATUSOVA, T. A., WAGNER, L. & YASCHENKO, E. 2007. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res,* 35**,** D5-12.

XU, W., XING, T., ZHAO, M., YIN, X., XIA, G. & WANG, M. 2015. Synonymous codon usage bias in plant mitochondrial genes is associated with intron number and mirrors species evolution. *PLoS One,* 10**,** e0131508.

XU, Y., MA, P., SHAH, P., ROKAS, A., LIU, Y. & JOHNSON, C. H. 2013. Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. *Nature,* 495**,** 116-20.

YANG, Z. & NIELSEN, R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol,* 25.

ZHANG, Y. M., SHAO, Z. Q., YANG, L. T., SUN, X. Q., MAO, Y. F., CHEN, J. Q. & WANG, B. 2013. Non-random arrangement of synonymous codons in archaea coding sequences. *Genomics,* 101**,** 362-7.

ZHOU, Z., DANG, Y., ZHOU, M., LI, L., YU, C. H., FU, J., CHEN, S. & LIU, Y. 2016. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci U S A,* 113**,** E6117-E6125.

Tables and Figures

Chapter 6 Tables

Table 6.1. Top 10 Codon Usage Bias Correlations

Virus Accession Number	Virus Name	Virus Protein Name	Protein Accession Number	Protein Name	Correlation %	P-value
NC_009334	Human herpesvirus 4	BALF5	NP_620124.1	RHOT2	93.6	8.64E-30
NC_007605	Human herpesvirus 4 (wild	BALF5	NP_620124.1	RHOT2	93.5	1.36E-29
NC_000898	Human herpesvirus 6B	U90	NP_112561.2	TEX15	93.1	6.40E-29
NC_014185	Human papillomavirus 121	E1	NP_940841.1	KBTBD3	92.8	2.53E-28
NC_001716	Human herpesvirus 7	IE1	NP_001073973.2	RBM44	92.8	3.03E-28
NC_016157	Human papillomavirus 126	Pos: 817-2640	NP_940841.1	KBTBD3	92.0	6.78E-27
NC_009333	Human herpesvirus 8	ORF75	NP_002891.1	RBP3	91.8	1.47E-26
NC_010329	Human papillomavirus 88	E1	NP_940841.1	KBTBD3	90.8	4.10E-25
NC_001806	Human herpesvirus 1	UL30	NP_055778.2	SBNO2	90.8	4.15E-25
NC_014955	Human papillomavirus 132	E1	NP_940841.1	KBTBD3	90.5	9.67E-25

Here we report the top-ten codon usage bias correlations (Pearson's r values) between a virus and a human protein with their respective p-values (all under 10^{-25}), demonstrating that viruses and proteins in their host (humans) share high codon biases. Unnamed viral proteins are designated by their position numbers in the following format— Pos: start position-stop position.

Table 6.2. A Selection of Viral Proteins and their Top Correlating Human Proteins, along with the Human Protein's Documented Area of Expression

Accession Number	Virus Name	Virus	Correlating Human	Protein's Expression Location
NC_004500	HPV 92	E1	MSH4	Testis
NC_022095	HPV 179	L1	HLTF	Testis
NC_014952	HPV 128	E1	TIGD4	Testis, vagina
NC_001691	HPV 50	E1	TEX15	Testis
NC_001405	HPV 18	L1	MRC2	Soft tissue, testis, endometrium
NC_001354	HPV 41	USP7	SLC12A2	Digestive tract, breast, placenta
NC_000898	HHV 6	U90	ELTD1	Gallbladder, breast, smooth muscle
NC_019023	HPV 166	E1	OTOGL	Cervix, testis
NC_009334	HHV 4	BALF5	SPTB	Epididymis
NC_010329	HPV 88	E1	RAD51AP2	Seminal Vesicle, Fallopian Tube
NC_004500	HPV 92	E1	USP9Y	Prostate

These results show that viral codon usage biases highly correlate with the codon usage biases of human proteins that are found within tissues that the viruses are known to promote symptomatic issues.

Chapter 6 Figures

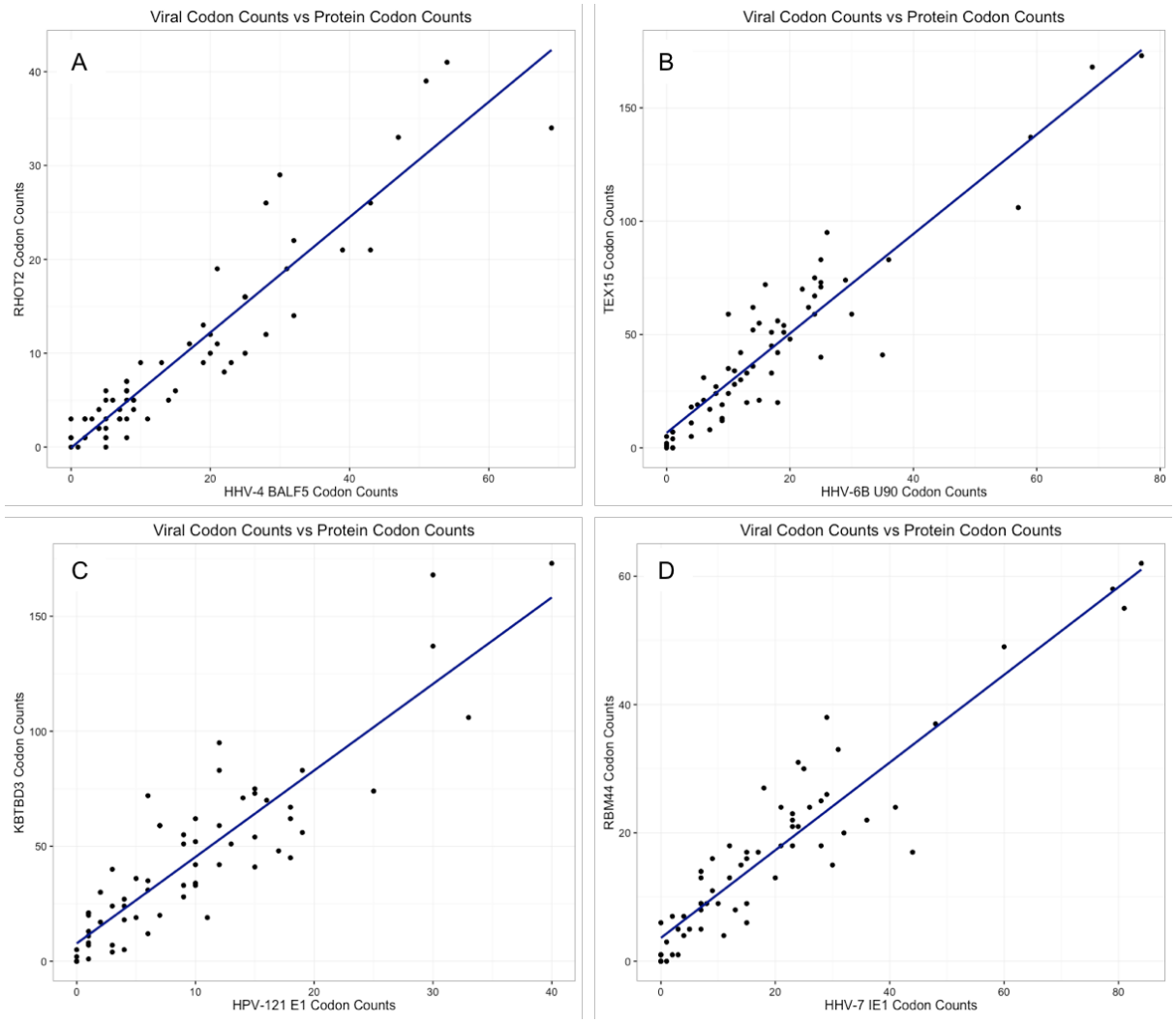


Figure 6.1. Codon Counts Four of the highest correlating virus-protein pairs found in Table 1 are displayed. We plotted codon counts for the viral protein (X-axis) against the human protein's codon counts (Y-axis). Each graph has 64 points, each representing a codon. Points near the top right are used at a higher rate than points near the bottom left. The line represents the result of a best-fit linear model, indicating that there is a strong correlation--as protein codon usage increases, so does the codon usage count of the respective virus. Residual plots of the linear regression were also analyzed and appear to fit the assumptions of the model. (A) displays RHOT2 vs HHV-4 (correlation of 93.6%), (B) shows TEX15 vs HHV-6B (correlation of 93.1%), (C) shows KBTBD3 vs HPV-121 (correlation of 92.8%), and (D) displays RBM44 vs HHV-7 (correlation of 92.8%). See Table 1 for more information on these pairs.

Manuscript published in *Bioinformatics*. 2018; https://doi.org/10.1093/bioinformatics/bty669

Chapter 7

JustOrthologs: a fast, accurate, and user-friendly ortholog identification algorithm

Justin B. Miller[1], Brandon D. Pickett[1], and Perry G. Ridge[1]

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

Abstract

Motivation: Orthologous gene identification is fundamental to all aspects of biology. For example, ortholog identification between species can provide functional insights for genes of unknown function and is a necessary step in phylogenetic inference. Currently, most ortholog identification algorithms require all-versus-all BLAST comparisons, which are time consuming and memory intensive.

Results: JustOrthologs is a novel approach to identifying orthologs that exploits the conservation of gene structure by using the lengths of coding sequence (CDS) regions as well as dinucleotide percentages to identify orthologs. In comparison to OrthoMCL, OMA, and OrthoFinder, JustOrthologs decreases ortholog identification runtime by more than 96% and achieves comparable precision and recall scores. The computational speedup allowed us to conduct pairwise comparisons of 1 197 complete genomes (780 eukaryotes and 417 archaea). We confirmed gene annotations for 384 120 genes, grouped 1 675 415 genes in previously unreported ortholog groups, and identified 51 429 potentially mislabeled genes across 622 843 ortholog groups.

Availability: JustOrthologs is an open source collaborative software package available in the GitHub repository: https://github.com/ridgelab/JustOrthologs/. All test FASTA files used for comparisons are freely available at https://github.com/ridgelab/JustOrthologs/comparisonFastaFiles/. Reference genomes used in this work are available for download from the NCBI repository: ftp://ftp.ncbi.nih.gov/genomes/.

Contact: perry.ridge@byu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Introduction

Ortholog identification has long been a daunting, yet critical, first step for many studies. Orthologs are gene sequences derived from the same ancestral gene present in two species' last common ancestor, and can provide support in phylogenetic tree reconstruction or insights into gene function (Koonin, 2005).

Unsurprisingly, many ortholog identification algorithms are currently available. Unfortunately, existing algorithms are complex and are hampered by poor performance. OrthoMCL requires a complicated 13-step process, which involves an all-versus-all BLAST comparison, a Markov Clustering (MCL) algorithm, and construction of a MySQL database to identify ortholog groups (Li, et al., 2003). OrthAgogue attempts to simplify the process by combining the MCL into a single step, and decreases the number of steps required in an OrthoMCL analysis from 13 to eight (Ekseth, et al., 2014); however, the eight-step process is still overwhelming for the average biologist. Using a different approach, OrthoFinder increases ortholog precision by taking into account a gene length bias associated with the all-versus-all BLAST scores (Emms and Kelly, 2015). While OrthoFinder is a single-step process, it still requires the installation of several software dependencies and is time-consuming to run. OMA evaluates the evolutionary relationships between proteomes through a pairwise comparison, with additional web interfaces and tools for querying their databases (Altenhoff, et al., 2015). OMA has over a dozen major releases, each of which increased the number of proteomes in the database. However, it requires a strict directory structure for independent ortholog identification and is not easily scriptable. Other algorithms, such as Inparanoid (Sonnhammer and Ostlund, 2015), EggNOG (Huerta-Cepas, et al., 2016), OrthoDB (Zdobnov, et al., 2017), and TreeFam (Schreiber, et al., 2014) take

a similar approach to OMA by maintaining a database of orthologous groups and providing tools to BLAST a query sequence against their respective database. While each software package implements a slightly different ortholog identification algorithm, each method is based on time-intensive all-versus-all BLAST comparisons for the initial scoring, which limits the typical dataset to a few specific genes of interest. Furthermore, external dependencies, intricate step-by-step processes, or a strict directory structure are often required, precluding inexperienced researchers from using these programs to identify orthologs. Therefore, comprehensive comparisons between algorithms require not only an analysis of accuracy, but also an evaluation of runtime complexity and ease of user experience. A comparison of the strengths and weaknesses of each of the three algorithms used for comparisons is found in Supplemental Table 1.

JustOrthologs is unlike any other ortholog identification algorithm. It exploits the conservation of coding sequence (CDS) region length to reduce the number of gene-gene sequence comparisons. By sorting each FASTA file by the number of CDS regions in each gene (i.e., the number of exons), fewer direct comparisons are required. Furthermore, rather than compare whole sequences (i.e. a BLAST comparison), JustOrthologs compares dinucleotide percentages to determine the level of sequence identify between two CDS regions. These innovations reduce runtime by at least 96% compared with other popular ortholog identification algorithms. Moreover, JustOrthologs has no external dependencies, has only a few, well-documented parameters, and requires only a single step at runtime.

Methods

Algorithm Design

Although JustOrthologs is run by a single command, the algorithm implements a two-step process. First, JustOrthologs utilizes a previously unreported conservation in CDS region length within orthologs. JustOrthologs compares CDS region lengths and requires that the two genes, with a couple exceptions, have CDS regions of the exact same lengths. JustOrthologs allows up to two CDS regions to differ in length within each sequence, thereby accommodating exon fusion and splitting events. Furthermore, since genes are sorted by the number of CDS regions, and only two fusion or splitting events are allowed, if the difference between the number of CDS regions in the query and subject sequences exceeds two, the remaining genes in the file are not compared. By limiting comparisons to only CDS regions, as described above, we significantly decrease the number of pairwise comparisons between genes.

Second, we further reduce computational complexity by completely avoiding BLAST comparisons in favor of dinucleotide usage percentages. A dinucleotide percentage is calculated by counting the occurrences of a dinucleotide pair in an exon and dividing by the total number of dinucleotide pairs in that exon. This process is repeated for each of the 16 possible combinations of dinucleotides (e.g. AG, CT, CC, etc.), and then repeated for each exon, creating dinucleotide motifs which can be compared between exons in other genes. If the difference in dinucleotide percentages between two sequences is lower than a threshold, and the lowest among possible orthologs in the subject file, then that gene is reported as orthologous to the query. Nucleotide bigrams were used to allow for greater sequence divergence within each CDS region, especially

at the third codon position. See Supplementary Figure 1 for an outline of the decision process for JustOrthologs.

We present three settings for JustOrthologs, each refined for a specific case: 1) comparison of closely related species, 2) comparison of distantly related species, and 3) a combination of the first two options to report the highest number of orthologs. Pseudocode for each of the three settings can be found in Supplementary Algorithms 1, 2, and 3 respectively.

Thresholds for dinucleotide percentages are set depending on which of the three use cases, described above, is set. For closely related species, the recommended threshold is 0.05, while distantly related species have a recommended threshold of 0.1. Both thresholds were tuned and calculated using species not shown in this paper so as not to inadvertently train our thresholds on our test cases. We tuned the threshold for closely related species by examining the precision and accuracy of recovered orthologs between *Alligator sinensis* and *Alligator mississippiensis* (52 MYA estimated time of divergence (Hedges, et al., 2006; Hedges, et al., 2015; Kumar and Hedges, 2011; Kumar, et al., 2017)) and *Myotis lucifugus* and *Myotis brandtii* (14.2 MYA estimated time of divergence (Hedges, et al., 2006; Hedges, et al., 2015; Kumar and Hedges, 2011; Kumar, et al., 2017)) for thresholds between 0.01 and 1.00, incremented by 0.01. The same process was completed for orthologs recovered from the more distantly related species, *Alligator sinensis* and *Myotis lucifugus* (312 MYA estimated time of divergence (Hedges, et al., 2006; Hedges, et al., 2015; Kumar and Hedges, 2011; Kumar, et al., 2017)). The threshold score is adjustable (see Supplementary Note for description on how to tune these thresholds using other species), although we have provided recommended thresholds based on our analyses.

All three settings of JustOrthologs are parallelized with the default setting to use as many cores as the system has available. Alternatively, the user may specify the number of cores. To improve the user experience, intuitive, well-documented argument parsing is included. A provided wrapper script allows users to extract all ortholog pairs from two FASTA files and two General Feature Format 3 (GFF3) files with options to extract all CDS regions, to sort based on the number of CDS regions, to filter based on gene annotation, and then to run any version of JustOrthologs and find all ortholog pairs between the two species. We provide a comprehensive README and README_WRAPPER for argument descriptions, as well as example FASTA and GFF3 files in the GitHub repository.

Ortholog Identification Across 1 197 Species

A common practice is to find orthologous genes across a group of species. Since JustOrthologs is designed for pairwise species comparisons, an independent Python script (combineOrthoGroups.py with accompanying documentation in README_OTHER_PROGRAMS) was written to combine the output from multiple JustOrthologs output files. CombineOrthoGroups takes as input a directory with the output files from one or more species comparisons completed using JustOrthologs. It reads each file, adding the pairwise ortholog groups to a dictionary of all ortholog pairs. It then finds all genes that belong to a group (e.g., if gene A in species 1 points to gene B in species 2, and gene B in species 2 points to gene C in species 3, then the ortholog group would contain genes A, B, and C). Because we are interested in identifying potentially mislabeled or previously unidentified orthologs, we applied a filter which requires one-to-one orthology (i.e., two genes from the same species cannot be reported as orthologous). While we realize that one-to-one orthology is not

always the best representation of phylogenetic history due to gene duplication, horizontal gene transfer, etc., one-to-one orthology ensures that orthologs are grouped based on the most-probable orthology and not because of paralogy or software error.

Generating Test Data

Since JustOrthologs requires DNA sequences and CDS annotations, we were unable to use traditional ortholog data sets, such as OrthoBench (Trachana, et al., 2011), which contain protein sequences without splice site annotations. Therefore, we relied on the Human Genome Organisation Gene Nomenclature Committee (HGNC) gene annotations and outline the creation of test data sets in Supplementary Figure 2. The HGNC uses ortholog annotations established by SWISS-PROT and the HGNC interacts with various nomenclature groups to ensure that orthologous genes between different species are assigned the same symbol. All FASTA sequence data for our main comparisons between 1 197 genomes and our pairwise comparisons between *Homo sapiens*, *Pan paniscus*, *Falco peregrinus*, and *Equus caballus*, were downloaded and extracted from the reference genomes and GFF3 files found in the NCBI database in September, 2017 (Pruitt, et al., 2014; Tatusova, et al., 2014). All 1 197 species are listed in Supplementary Table 2.

Three types of test data sets were created, each outlined in Supplementary Figure 2: 1) original, in which all genes included from species 1 have their true ortholog in species 2 included in the test set (i.e. everything in these test sets are true positives (TP)), 2) mismatch, which contains a mix of genes and their true orthologs, and genes with no orthologs in the data set—these test sets most closely approximate an unfiltered data set that might be used in research because they have

a mix of TPs and false positives (FP), and 3) error, which contain no TP orthologs (i.e. any orthologs identified in these test sets are FPs). Each test set includes up to 1 000 genes. Once a test set has 1 000 genes, a new test set is created starting where the last test set left off. In our test sets, mismatch test sets had 50-90% TPs. This process resulted in 33 test sets (11 of each type) for human versus falcon, 39 test sets (13 of each type) for human versus horse, and 45 test sets (15 of each type) for human versus bonobo.

We report estimated species divergence times between *Homo sapiens* (GCF_000001405.28), *Pan paniscus* (GCF_000258655.2), *Falco peregrinus* (GCF_000337955.2), and *Equus caballus* (GCF_000002305.2) in Table 1 to show that our comparisons span both closely and distantly related species. Filters were applied to these data to remove annotated translational errors, suspected errors, and unclassified transcription discrepancies. Similar to previous studies (Camiolo, et al., 2015), we included only the longest isoform of each gene in our analyses. To generate our test data, we relied on an upper and lower case insensitive review of gene names that were annotated by the HGNC (Gray, et al., 2015) to divide genes into several groups for testing as described below. Orthologs were considered TP if they matched the HGNC annotations, FP if they did not match HGNC annotations, and false negatives (FN) if genes with matching HGNC annotations were not reported. Any orthologs reported for the error data set were by definition FPs, as no TPs were possible.

We recognize that some HGNC gene annotations are potentially incorrect. However, these annotations are reliable for our testing and algorithm comparisons for two reasons. First, it is likely that a large majority of the annotations are correct, and since we use a total of 51 721

genes between the four species for testing, a small fraction of incorrect labels is unlikely to significantly affect the results. Second, all algorithms were evaluated using the same data sets, so all algorithms are subject to the same potentially incorrect annotations present in the test data sets.

Comparisons to OrthoMCL, OMA, and OrthoFinder

The OrthoMCL pipeline (Li, et al., 2003) has many steps and can be difficult to use. Nevertheless, the process is relatively well documented. During the all-versus-all BLAST, we used the NCBI BLAST+ suite version 2.2.28 (Camacho, et al., 2009) instead of the legacy BLAST suite (Altschul, et al., 1990). We used the BLAST+ provided Perl script, legacy_blast.pl, to convert the BLAST command to the correct form for BLAST+. Further modifications were required to obtain the desired output because the provided script is intended only as a starting point. After carefully reading the BLAST+ documentation, parameters for the final BLAST+ command were: -evalue 1e-5 -seq "yes" -num_descriptions 10000 -soft_masking true -outfmt 6. All other commands for OrthoMCL were as outlined in the original manuscript (Li, et al., 2003) and the step-by-step processes for OMA (Altenhoff, et al., 2015) and OrthoFinder (Emms and Kelly, 2015) were executed without modification.

Performance Measurements

Similar to the method outlined by Emms et al. (Emms and Kelly, 2015), we used precision and recall to evaluate our algorithms. In our study, precision is the ratio of true positive orthologs reported to total orthologs reported, while recall is the ratio of true positive orthologs reported to all possible real orthologs in each data set:

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN}$$

Some algorithms that we compared also searched for orthologs within the same species. JustOrthologs does not have this functionality, due to the high similarity of isoforms within a species and the rarity of such orthologs. Therefore, to ensure a fair comparison between algorithms, if an algorithm reported orthologs that did not include a sequence from each of the two species being compared, those specific orthologs were excluded from evaluation (e.g. in the *Homo sapiens* and *Pan paniscus* comparison, one gene from each of the species is required, as opposed to both genes being from a single species). Remaining groups were considered TPs if the groups had exactly one sequence from each species (as opposed to two or more from one or both species) and the gene names matched (i.e., the group exhibited one-to-one orthology between the two species). All tests were performed on an Intel Haswell (2.3 GHz) node with 24 cores. We allocated one node and 16 cores to each algorithm.

Results

Comparisons

Precision

Precision evaluates the confidence that ortholog pairs are correct. JustOrthologs had the best precision of the algorithms tested, with nearly 100% precision for each test data set. OrthoFinder also had 100% precision for all test sets, except human versus falcon, for which no ortholog pairs were reported. OrthoMCL had the lowest precision (~55-80%) for all test sets, while OMA had high precision (~100%) when only orthologs are present, but lower precision (~96%) when mismatches are present in the test data (Supplementary Figures 3 and 4).

Recall

Recall measures the number of correctly reported ortholog pairs out of the number of possible real ortholog pairs. JustOrthologs, OMA, and OrthoMCL had nearly 100% recall for human versus bonobo. For all three test sets, recall for JustOrthologs was much higher than OrthoFinder. Recall for JustOrthologs is comparable with the recall from OrthoMCL and OMA for closely related species, but JustOrthologs' recall was significantly lower for more distantly related species (Supplementary Figures 5 and 6). As expected from the algorithm's implementation, recall for JustOrthologs increases when more CDS regions are present in a gene because significant mutations within a few CDS regions can indicate speciation events while the remaining CDS regions remain relatively unchanged.

False Positive Rate

We used the error data sets to assess false positive rates. OrthoFinder did not report any false positives in any of the data sets. Likewise, JustOrthologs reported no false positives for human versus bonobo and human versus falcon test cases, but had a false positive rate of 0.008% for the human versus horse test cases. All other algorithms had high false positive rates: OrthoMCL (27-42%) and OMA (11-12.5%) (Supplementary Figures 7 and 8).

Performance

Since all-versus-all BLAST requires comparing all sequences within the same file (once for each file), and all sequences between files (using each file once as the subject), big-O time complexity for ortholog pair identification using all-versus-all BLAST based algorithms (i.e. all algorithms except JustOrthologs) is typically $O(n^4)$, where n is the number of sequences analyzed. In

contrast, the time complexity of JustOrthologs is a function of the number of genes with similar numbers of CDS regions (c) and the lengths of the compared CDS regions (l). Both values are usually significantly smaller than the total number of genes or the total number of CDS regions, and have very small constant factors. For the dinucleotide percentages that are actually compared, they are compared in a pairwise manner, leaving the maximum time complexity as $O(c^2l^2)$. In real-world scenarios, where relatively few genes contain similar numbers of CDS regions, the time complexity is more similar to a logarithmic function because of the initial sorting step limits sequence comparisons to only sequences with similar numbers of CDS regions. The dinucleotide comparisons also reduce complexity because the actual sequences are never aligned. The third setting of JustOrthologs, which is a combination of the first two, is twice as computationally intensive ($O(2c^2l^2)$) because it requires running both algorithms before combining the output from each.

We compared the user time, which accounts for execution time of each thread, (i.e. JustOrthologs gained no advantage in this comparison by having more efficient multi-threading) for each of the algorithms across all test data sets. JustOrthologs was substantially faster than all other algorithms, even in its slowest setting. The slowest setting of JustOrthologs was on average 28x faster than OrthoMCL, 96x faster than OMA, and 4900x faster than OrthoFinder. The two faster settings of JustOrthologs were always at least 58x faster than all other algorithms (Supplementary Figure 9).

Furthermore, the multiprocessing capabilities of JustOrthologs surpasses all other algorithms, with an average core utilization of 11.3 out of the 16 allocated cores. In comparison, OMA

averaged approximately 1.25 cores, OrthoFinder averaged approximately 5.0 cores, and OrthoMCL averaged approximately 6.6 cores out of 16 allocated cores, thus when comparing the use of each algorithm in a realistic setting (i.e. multi-threaded), JustOrthologs provides a more substantial advantage than reported here.

Results for Individual Tests

Precision, recall, and user time for individual tests for each algorithm are found in Supplementary Figures 10-33.

Ortholog Identification in 1 197 Species

Finally, as proof-of-concept, we used JustOrthologs to perform a pairwise comparison of all genes in 1 197 species. JustOrthologs finished each genome-wide pairwise comparison in 0-24 hours, depending on the number/length of annotated genes. In total, all pairwise comparisons took 45 476 hours to complete. We identified 1 675 415 currently unnamed genes that were classified as orthologous to other genes in different species. We also identified 51 429 potentially mislabeled genes, which we report. We report the first 30 ortholog groups identified by JustOrthologs in Table 2 and examine potentially mislabeled genes within those groups. In Table 2, several ortholog groups have poor sequence alignments. In Supplementary Note 2, we explain why a poor alignment might occur and give an example of two simulated sequences with a poor alignment that would be identified as orthologous. We have included a comprehensive list of all orthologous gene groups identified in these comparisons in Supplementary Table 3. Supplementary Tables 4 and 5 analyze the composition of these groups by reporting the

annotations and group sizes, respectively. We propose that the annotations of each of these genes should be examined and updated by the HGNC.

All ortholog identification algorithms are limited by their ability to successfully differentiate between paralogs and orthologs. Therefore, individual species comparisons where whole genome duplications occurred or where many homologs exist generally cause algorithms to report a higher number of false positive orthologs. In our comparison of 1 197 species, we also analyzed specific pairwise gene comparisons. We show 15 pairwise comparisons of complete genomes across diverse taxa in Table 3. We did not subsample genes from these data, which allows of a more complete view of how JustOrthologs performs on real-world datasets. Although recall is significantly affected in some species comparisons, JustOrthologs maintains high precision in all instances. Furthermore, thousands of previously unnamed genes are identified in orthologous pairs, facilitating the evaluation of their orthologous relationship. In the aforementioned orthology groups, we performed a strict one-to-one orthology filter to combine these pairwise relationships to minimize compounding false positive relationships.

Discussion

JustOrthologs significantly decreases ortholog classification runtimes, allowing faster ortholog comparisons on larger gene data sets than any other ortholog identification algorithm. The higher precision of JustOrthologs offers users more confidence in ortholog pairs identified by JustOrthologs than orthologs identified by OrthoMCL or OMA. JustOrthologs also offers higher recall in genes from closely related species with many CDS regions than any other algorithm, allowing better identification of orthologs with many splice sites. As might be expected, all ortholog identification algorithms perform best when analyzing closely related species such as

Homo sapiens versus *Pan paniscus*. Compared to other algorithms, JustOrthologs had a higher combined precision and recall score than any other algorithm for all test sets for closely related species. In more distantly related species, such as *Homo sapiens* versus *Equus caballus*, only OrthoFinder was more precise than JustOrthologs, but OrthoFinder had much lower recall—JustOrthologs identified over 6 000 ortholog groups that OrthoFinder missed. For more distantly related species, such as *Homo sapiens* versus *Falco peregrinus*, OrthoFinder reported no ortholog pairs, but JustOrthologs reported over 1 000 ortholog pairs, while maintaining approximately 99% precision. In contrast, less precise methods, such as OrthoMCL, reported only 70-80% precision on the same data sets. Overall, JustOrthologs is the most consistent performer among tested algorithms, and is significantly faster.

The decreased runtime allows JustOrthologs to perform whole genome analyses of diverse species that were previously impossible to perform. Since JustOrthologs uses a unique algorithm that does not rely on time-consuming all-versus-all BAST comparisons, it enables researchers to quickly identify potential orthologs using whole genome analyses. Since we opted to have higher precision than recall, orthologs reported by JustOrthologs have high precision, which allows researchers to have confidence in the reported ortholog pairs.

Moreover, JustOrthologs has comprehensive documentation and, compared to other algorithms, is easy to use. These characteristics, the provided wrapper scripts, and the single-step command line process that does not require any external software, make JustOrthologs accessible to even individuals with limited programming experience.

Although JustOrthologs is a novel approach that accurately and precisely recovers orthologous gene relationships without a sequence alignment, a sequence alignment could be used to evaluate proposed orthologous relationships identified by JustOrthologs. Since all-versus-all BLAST searches are computationally intractable when the number of sequences is large (e.g., whole genome analyses), using BLAST to evaluate the sequence alignments of the proposed orthologous pairs could be used to further improve accuracy with a limited computational cost. However, we opted not to include an alignment step in our algorithm to illustrate the predictive power of our novel approach. Furthermore, our approach allows for structural variants and rearrangements that a sequence alignment might miss.

Since JustOrthologs exploits CDS region length conservation, the algorithm works only with annotated CDS. However, as whole genome and transcriptome sequencing is becoming increasingly common, owing to reduced prices and better assembly/annotation software, this limitation is likely to decrease with time. Furthermore, JustOrthologs is better suited than any existing algorithm to handle the large data sets that have become the norm in biology. As evidence of the potential utility of JustOrthologs, we identified orthologous groups within 1 197 species, in 45 000 hours of real time using 16 processing cores (we farmed the analysis out to multiple processing nodes, so real time was calculated by summing the real time from each of the nodes). Extrapolating from measured times, such a comparison would not have been possible for any of the other algorithms compared in this manuscript.

The gold standard in science is perfectly accurate and complete data; however, few algorithms are capable of delivering both. We deliberately opted for JustOrthologs to have higher precision

than recall, because as biologists we prioritize confidence in the accuracy of our data as opposed to being comprehensive. For closely related species, the tradeoff is almost unnoticeable. However, similar to OrthoFinder, greater evolutionary distance between genes significantly decreases the recall of JustOrthologs. Nevertheless, recall for JustOrthologs significantly outperforms OrthoFinder for distantly related species.

JustOrthologs is a unique algorithm for ortholog identification as it departs from the traditional all-versus-all BLAST search algorithms that have saturated ortholog identification for the past decade. While all-versus-all BLAST has proven useful for small-scale analyses, its $O(n^4)$ runtime is prohibitive for species-wide ortholog identification. In fact, two algorithms, OMA and OrthoFinder, are incapable of completing a genome-wide ortholog comparison in a week. In an era of high throughput sequencing, an algorithm capable of efficiently searching entire genomes is necessary.

Acknowledgements

We appreciate the Fulton Supercomputing Laboratory at Brigham Young University (https://marylou.byu.edu), without which these analyses would not have been possible. We acknowledge the contributions of Brigham Young University for sponsoring our research and providing a facility in which to work.

Funding

No external funding was used.

Conflict of Interest: none declared.

References

Altenhoff, A.M.*, et al.* The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic acids research* 2015;43(Database issue):D240-249.

Altschul, S.F.*, et al.* Basic local alignment search tool. *J Mol Biol* 1990;215(3):403-410.

Camacho, C.*, et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.

Camiolo, S., Melito, S. and Porceddu, A. New insights into the interplay between codon bias determinants in plants. *DNA Res* 2015;22(6):461-470.

Ekseth, O.K., Kuiper, M. and Mironov, V. orthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics* 2014;30(5):734-736.

Emms, D.M. and Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome biology* 2015;16:157.

Gray, K.A.*, et al.* Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res* 2015;43(Database issue):D1079-1085.

Hedges, S.B., Dudley, J. and Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 2006;22(23):2971-2972.

Hedges, S.B.*, et al.* Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol* 2015;32(4):835-845.

Huerta-Cepas, J.*, et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 2016;44(D1):D286-293.

Koonin, E.V. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 2005;39:309-338.

Kumar, S. and Hedges, S.B. TimeTree2: species divergence times on the iPhone. *Bioinformatics* 2011;27(14):2023-2024.

Kumar, S.*, et al.* TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* 2017;34(7):1812-1819.

Li, L., Stoeckert, C.J., Jr. and Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* 2003;13(9):2178-2189.

Pruitt, K.D.*, et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* 2014;42(Database issue):D756-763.

Schreiber, F.*, et al.* TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res* 2014;42(Database issue):D922-925.

Sonnhammer, E.L. and Ostlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* 2015;43(Database issue):D234-239.

Tatusova, T.*, et al.* RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res* 2014;42(Database issue):D553-559.

Trachana, K.*, et al.* Orthology prediction methods: a quality assessment using curated protein families. *Bioessays* 2011;33(10):769-780.

Zdobnov, E.M.*, et al.* OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* 2017;45(D1):D744-D749.

Tables and Figures

Chapter 7 Tables

Table 7.1. Estimated Time of Species Divergence

Species 1	Species 2	Estimated Time	Median Time	Confidence Interval
Homo sapiens	*Pan paniscus*	6.65 MYA	6.4 MYA	6.23-7.07 MYA
Homo sapiens	*Equus caballus*	96 MYA	94 MYA	91-102 MYA
Homo sapiens	*Falco peregrinus*	312 MYA	320 MYA	297-326 MYA
Pan paniscus	*Equus caballus*	96 MYA	94 MYA	91-102 MYA
Pan paniscus	*Falco peregrinus*	312 MYA	320 MYA	297-326 MYA
Equus caballus	*Falco peregrinus*	312 MYA	320 MYA	297-326 MYA

Species Divergence taken from the average estimate from various studies included in TimeTree (Hedges, et al., 2006; Hedges, et al., 2015; Kumar and Hedges, 2011; Kumar, et al., 2017).

Table 7.2. Ortholog Groups recovered using JustOrthologs and CombineOrthoGroups

Genes with the Same Annotation	Genes with Other Annotations	Genes with Unknown Annotations	Total Genes	Reason for Other Annotations
127	0	63	190	N/A
178	0	7	185	N/A
172	1	7	180	XP_018109801.1 has 100% BLAST identity with NP_001087532.1, which is annotated the same as the other 172 genes
155	2	21	178	The nucleotide composition and exon length of XP_001959559.1 and XP_002071834.1 are similar to XP_010179458.1. However, the alignment is very different. These two genes are probably incorrectly reported as orthologous by JustOrthologs.
169	0	9	178	N/A
169	1	5	175	XP_414807.2 has a 99% BLAST identity with XP_015732072.1 from a closely related species, which is annotated the same as the other 169 genes.
166	0	5	171	N/A
165	1	5	171	NP_068697.1 is annotated Trp53inp1 instead of TP53INP1.
163	1	6	170	XP_014347657.1 is annotated LRRC8E instead of LRRC8C
165	0	4	169	N/A
161	0	7	168	N/A
162	0	5	167	N/A
161	1	4	166	XP_020368157.1 is incorrectly reported as orthologous by JustOrthologs. The CDS region lengths matched some exons in XP_005866852.1, but the alignment of the sequences was very poor.
163	0	3	166	N/A
152	1	13	166	XP_018123052.1 is annotated grb10.L instead of GRB10
161	0	4	165	N/A

156	0	9	165	N/A
159	0	6	165	N/A
160	0	5	165	N/A
160	0	4	164	N/A
159	0	5	164	N/A
158	0	5	163	N/A
156	1	5	162	XP_017312051.1 is incorrectly reported as orthologous by JustOrthologs. The CDS region lengths matched several exons within XP_020920808.1, but the alignment of the sequences was poor.
156	0	5	161	N/A
158	0	3	161	N/A
153	0	7	160	N/A
149	0	9	158	N/A
154	0	3	157	N/A
146	0	11	157	N/A
153	0	4	157	N/A

The first 30 ortholog groups are ordered from the most genes to the fewest genes. The first column shows the number of genes with the same annotations. The second column shows the number of genes with a different annotation than the genes in the first column. The third column shows the number of genes without annotations. The fourth column shows the total number of genes in the ortholog group. The fifth column is an analysis of why genes in the second column were not annotated the same as genes in the first column but were reported as orthologous by JustOrthologs. Each gene comes from a different species.

Table 7.3. Whole Genome Comparison of Different Species

Species 1	Species 2	Number of Genes in Species 1	Number of Genes in Species 2	Number of Shared Ortholog Annotations from HGNC	True Positives Reported	False Positives Reported	Unnamed genes reported in orthologous pairs	Precision (%)	Recall (%)
Homo sapiens	*Pan paniscus*	20 088	17 900	14 653	14 119	462	905	96.83	96.36
Homo sapiens	*Equus caballus*	20 088	16 691	12 725	8 229	150	246	98.21	64.67
Homo sapiens	*Falco peregrinus*	20 088	12 643	10 659	841	38	35	95.68	7.89
Gallus gallus	*Falco peregrinus*	16 420	12 643	9 163	5 132	139	597	97.36	56.01
Astyanax mexicanus	*Danio rerio*	21 920	22 408	5 832	683	296	688	69.77	11.71
Cynoglossus semilaevis	*Danio rerio*	19 450	22 408	5 699	199	104	205	65.68	3.49
Oncorhynchus kisutch	*Salmo salar*	30 680	40 642	2 800	2 424	183	18 300	92.98	86.57
Oreochromis niloticus	*Pundamilia nyererei*	27 785	21 832	8 645	8 326	94	9 857	98.88	96.31
Alligator mississippiensis	*Crocodylus porosus*	17 492	13 837	10 993	10 238	4	1615	99.96	93.13
Mus musculus	*Rattus norvegicus*	21 815	21 481	15 199	12 183	720	279	94.42	80.16
Bos taurus	*Capra hircus*	17 980	19 208	12 894	11 929	97	1 337	99.19	92.52
Bos taurus	*Vicugna pacos*	17 980	16 297	11 411	7 991	18	502	99.78	70.03
Calypte anna	*Haliaeetus leucocephalus*	12 225	14 150	9 825	7 041	15	662	99.79	71.66
Calypte anna	*Chaetura pelagica*	12 225	11 852	8 770	6 565	14	695	99.79	74.86
Prunus avium	*Prunus mume*	24 179	22 628	0	0	0	14 004	N/A	N/A

All available genes are compared between various species. The first two columns are the names of the species being compared. Columns three and four indicate how many genes are present in each species. Column five shows how many genes have the same ortholog annotations in both species. Column six shows the number of true positives JustOrthologs identifies. Column seven shows the number of false positives identified by JustOrthologs. Column eight shows the number of genes reported as orthologous by JustOrthologs but not named by the HGNC. Columns nine and ten report the precision and recall of the compared species, respectively.

Manuscript submitted to *Nucleic Acids Research*

Chapter 8

ExtRamp: a novel algorithm for extracting the ramp sequence based on the tRNA adaptation index or relative codon adaptiveness

Justin B. Miller[1,+], Logan R. Brase[1,+], and Perry G. Ridge[1]

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

[+]*Contributed equally to this work*

Abstract

Different species, genes, and locations within genes use different codons to fine-tune gene expression. Within genes, the ramp sequence assists in ribosome spacing and decreases downstream collisions by incorporating slowly-translated codons at the beginning of a gene. Although previously reported as occurring in some species, no previous attempt at extracting the ramp sequence from specific genes has been published. We present ExtRamp, a software package that quickly extracts ramp sequences from any species using the tRNA adaptation index or relative codon adaptiveness. Different filters facilitate the analysis of codon efficiency and enable identification of genes with a ramp sequence. We validate the existence of a ramp sequence in most species by running ExtRamp on 229 742 339 genes across 23 428 species. We evaluate differences in reported ramp sequences when we use different parameters. Using the strictest ramp sequence cut-off, we show that across most taxonomic groups, ramp sequences are approximately 20-40 codons long and occur in about 10% of gene sequences. We also show that in Drosophila melanogaster as gene expression increases, a higher proportion of genes have ramp sequences. We provide a framework for performing this analysis on other species. ExtRamp is freely available at https://github.com/ridgelab/ExtRamp.

Introduction

The central dogma of biology shows that three consecutive nucleotides of coding DNA, called codons, are transcribed into messenger RNA (mRNA), mRNA is translated into amino acids, and amino acids form proteins (Crick, 1970). There are 61 canonical codons plus three stop codons that form and regulate the creation of 20 amino acids (Crick et al., 1961). Since there are more codons than amino acids, in many cases multiple synonymous codons encode the same amino acid. Although originally presumed to be identical in function, unequal distributions of synonymous codons quickly led to two non-mutually exclusive hypotheses: 1. non-random mutations occur particularly at the third codon position, and 2. selection for codon bias persists (Hershberg and Petrov, 2008; Quax et al., 2015). Furthermore, highly expressed genes display more prominent codon usage biases, suggesting that synonymous codons might play different roles in species fitness (Sharp and Li, 1986). The unequal abundance of tRNA anticodons led to the wobble hypothesis: tRNA anticodons do not need to latch onto all three codon nucleotides (Crick, 1966). However, codon usage is highly associated with the most abundant tRNA present in the cell (Post et al., 1979). Furthermore, codon usage patterns affect gene expression, with codons latching onto fewer than all three tRNA anticodons being considered suboptimal for gene expression (Gutman and Hatfield, 1989).

Although increased gene expression is often considered optimal, suboptimal codons are preferred in certain genes or parts of genes because they slow translation and reduce translational errors. For instance, a short set of 30-50 slowly-translated, suboptimal codons was identified at the 5' end of many protein coding sequences, which serves to evenly space ribosomes (Tuller et al., 2010) and reduce mRNA secondary structure (Goodman et al., 2013) at translation initiation.

This region has codons that are less adapted to the tRNA pool and consequently the ramp sequence has a slower elongation speed relative to the rest of the gene (Tuller and Zur, 2015). This ramp could be caused by any of three features correlating with slower translation elongation speed: codon adaptation to the tRNA pool, amino acid charge, and mRNA folding energy (Tuller et al., 2011). The ramp sequence was discovered by using a sliding window of 15 codons (although verification was done with sliding windows ranging from 10 to 20 codons), representing the length of the ribosome footprint (Tuller et al., 2010). However, more recent estimates of the ribosome footprint range from 15 nucleotides (5 codons) to about 45 nucleotides (15 codons) with a commonly accepted length of 28 nucleotides (about 9 codons) (Martens et al., 2015). Therefore, any algorithm for extracting the ramp sequence must be capable of adapting to different ribosome footprints by changing the size of the sliding window. Finally, since the ramp sequence is a relative measure of codon efficiency for each gene and not an absolute measure for the whole genome, each gene sequence must be analyzed individually (Tuller and Zur, 2015).

Several methods have been used to calculate the effect that each codon has on overall translation efficiency. Two of the most common approaches are the *Codon Adaptation Index* (CAI) (Sharp and Li, 1987a), which calculates a normalized value for each codon based on a set of highly expressed genes from the organism, and the *Effective Number of Codons* (N_c) model (Wright, 1990), which uses a population genetics approach to calculate the efficiency of each codon based on its overall usage in the species. To calculate CAI, two relative adaptiveness measures are used. First, the relative synonymous codon usage (RSCU) is calculated by dividing the observed frequency of a codon by the frequency of each codon encoding the same amino acid, assuming equal usage. Second, the relative adaptiveness of a codon (w_{ij}) is calculated for the j-th codon in

the i-th amino acid. The w_{ij} metric is the ratio of $RSCU_{ij}$ to $RSCU_{imax}$ for the i-th amino acid (Sharp and Li, 1987b). The *tRNA Adaptation Index* (tAI) (dos Reis et al., 2003) more accurately reflects changes in overall translational efficiency due to wobble interactions, tRNA composition, and synonymous codon position within a gene (Brule and Grayhack, 2017). However, most species do not have annotated tAI values. Since the tAI and w_{ij} both measure overall translational efficiency, w_{ij} can be used as a proxy for tAI when only sequence data are available.

We present ExtRamp, the first algorithm that can identify areas of decreased translational efficiency at the start of individual genes using tAI, w_{ij}, or any other codon efficiency table. No existing algorithm can identify ramp sequences in individual genes. We validate our approach by recreating the whole genome trends identified by Tuller et al. (2010) in *Saccharomyces cerevisiae, Drosophila melanogaster,* and *Caenorhabditis elegans* using tAI values. Moreover, we demonstrate the effectiveness of w_{ij} as a proxy for tAI by using it to detect the same patterns. Finally, we provide statistics of ramp usages and relative codon adaptiveness in 23 428 species across all domains of life.

Materials and Methods

Data Collection and Processing

We use the coding sequences (CDS) from 23 428 species from the following taxonomic groups with some overlap between viruses and bacteria: 418 archaea, 15 063 bacteria, 234 fungi, 149 invertebrates, 89 plants, 75 protozoa, 107 mammalian vertebrates, 123 other vertebrates, and 7 233 viruses. All CDS regions were downloaded from the National Center for Biotechnology

Information (NCBI) in September, 2017 (Wheeler et al., 2007; Pruitt et al., 2014; Tatusova et al., 2014). The reference sequences for each gene were used because they are the most complete compilation of the alleles in a given species (Wheeler et al., 2007). We always used the longest isoform, when given a choice, and we filtered out partial gene sequences and sequences with annotated exceptions (i.e., unclassified transcription discrepancy, suspected errors, translational exception, etc.).

The tAI values were downloaded from the tAI Calculator (http://tau-tai.azurewebsites.net) (Sabi et al., 2017). We provide tAI data for *Escherichia coli* and *Saccharomyces cerevisiae* in the GitHub repository as examples of the two comma-separated values (CSV) file formats accepted by ExtRamp. Since tAI values reflect the overall translational efficiency of a species better than w_{ij}, we recommend using tAI values, where available.

Extracting the Ramp Sequence

ExtRamp has two options to extract the ramp sequence, determined by user input (see Figure 1). The first option uses tAI values (or other codon efficiency values). ExtRamp first removes the start and stop codons. Then the algorithm walks over each gene, codon by codon, and matches the associated tAI to each codon, creating an ordered list of codon efficiencies within that gene.

Optionally, local codon efficiency bottlenecks are then calculated by taking a sliding window the size of a ribosomal footprint (default nine codons (Ingolia et al., 2009)) of codon efficiencies, finding the middle (default harmonic mean) of each sliding window, and then determining where in the gene these bottlenecks occur, similar to the methods in Navon and Pilpel (2011) (see

Figure 2 for a detailed explanation with an example). Next, ExtRamp optionally determines regions across all genes in the input FASTA file where more local bottlenecks occur than expected by random chance (default is true outliers). Using this method, we take the most conservative approach to determine in which percentage of the gene the local minimum must occur for a ramp sequence to be identified by ensuring that all percentages (from 1 to n) are outliers (e.g., if 1,2,3,5,7 are outlier regions, then the local bottleneck must occur in the first 3% of the gene because 4 was not an outlier region). The user can specify outlier regions as well, in which case the bottleneck must occur within the user-defined outlier region. If a bottleneck occurs within this region, then the mean codon efficiency of the entire sequence is calculated. The ramp is extended beyond the bottleneck until the sliding window codon efficiency exceeds the mean codon efficiency of the whole sequence.

If the user specifies the arithmetic mean, geometric mean, or median, the local translational bottleneck method outlined above is used as default. However, standard deviations can be used instead of local bottlenecks. If standard deviations are used, then the mean and standard deviation of all codon efficiencies within the sequence are calculated. Using a sliding window starting from the beginning of the gene, the ramp sequence extends until the mean codon efficiency within the sliding window exceeds the mean codon efficiency of the entire gene sequence minus the standard deviations specified by the user. However, since codon efficiencies have a large degree of variance and the sequences are relatively small, typically standard deviations must be smaller than 1.0 in order to identify ramp sequences.

An optional quality control step ensures that reported ramp sequences have similar lengths by calculating the average ramp length across all identified ramp sequences and removing ramp sequences that are in the tailing regions outside of a user-defined number (recommended two) of standard deviations above or below the mean length. Each step is multithreaded and by default uses all available processing cores, although any number of processing cores can be specified by the user.

The second option is used when the user does not supply the codon efficiency values for the species (i.e., the tAI values are not available for the species). This option uses either the input FASTA file or a user-supplied input FASTA file (typically containing highly expressed genes) to calculate the RSCU for each codon using the following formula, where x_{ij} is the occurrences of the j-th codon in the i-th amino acid, and n_i is the number of alternative codons for the i-th amino acid (Sharp and Li, 1987b):

Equation (1):

$$RSCU_{ij} = \frac{x_{ij}}{\frac{1}{n}\sum_{j=1}^{n_i} x_{ij}}$$

Next, w_{ij} is calculated using the following formula:

Equation (2):

$$w_{ij} = \frac{RSCU_{ij}}{RSCU_{imax}}$$

These ratios estimate codon adaptiveness for a species, with smaller ratios associated with less adaptive (efficient) codons. Once these efficiencies are calculated, the analysis is the same as the tAI method, with w_{ij} substituting the tAI values. This second method extends the utility of ExtRamp to non-model organisms that are not yet included in the tAI library.

Program Options

ExtRamp is written in Python 3.5 and requires a few standard libraries which can easily be installed using pip3 (process outlined in the GitHub README). To increase the versatility of ExtRamp, we include several options that can be split into two categories: controlling input and output files, and specifying variables used in the algorithm. To see real-time progress of the algorithm at runtime, the `-v` (verbose) option can be used.

An input FASTA file with CDS sequences is required using the `-i` option. By default, DNA sequences are expected, although RNA can be provided using the `-r` flag. An optional input file containing tAI values (or other codon efficiency values) for each codon can also be provided using the `-a` option. If tAI values are not provided, ExtRamp will calculate ramp sequences based on w_{ij} for the input FASTA file. However, w_{ij} can be calculated on a different FASTA file using the `-u` option. By default, ramp sequences are printed to standard out (terminal) in FASTA format to facilitate piping the results into additional analysis tools. To print the ramp sequences to a file, the `-o` option can be provided. The list of local translation efficiencies for each sequence can be printed to a CSV file using the `-l` option. Each of the efficiency sequences, are smoothed using the ribosomal window length (discussed below) and the data are printed in 'tidy' format (Wickham, 2014) for easy graphing using R. An unsmoothed list of all codon efficiency speeds for each codon can also be written to a file using the `-p` option. A list of the gene names that did not contain any calculable ramp sequence can be written to a text file using the `-n` option and sequences that are removed because they are not divisible by three or do not exceed the minimum sequence length can be written to another file using the `-z` option.

There are nine options that control variables used in the analysis performed in ExtRamp. The `-t` option controls the number of threads used, with the default being all available processing cores. The `-q` option sets the minimum length of a sequence to be analyzed. Similar to the methods used by Navon and Pilpel (2011), the default is 300 nucleotides (100 codons). Since there are several methods to determine the middle of a dataset, we provide the `-m` option with inputs of mean (arithmetic mean), median, gmean (geometric mean), and the default of hmean (harmonic mean). The `-s` option controls the number of standard deviations below the average of the consensus codon efficiency list for the maximum codon efficiency within a ramp sequence (typically less than one because codon efficiencies have large variances). The `-d` option controls the number of standard deviations above or below the mean ramp sequence length for all reported ramp sequences (if used, we recommend two standard deviations). The `-w` option controls the ribosomal window length that is used to smooth the proposed ramp sequences to minimize excess noise from spikes and dips in individual codon efficiencies. The default ribosomal window length is nine codons. The `-f` flag determines the outlier bottleneck regions based on sequences included in the input FASTA file. By default, the `-f` flag finds true outliers in the dataset. However, this can be modified using the `-e` option to find regions above a percentile (e.g., 75 would find places in a gene that have bottlenecks in the 75th percentile or above). The `-c` option sets the outlier region percentage (e.g., 10 would mean that the bottleneck must occur in the first 10% of the gene sequence). The default outlier region is in the first eight percent of the gene sequence.

Algorithm Validation

To validate our approach, we compared the consensus efficiencies calculated by ExtRamp for *Saccharomyces cerevisiae, Drosophila melanogaster,* and *Caenorhabditis elegans* to results by Tuller et al. (2010). We used the tAI values published in that study rather than updated values to enable accurate comparisons. We found the consensus efficiency for each species using the ExtRamp algorithm and graphed the results. We ran the algorithm using the `-m mean` option to match the method used by Tuller et al. (2010). The local efficiency values were also smoothed with a window size of four for consistency with their methods.

FlyBase Comparison

We used RNA-Seq gene expression values reported in FlyBase (http://flybase.org/rnaseq/profile_search) (Gramates et al., 2017) to determine if reported ramp sequences were associated with gene expression values (see Figure 3). We combined all expression data from both males and females at 1, 5, and 30 days old. Using the 'Expression On' utility, we pulled the FlyBase gene names for each of the eight expression level bins: 'No/Extremely low', 'Very low', 'Low', 'Moderate', 'Moderately high', 'High', 'Very high', and 'Extremely high'. These gene names were converted to protein names using the provided FlyBase 'convert' tool to facilitate comparisons with our dataset. The RNA-Seq Profile tool uses a 'not less than' approach, so by default the 'No/Extremely low' bin contains all the genes that are identified by the higher expression bins as well. We ensured that each bin contained only genes with a certain expression level by removing all genes reported in bins with higher expressions.

We ran ExtRamp on the *Drosophila melanogaster* CDS regions using the default options with tAI values. We then counted the number of ramp sequences for each expression level. Converting from gene names to protein names amplifies the number of sequences because there are multiple isoforms for each gene. Since we used the longest isoform of each gene, we used the number of gene names for the total number of sequences possible, instead of the number of protein names. Using a Chi-squared test, we checked if the number of hits for each expression level significantly differed from random.

W_{ij} versus tAI Option Comparison

To determine if running ExtRamp with and without tAI values produces similar results, we ran ExtRamp with and without tAI values on five species: *Acidilobus saccharovorans, Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster,* and *Saccharomyces cerevisiae.* We calculated the number of shared ramp sequences between the two techniques. We then used phyper, a mark and recapture statistical test built into R, to determine if the number of common elements was statistically significant. The following options were used: p = number of common sequences, m = number of tAI extracted sequences, n = total number of CDS tested – m, k = number of extracted sequences using w_{ij}, and *lower.tail* = FALSE.

Comparison Across All Domains of Life

To further validate our approach, we used ExtRamp to extract ramp sequences from 229 742 339 gene sequences found in 23 428 species. We used the w_{ij} method instead of tAI values for this analysis because tAI values are not available for most species. After extracting the ramp sequence from each gene, we determined the length of each ramp. For each species partition, we

plot the percentage of genes with a ramp sequence and the length of the identified ramp sequences.

Results

We first tested the accuracy of our algorithm by replicating the consensus translation efficiency of species reported in *Tuller et al.* Using parameters specified in their manuscript, ExtRamp reports identical codon efficiencies at each position (Figure 4).

We then determined if ramp sequences were associated with gene expression values using the tAI method (Table 1). Using the detailed gene expression data available for *Drosophila melanogaster*, we compared the isolated ramp sequences to their respective expression level bin. Using a Chi-squared test, we compared the number of genes found in each bin to the expected number if the ramp sequences were proportionally distributed between the expression bins. The reported Chi-squared value was 58.2 with seven degrees of freedom and a p-value of 3.45×10^{-10}. A clear progression of increasing standard residuals is seen from genes with low expression to extremely high expression. Very low and extremely low expression genes have slightly higher standard residuals that low expression genes (0.19 and -1.28, respectively). However, the residuals are much lower than very high and extremely high expression genes (2.87 and 6.24, respectively). We plot the standard residuals in Figure 5 to show the trend toward more ramp sequences in more highly expressed genes in *Drosophila*.

We compared the w_{ij} and the tAI approaches on identifying ramp sequences. The number of ramp sequences extracted from each species varied between these approaches, so we calculated

if the number of common sequences between the approaches was random or if both options were targeting the same sequences. Using a Mark and Recapture statistical approach on five species, four of the five species had very significant p-values (less than $1x10^{-6}$), indicating that the two approaches typically identified ramp sequences for the same genes (Table 2).

Finally, we identified ramp sequences in all genes from 23 428 different species across all domains of life using the w_{ij} method. In all instances, similar ramp sequences were reported using any of the four middle values: geometric mean, harmonic mean, arithmetic mean, and median. The first 5-10% of the gene was typically considered an outlier region, with protozoa having a slightly lower average (2-5%) and viruses reported almost no outlier regions (see Figure 6). Reported ramp lengths in sequences with a ramp typically ranged from about 60 to 120 nucleotides (20-40 codons), with plants having a slightly higher average length (about 25-55 codons) and viruses having a slightly lower average length (about 10-20 codons) (see Figure 7). Bacteria and plants reported the highest percentage of genes with ramp sequences (15-30%), while viruses reported almost no genes with ramp sequences (see Figure 8). Using the translation bottleneck technique with the strictest filter for outliers, most taxonomic groups report ramp sequences for about 10% of all species genomes (Figure 8).

We also analyzed the outlier regions that were identified by ExtRamp. Since all middle values report similar ramp sequences, we chose the default harmonic mean to analyze the outlier percentiles. We first removed the start and stop codon from each gene in the genome. Then we divided each gene into 100 equal parts and determined in which part the translational bottleneck occurred. Where multiple equal bottlenecks were identified, all bottlenecks were included in the

analysis. We show that in bacteria, invertebrates, mammals, other vertebrates, and plants, 100% of the species had an outlier region in the first percentile of their genes, and the outlier region extends to the tenth percentile in most taxonomic groups (see Figure 9). We also found that all taxonomic groups have an outlier region in the last percentile (99^{th} percentile) of the gene. Each taxonomic group except viruses clearly shows an outlier region at the beginning of the gene sequence with very few outlier regions between the first 10% of the gene and the end of the gene.

Discussion

Using the strictest settings on ExtRamp, most taxonomic groups had similar percentages of ramp sequences (approximately 10% of genes) and ramp sequence lengths (20-40 codons). However, bacteria and plants reported significantly more ramp sequences (approximately 25% of genes), while viruses reported almost no genes with a ramp sequence. In plants, the reported cutoff value was higher than the other taxonomic groups (Figure 6), indicating the outlier regions extended farther into the gene sequence. More outlier regions could indicate that in plants, translation bottlenecks occur in a wider region at the beginning of gene sequences than in other taxonomic groups. A higher reported cutoff percentage would increase the number of ramp sequences identified and could account for the increased number of ramp sequences shown in Figure 8. In bacteria, the mean cutoff percentage is slightly higher than in other taxonomic groups. However, the density distribution is much more tightly concentrated around the mean (see Figure 6). A tighter density distribution indicates that bacteria report more similar cutoff percentages between species than inter-species comparisons in other taxonomic groups. This tighter density distribution could also indicate that more selective pressure exists in bacteria to maintain a ramp

sequence than in other taxonomic groups. Unsurprisingly, viruses reported almost no ramp sequences. Viral genes are populated with regulatory sequences at the beginning of the coding region and host-repeating substrings throughout the genetic code (Goz et al., 2018), which potentially limits the applicability of selection for ramp sequences in viruses.

We also present evidence that a clear progression of increasing proportions of ramp sequences are identified from low expressed genes to extremely high expressed genes in *Drosophila melanogaster* (Table 1). We plot the standard residuals for each expression bin (Figure 5) and show that the highest standard residual (6.24) is found in the 'Extremely high' expression bin, while all expression bins less than or equal to 'Moderately high' expression have standard residuals at or below zero. This analysis complements previous studies indicating that ramp sequences are more prevalent in highly expressed genes.

Although the w_{ij} and tAI methods detect different numbers of sequences as containing ramps, they largely target the same sequences, with four of the five species analyzed having p-values less than $1x10^{-6}$ (Table 2). Since tAI values are not available for most species, further evaluation with a more robust tAI library might indicate systematic biases of tAI, whether from a phylogenetic or algorithmic standpoint. It is probable that tAI is more accurate in some species, or the correlation between tAI and w_{ij} is not universal. However, through our analysis, we show that although w_{ij} and tAI recover different numbers of ramp sequences, both methods typically target the same sequences.

ExtRamp can also aid in the analysis of a gene as a whole. Because the ramp sequences behave differently than the rest of the gene, it can skew the results of certain analyses. Some studies have avoided the problem by removing the first 50 codons of all the sequences before performing their analyses (Yang et al., 2014). However, this practice removes potentially valuable data and is not universally accurate for all sequences or species. At least two solutions to this predicament are as follows: 1. Determine the exact ramp sequence for each gene (possibly none) and remove only those portions, thereby keeping more of the sequence data for downstream analysis. 2. Incorporate the annotated ramp sequences in the downstream analysis tools.

We also provide the option to view the local translation efficiencies for each sequence that can easily be plotted using R. With these data, analyses can extend beyond the ramp sequence into the body of the gene. Furthermore, the option to view codon efficiency at each position allows for more extensive analyses involving local translational bottlenecks and codon usages. Future analyses could evaluate if there are correlations between physical characteristics such as functional domains of the gene and the translational efficiency of that section of the gene.

Very few studies have been performed on ramp sequences because software for extracting individual ramp sequences does not exist. We developed this algorithm to fill this need and improve the study of ramp sequences. Many studies look at ramp sequences on a high level, either evaluating the average length of the sequences in a species or determining the codon usage bias that influences the ramp. ExtRamp is the first algorithm to isolate the ramp sequence from individual genes, and it is the first attempt to analyze ramp sequences in non-model organisms.

Future research can determine which codons, specifically, are targeted in the ramp sequence, if ramps have a different mutation rate than the rest of the gene, if ramp sequences are associated with DNA structure, and if the length of the ramps can be used as a predictor for expression levels. We anticipate that ExtRamp will make ramp sequence research more accessible and assist in uncovering more biologically meaningful interpretations of the ramp sequence.

Availability

ExtRamp is an open source collaborative project available in the GitHub repository (https://github.com/ridgelab/ExtRamp)

Acknowledgement

We appreciate the contributions of Brigham Young University and the Fulton Supercomputing Laboratory at Brigham Young University for supporting our research. We would also like to thank Hannah Wadham for her expert suggestions on statistically validating ExtRamp.

Funding

No external funding was used for this research.

Conflict of Interest

The authors declare no conflict of interests.

References

Brule, C. E., Grayhack, E. J. 2017. Synonymous Codons: Choose Wisely for Expression. Trends Genet 33, 283-297.

Crick, F. 1970. Central dogma of molecular biology. Nature 227, 561-563.

Crick, F. H. 1966. Codon--anticodon pairing: the wobble hypothesis. J Mol Biol 19, 548-555.

Crick, F. H., Barnett, L., Brenner, S., Watts-Tobin, R. J. 1961. General nature of the genetic code for proteins. Nature 192, 1227-1232.

dos Reis, M., Wernisch, L., Savva, R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. Nucleic Acids Research 31, 6976-6985.

Goodman, D. B., Church, G. M., Kosuri, S. 2013. Causes and effects of N-terminal codon bias in bacterial genes. Science 342, 475-479.

Goz, E., Zafrir, Z., Tuller, T. 2018. Universal evolutionary selection for high dimensional silent patterns of information hidden in the redundancy of viral genetic code. Bioinformatics.

Gramates, L. S., Marygold, S. J., Santos, G. d., Urbano, J.-M., Antonazzo, G., Matthews, B. B., Rey, A. J., Tabone, C. J., Crosby, M. A., Emmert, D. B., Falls, K., Goodman, J. L., Hu, Y., Ponting, L., Schroeder, A. J., Strelets, V. B., Thurmond, J., Zhou, P., the FlyBase Consortium. 2017. FlyBase at 25: looking to the future. Nucleic Acids Research 45, D663-D671.

Gutman, G. A., Hatfield, G. W. 1989. Nonrandom utilization of codon pairs in Escherichia coli. Proc Natl Acad Sci U S A 86, 3699-3703.

Hershberg, R., Petrov, D. A. 2008. Selection on codon bias. Annu Rev Genet 42, 287-299.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., Weissman, J. S. 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science (New York, N.Y.) 324, 218-223.

Martens, A. T., Taylor, J., Hilser, V. J. 2015. Ribosome A and P sites revealed by length analysis of ribosome profiling data. Nucleic Acids Res 43, 3680-3687.

Navon, S., Pilpel, Y. 2011. The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. Genome Biol 12, R12.

Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H., Dennis, P. P. 1979. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. Proc Natl Acad Sci U S A 76, 1697-1701.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., Tully, R. E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D. R., Murphy, T. D., Ostell, J. M. 2014. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42, D756-763.

Quax, T. E., Claassens, N. J., Soll, D., van der Oost, J. 2015. Codon Bias as a Means to Fine-Tune Gene Expression. Mol Cell 59, 149-161.

Sabi, R., Volvovitch Daniel, R., Tuller, T. 2017. stAIcalc: tRNA adaptation index calculator based on species-specific weights. Bioinformatics 33, 589-591.

Sharp, P. M., Li, W. H. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24, 28-38.

Sharp, P. M., Li, W. H. 1987a. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research 15, 1281-1295.

Sharp, P. M., Li, W. H. 1987b. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15.

Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K., Tolstoy, I. 2014. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res 42, D553-559.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., Pilpel, Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141, 344-354.

Tuller, T., Veksler-Lublinsky, I., Gazit, N., Kupiec, M., Ruppin, E., Ziv-Ukelson, M. 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol 12, R110.

Tuller, T., Zur, H. 2015. Multiple roles of the coding sequence 5' end in gene expression regulation. Nucleic Acids Res 43, 13-28.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., Yaschenko, E. 2007. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35, D5-12.

Wickham, H. 2014. Tidy Data. Journal of Statistical Software; Vol 1, Issue 10 (2014).

Wright, F. 1990. The 'effective number of codons' used in a gene. Gene 87, 23-29.

Yang, J. R., Chen, X., Zhang, J. 2014. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. PLoS Biol 12, e1001910.

Tables and Figures

Chapter 8 Tables

Table 8.1. tAI Ramp Sequences for FlyBase Expression Bins

Expression Level	Observed Ramps	Total Sequences	Expected Ramps	Standard Residuals
No/Extremely low	73	726	84.82014734	-1.28
Very low	182	1536	179.454196	0.19
Low	181	1830	213.8028507	-2.24
Moderate	337	3162	369.4232864	-1.69
Moderately high	312	2655	310.1893818	0.1
High	177	1383	161.5788757	1.21
Very high	155	1054	123.1410955	2.87
Extremely high	42	142	16.59016656	6.24
Total	1459	12488	1459	

Ramp sequences were extracted from all genes reported in FlyBase using ExtRamp and tAI values. For each expression level (No/Extremely low to Extremely high), the number of observed ramp sequences was compared to the expected number of ramp sequences if ramp sequences were not associated with expression (i.e., the total proportion of ramp sequences multiplied by the total number of sequences in a bin reported by FlyBase).

Table 8.2. Mark and Recapture Analysis

Species	Number of Ramps Identified tAI Method	Number of Ramps Identified w_{ij} Method	Number of Total Sequences	Number of Identical Ramps Captured	Mark and Recapture P-Value
Arabidopsis thaliana	1974	3672	25101	239	0.999
Acidilobus saccharovorans	191	261	1354	63	2.95×10^{-7}
Caenorhabditis elegans	3294	2897	18901	640	9.78×10^{-13}
Drosophila melanogaster	1848	2302	12920	850	2.21×10^{-210}
Saccharomyces cerevisiae	823	767	5649	256	1.66×10^{-47}

The number of ramps extracted using the tAI and w_{ij} options was determined from the total number of gene sequences. The 'Number of Identical Ramps Captured' indicates the number of sequences that contained ramps using both tAI and w_{ij} methods of ramp extraction. The p-value indicates the probability that the amount of overlap ('Number of Identical Ramps Captured') could occur randomly. The phyper function in R was used for these calculations.

Chapter 8 Figures

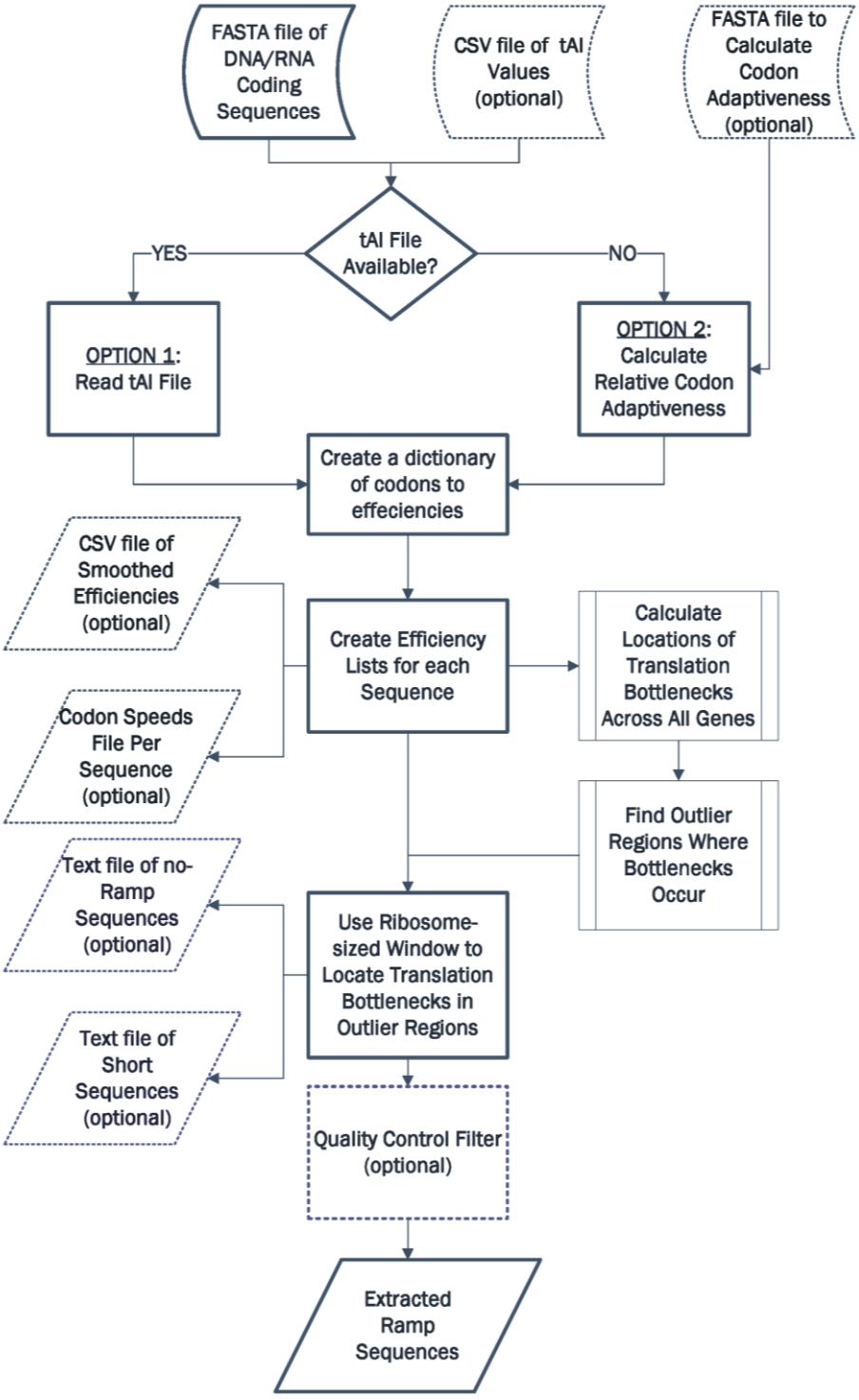


Figure 8.1. ExtRamp Algorithm Flowchart Outlines the algorithm steps, including the optional input/output arguments (dotted lines) and the two options (bold text) for calculating codon efficiencies.

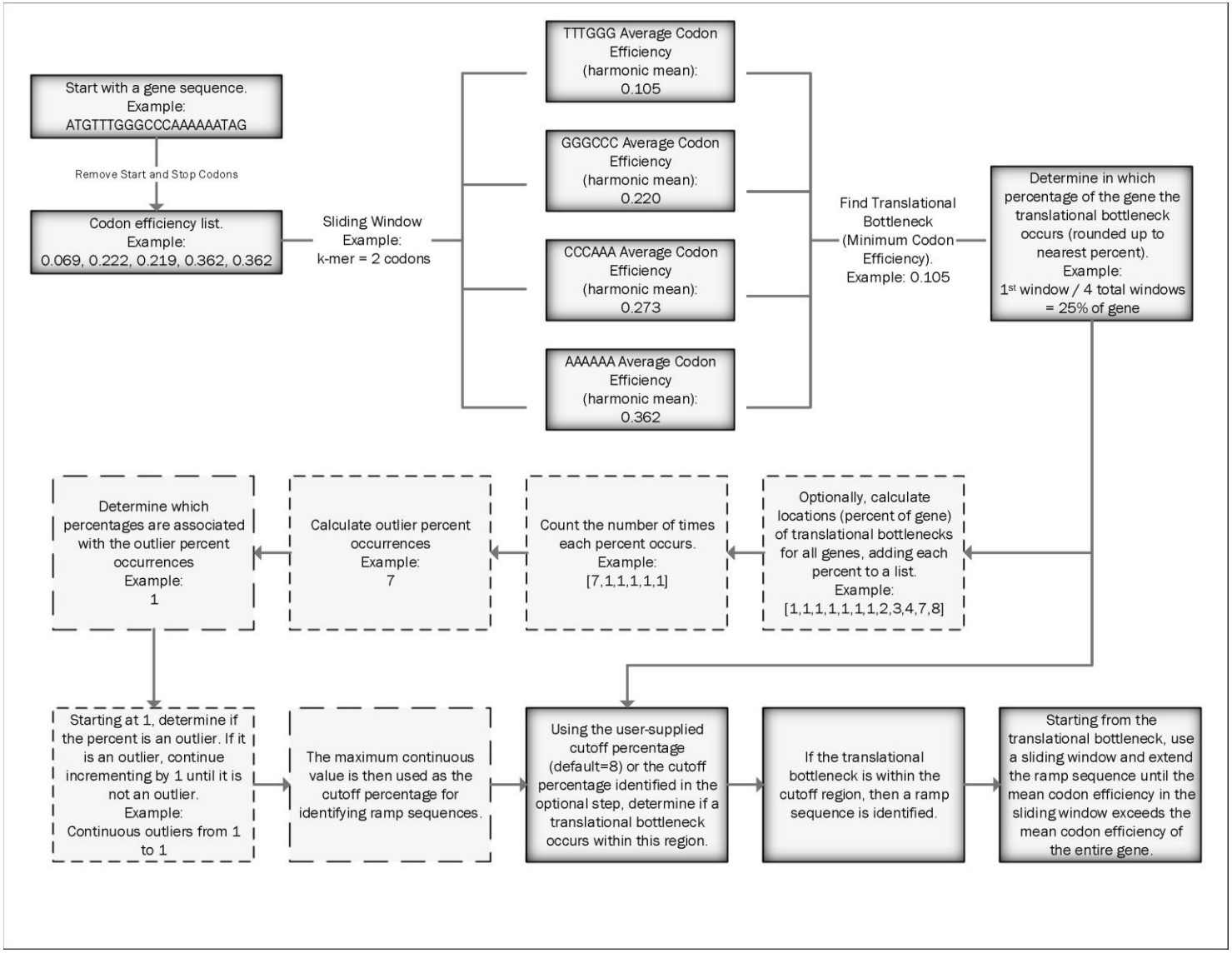


Figure 8.2. Translational Bottleneck Calculation and Usage A detailed example of how translational bottleneck outlier regions are calculated and used to identify ramp sequences.

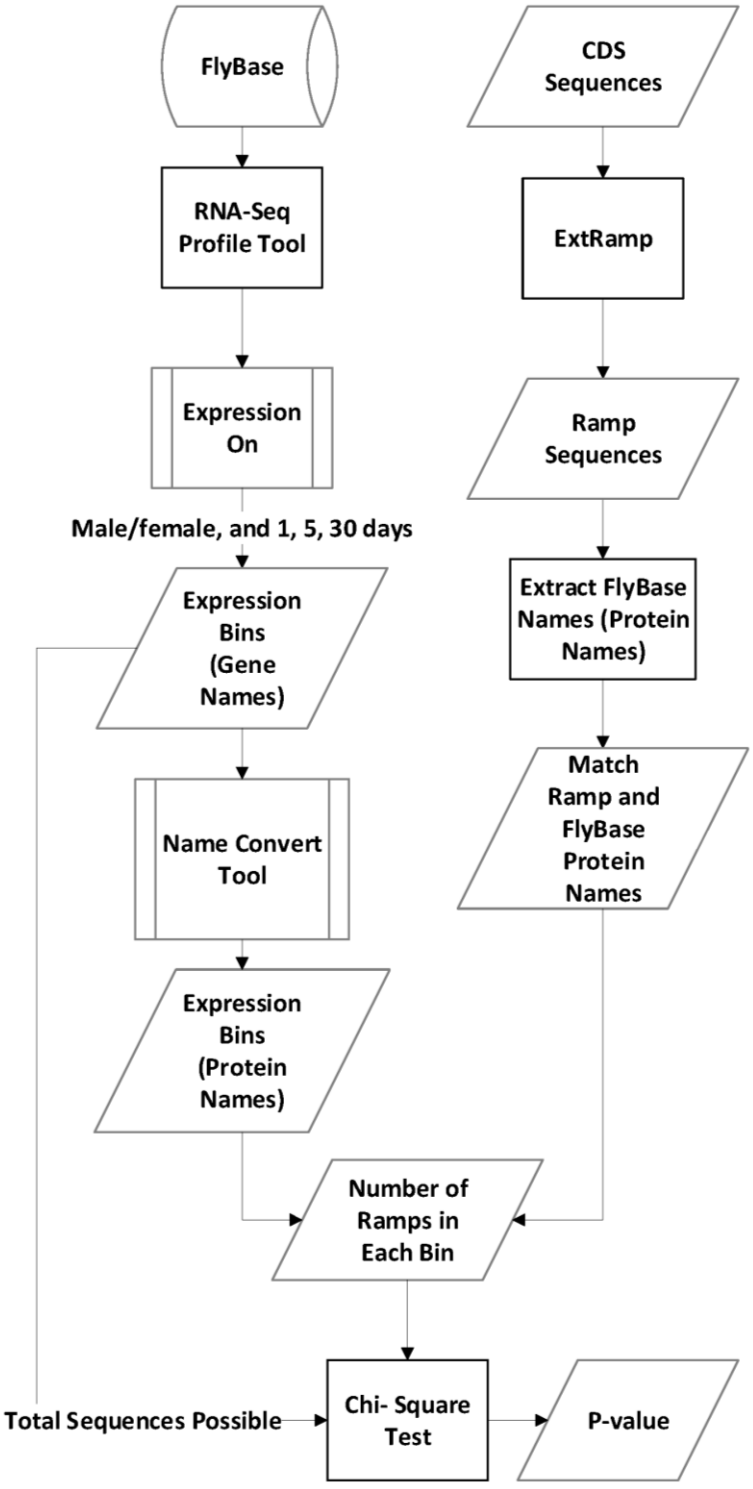


Figure 8.3. FlyBase Analysis Flowchart Data were collected from both the FlyBase database and by running ExtRamp on *Drosophila melanogaster* coding sequences. The number of ramps that fell into each expression level bin was tested with a Chi-squared test to determine if the distribution was random.

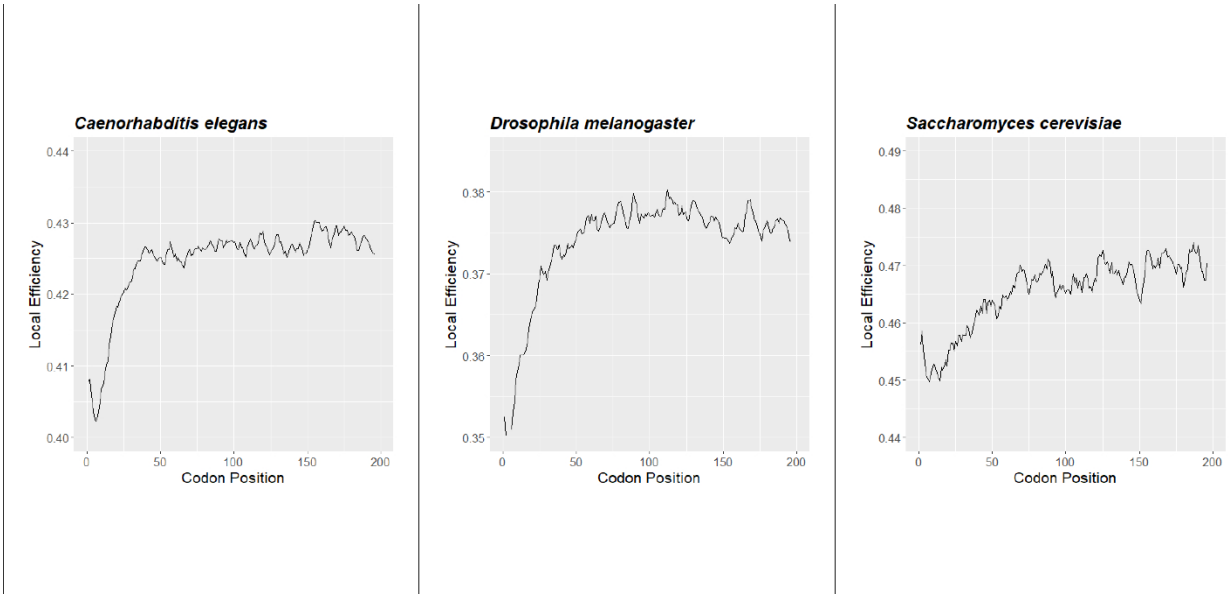


Figure 8.4. Consensus tAI Efficiencies The averaged local tAI values across all CDS regions mapped to the codon position for *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Saccharomyces cerevisiae*. The local efficiency values were smoothed with a window size of four. These graphs are identical to charts reported in *Tuller et al*.(9).

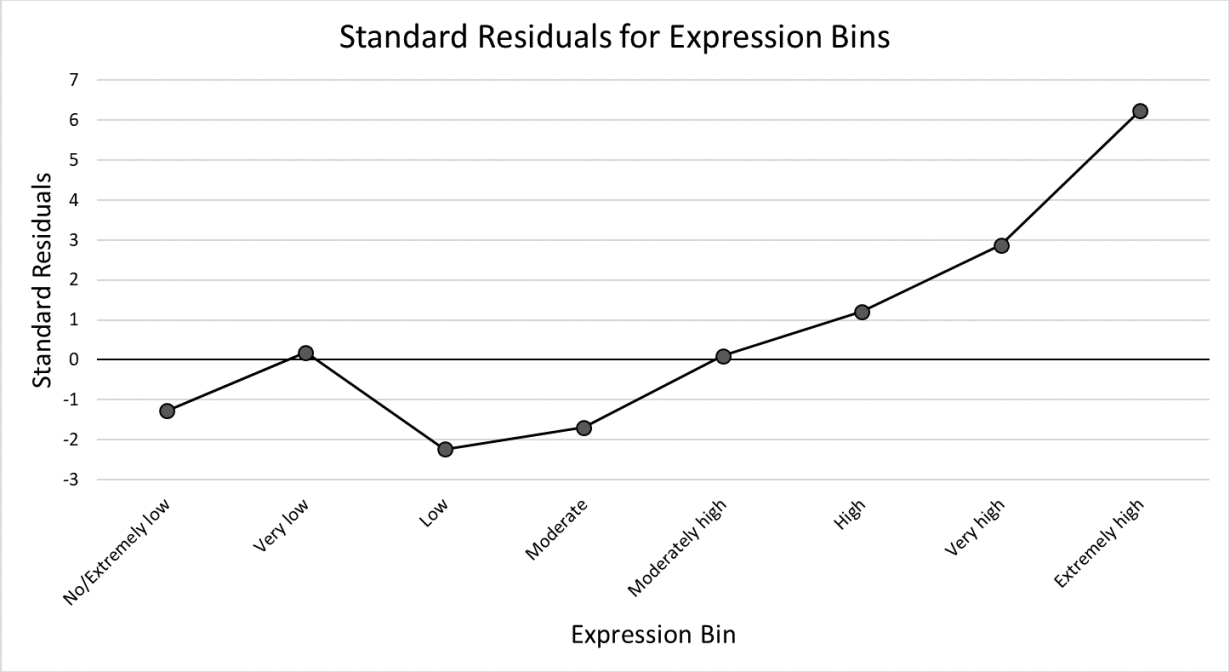


Figure 8.5. Standard Residuals of Expression Bins Using a Chi-squared test, we calculated the standard residuals for each expression bin and plotted these values, ordered from the bin with the lowest expression to the bin with the highest expression.

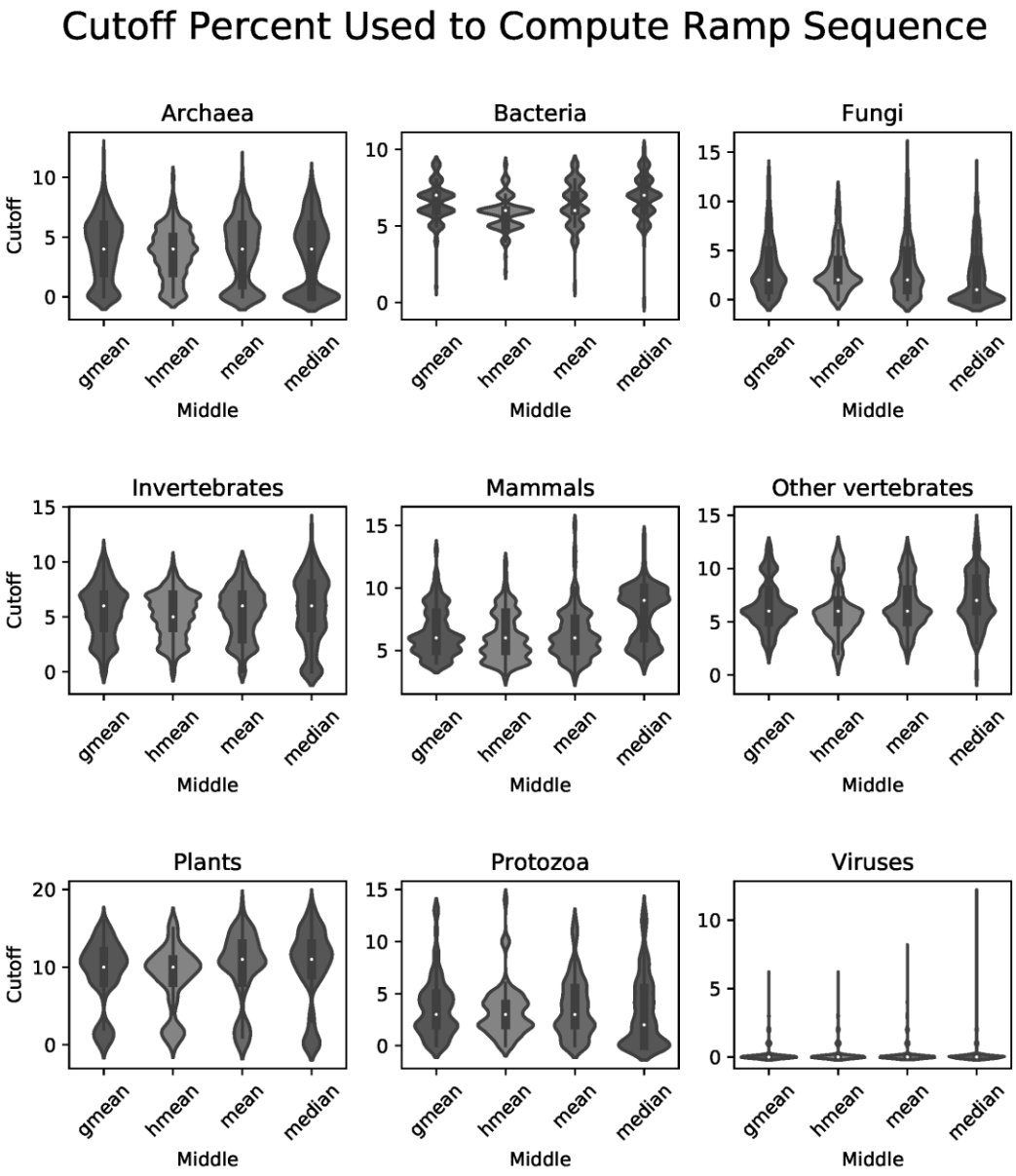


Figure 8.6. Cutoff Percent Used to Compute Ramp Sequence For each taxonomic group, violin plots for each of the four middle values show the cutoff percentages used to compute the ramp sequences. Cutoff percentages are defined as the last consecutive gene region before the number of translation bottlenecks is no longer an outlier, starting from the first percentile (i.e., if the cutoff is 5, then 1-5 are all outlier regions). Each of the nine subplots show means in the following order: geometric mean, harmonic mean, arithmetic mean, and median.

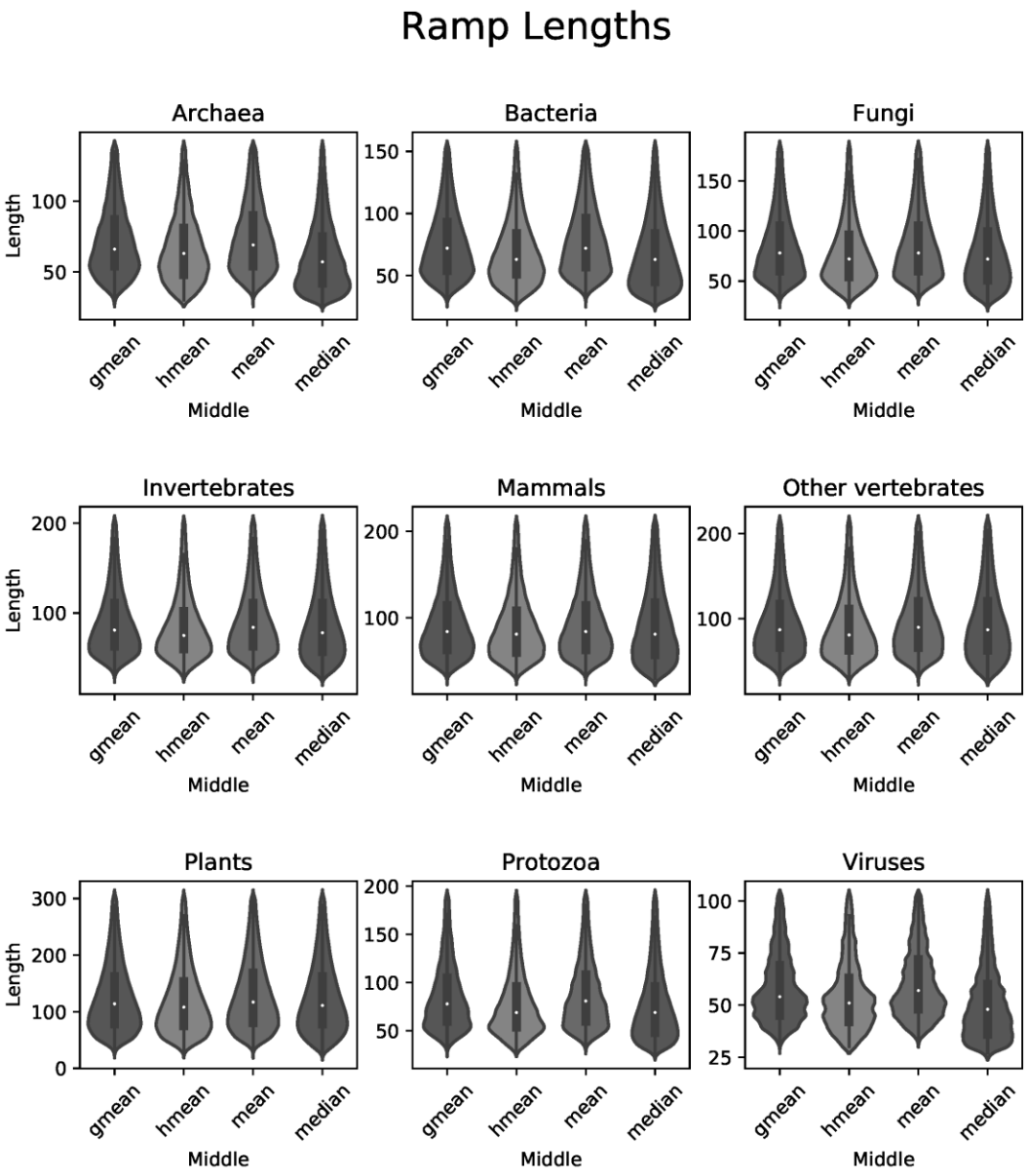


Figure 8.7. Ramp Lengths After removing outliers, we plot the ramp lengths for all ramp sequences in each taxonomic group. Each of the nine subplots show means in the following order: geometric mean, harmonic mean, arithmetic mean, and median.

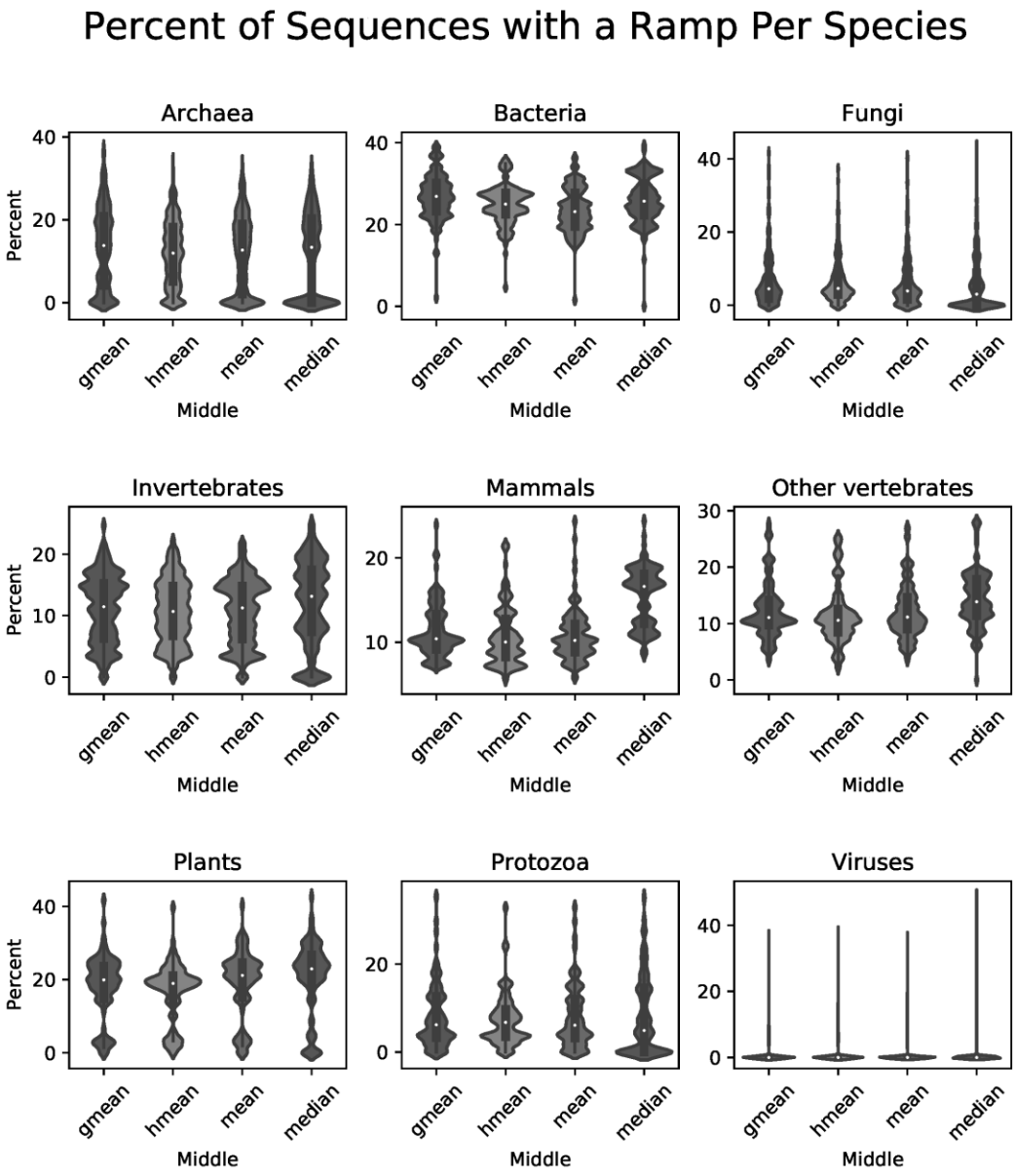


Figure 8.8. Percentage of Sequences with a Ramp Per Species For each taxonomic group, violin plots for each of the four middle values show the percent of sequences in each species that contained a ramp. Each of the nine subplots show means in the following order: geometric mean, harmonic mean, arithmetic mean, and median.

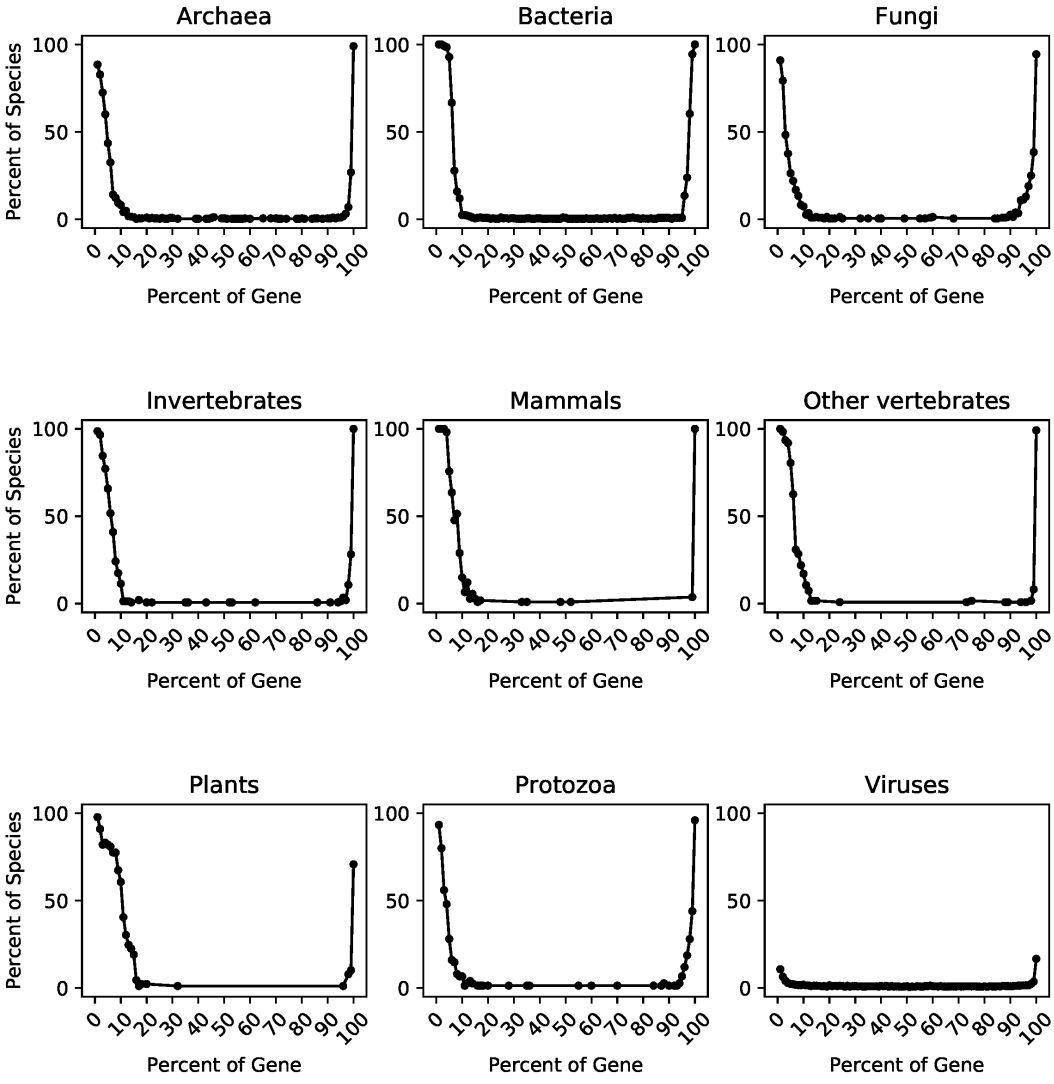


Figure 8.9. Percent of Species with Outliers at Each Gene Percent After dividing each gene into 100 equal parts, we determined where translation bottlenecks occur in the gene. We then identified all outlier regions using the harmonic mean. For each taxonomic group, we counted the number of species with an outlier region at each of the 100 percentiles, and we divided that number by the total number of species in the taxonomic group. We plot these percentiles. When no species had an outlier region, points are not plotted.

Chapter 9

Future Directions

Justin B. Miller

Although the Open Tree of Life (OTL) provides taxonomic relationships for millions of species, the consensus tree hides the confidence level of each reported node, giving researchers a false sense of certainty in reported species relationships. Originally, species were classified based on changes in morphological character states. *Current techniques* evaluate differences in homologous regions from multiple sequence alignments of a few genes across a subset of species. To span more species, the OTL pulls from various studies. Reported species relationships are then combined to form a consensus tree that may or may not be supported by all studies or by all tree reconstruction techniques. Presently, there is a *lack of tools* to adequately asses the accuracy of clades reported in the OTL. For instance, many intractable clades are depicted in the OTL as resolved because alternative hypotheses are given less weight. Since different tree structures give different hypotheses of evolution, all studies with a phylogenetic component (e.g., endangered species classifications, trait analyses, crop studies, drug therapeutics, etc.) are subject to the accuracy of the tree structure. *There exists a critical need to ensure the accuracy of proposed phylogenetic relationships and to provide researchers with alternative tree hypotheses.* Without providing support values for consensus trees depicted in the OTL, studies utilizing this tool will continue to operate on a potentially biased evolutionary hypothesis where controversial nodes are given equal weight as more established nodes.

Our *long-term goal* is to provide a novel method for evaluating species trees from different studies by establishing an open-source database that connects to the OTL and stores support values for each node. To the extent that we accomplish our goals, our *overall objective* is to allow researchers to evaluate the accuracy of the OTL at each clade based on information from both supporting and alternative tree hypotheses. Additionally, we expect to *use techniques that we have already developed* that use codon usage bias to span larger groups of species and resolve controversial species relationships. Our *central hypothesis* is that providing support values for each clade will enable researchers to determine which species can be evaluated using one evolutionary model and which species should be evaluated using multiple evolutionary models. The *rationale* underlying the proposed research is that *some species relationships are more established than others and providing easy access to alternative tree hypotheses enables researchers to evaluate their data on different models of evolution.* In addition to developing several techniques to evaluate species relatedness across thousands of species, we are particularly *well prepared* to address these issues because we have published or submitted several papers directly related to this research and we have extensive experience managing databases. We plan to pursue the following three *specific aims* to attain our overall objective:

1. **Develop a more accurate method to recover phylogenies from larger groups of species.**
   Our *working hypothesis* is that properties of codon usage bias can accurately recover species relationships across thousands of species faster than traditional techniques.
2. **Provide a framework for evaluating nodal support for different tree topologies.**
   We *hypothesize* that an open-source database can store normalized support values (e.g., bootstrap, posterior probabilities, homologous changes, distances) for tree topologies reported in various studies.

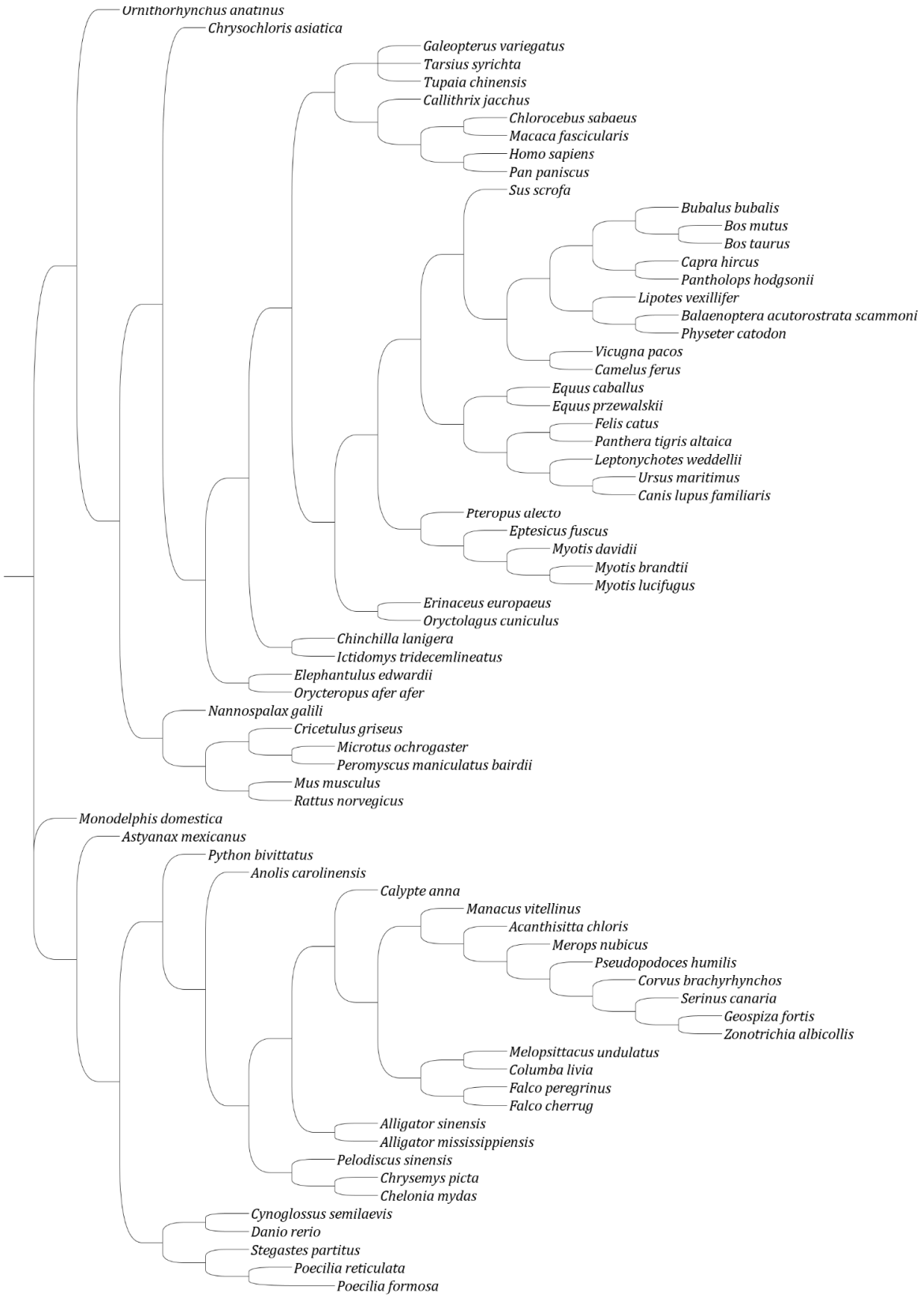
3. **Systematically identify controversial nodes and present alternative tree hypotheses.**

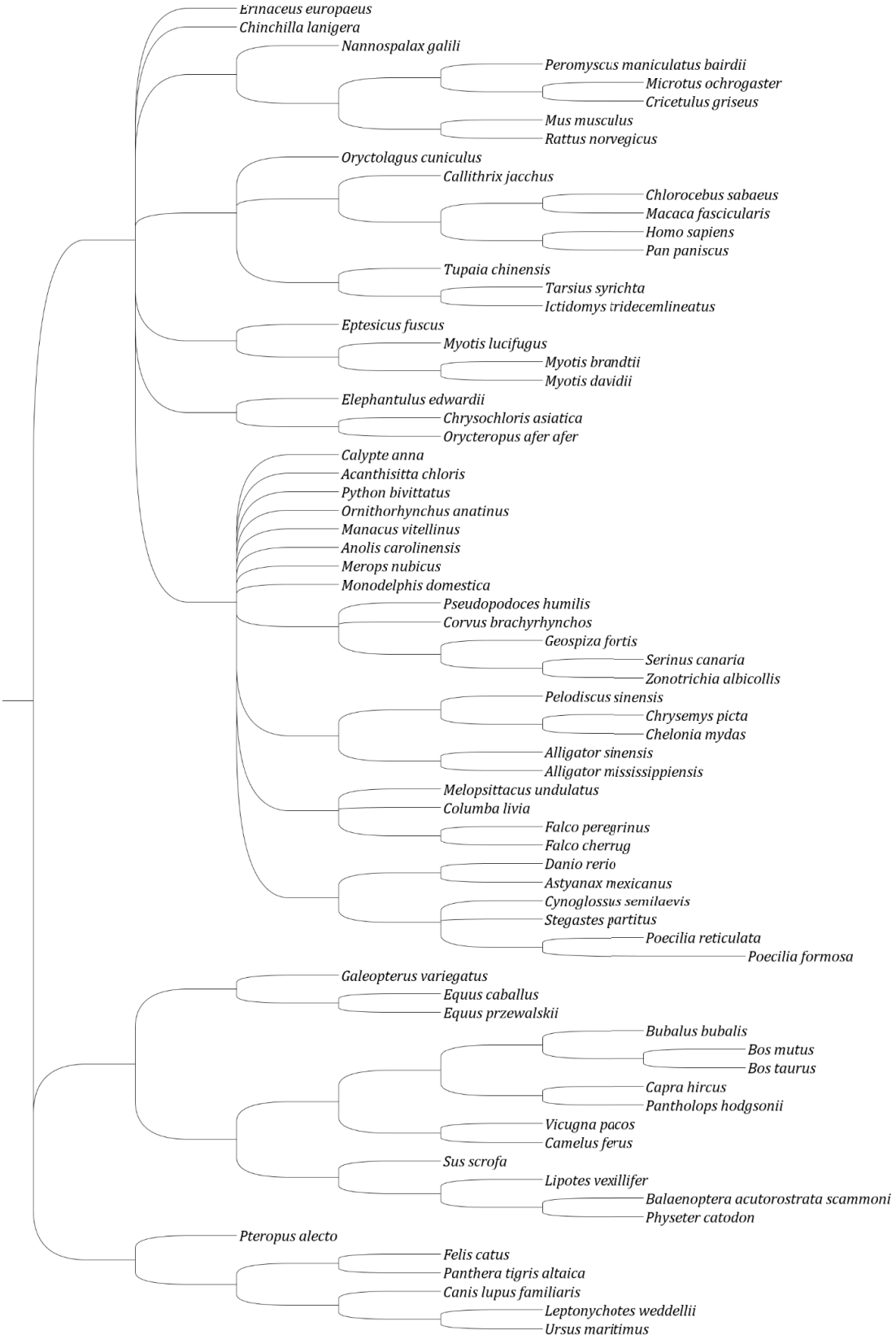
   We *hypothesize* that competing tree hypotheses can be identified and ranked based on nodal support.

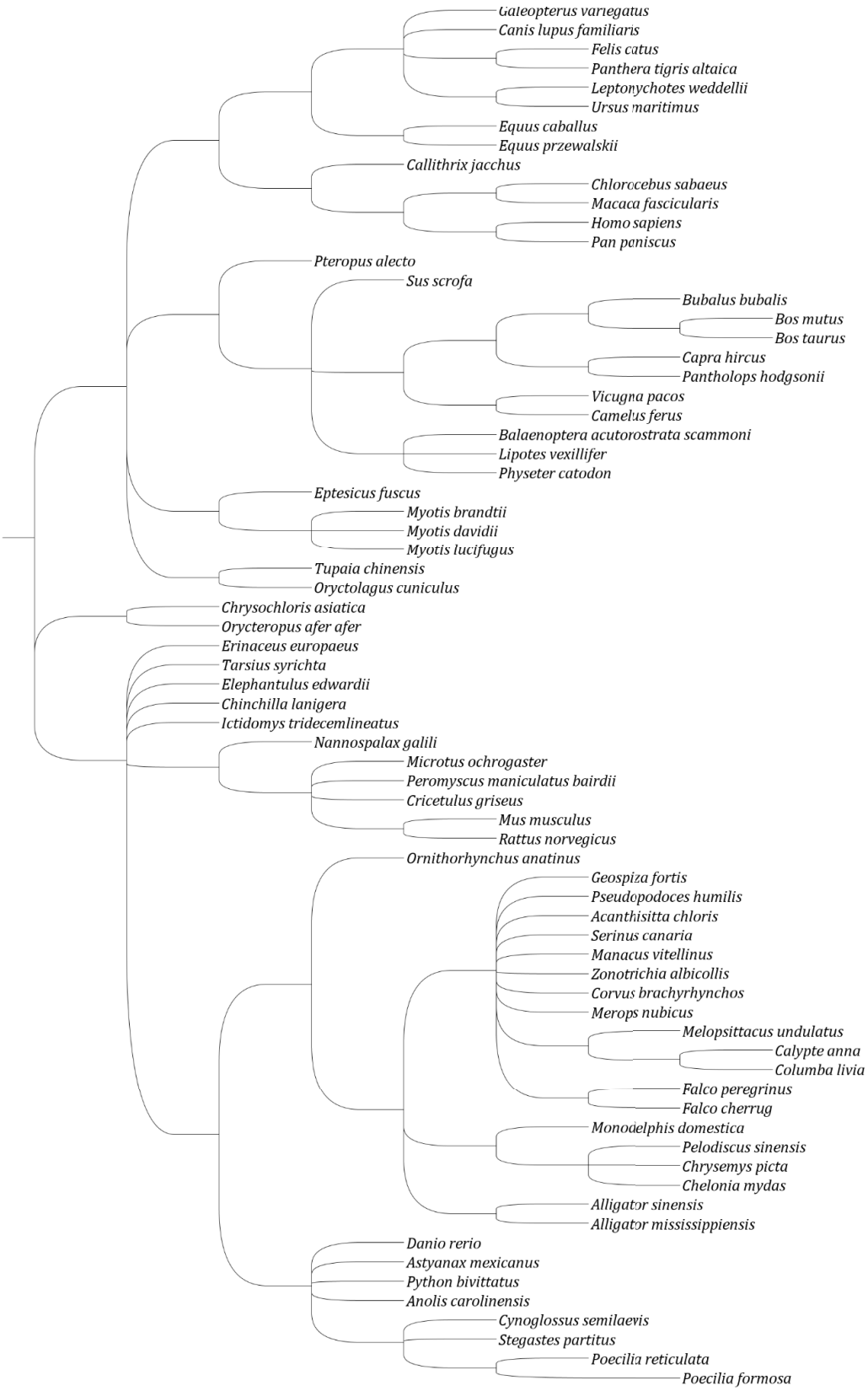
Our *expected outcomes* are to branch larger groups of species using a single method (aim 1), allow researchers to evaluate nodal support of the consensus tree (aim 2), and provide researchers with ranked alternative tree hypotheses (aim 3). These aims collectively attain the overall objective of providing an environment to assess the quality of each species relationship reported in the OTL. Collectively, this research represents a *major vertical advancement* in the field of phylogenetics by providing researchers with a better understanding and more efficient means of evaluating different tree topologies.

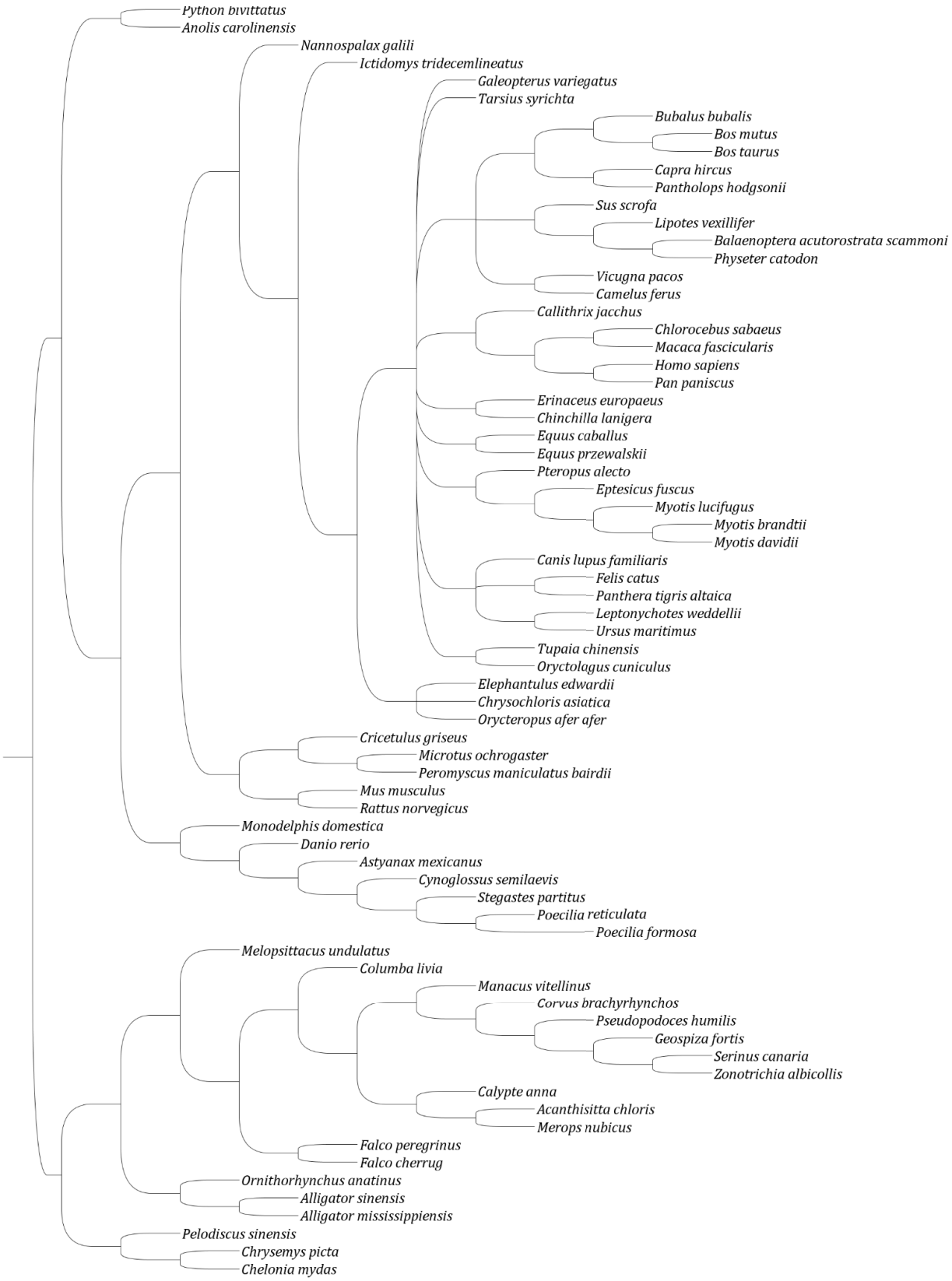
Appendix 1: Supplementary Figures and Tables for Chapter 2

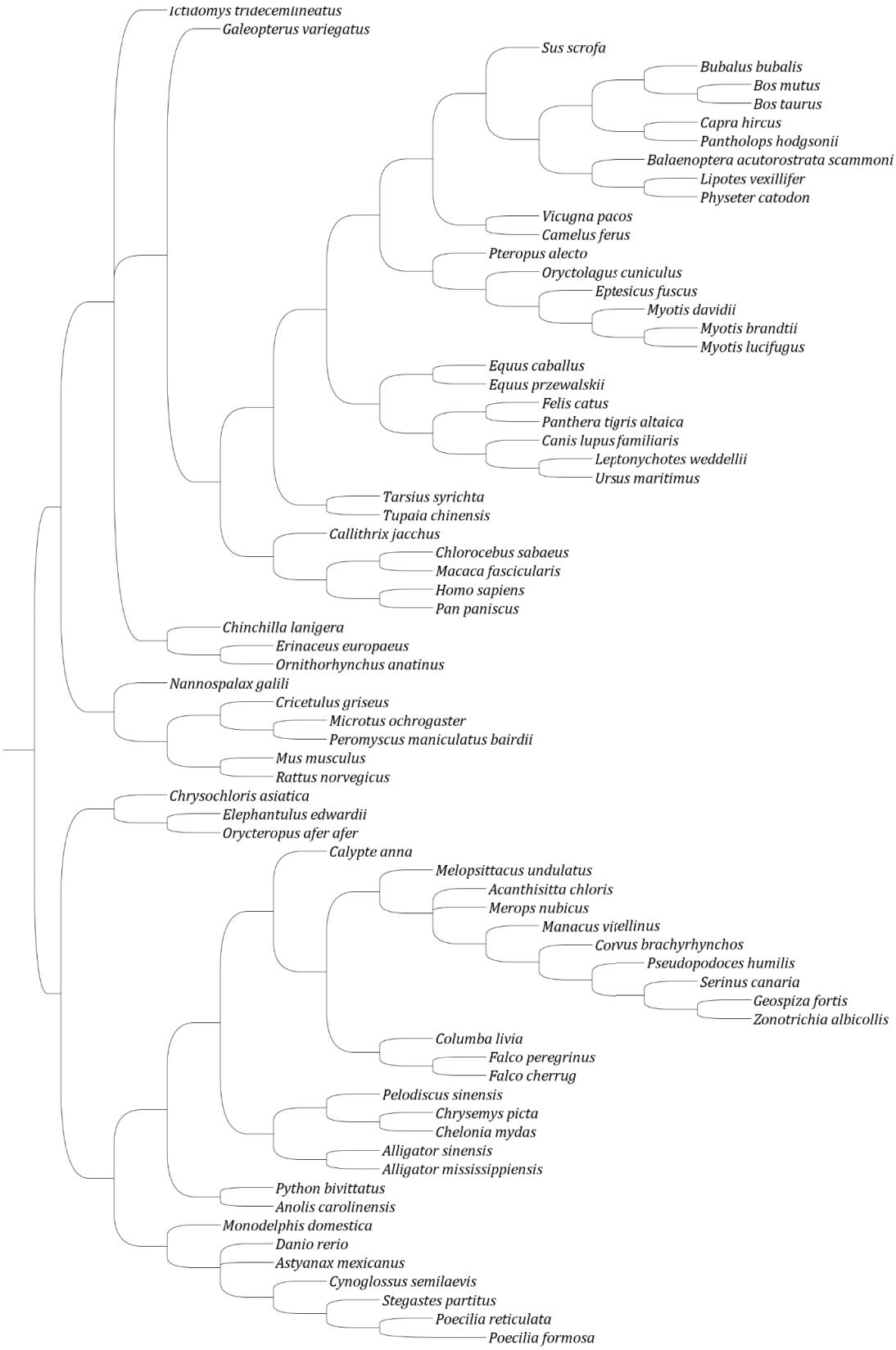
Supplemental Figure S1, Chapter 2: 64 phylogenetic strict consensus trees recovered using TNT. Each phylogeny represents the codon usage and non-usage of a given codon, coded as described in Figure 1. Each phylogeny has a header, which shows the codon used in tree reconstruction. The stop codons TAA and TGA depicted the strongest phylogenetic signal when compared with the Open Tree of Life project.

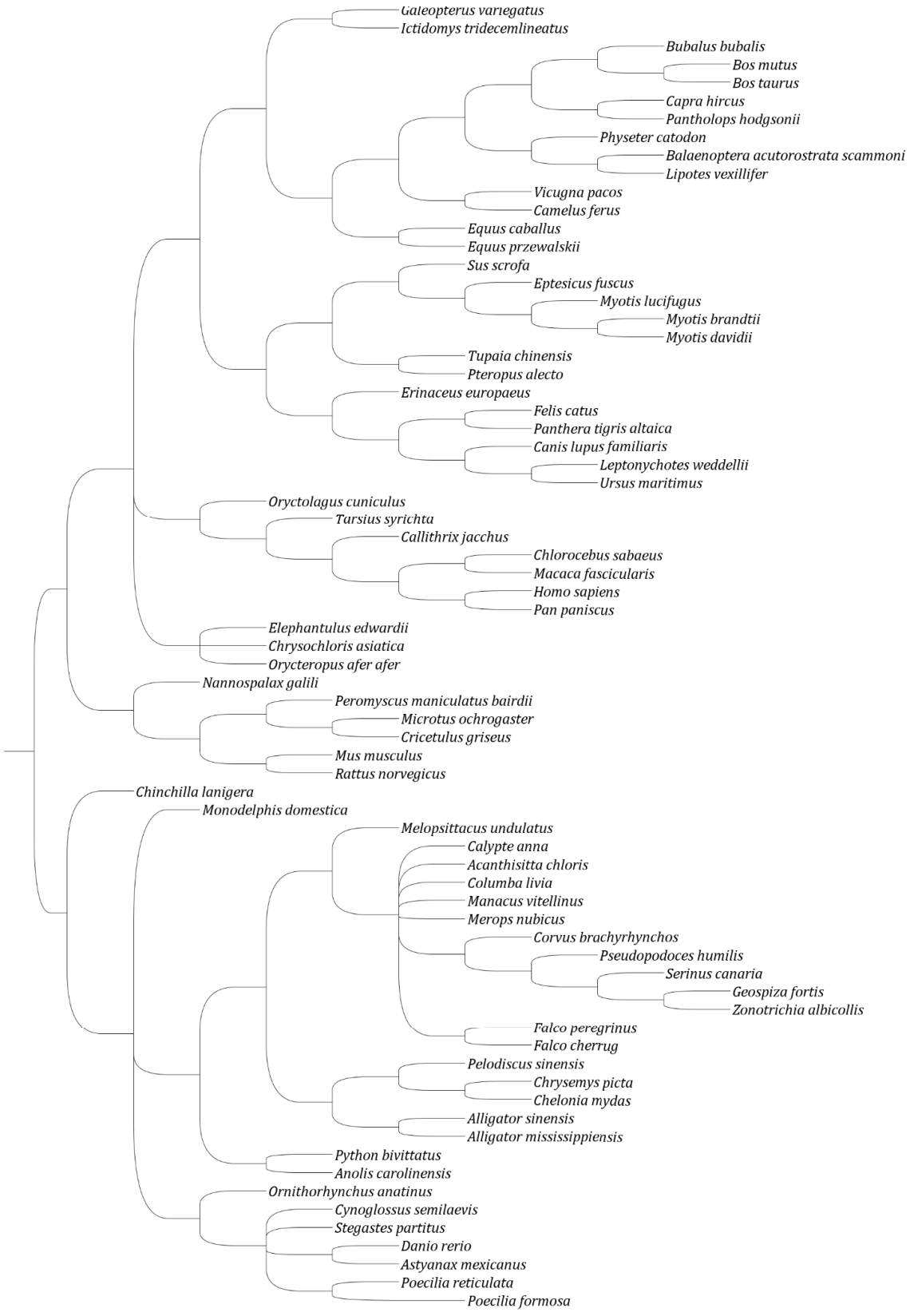

Ornithorhynchus anatinus
Chrysochloris asiatica
Galeopterus variegatus
Tarsius syrichta
Tupaia chinensis
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Canis lupus familiaris
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Erinaceus europaeus
Oryctolagus cuniculus
Chinchilla lanigera
Ictidomys tridecemlineatus
Elephantulus edwardii
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Monodelphis domestica
Astyanax mexicanus
Python bivittatus
Anolis carolinensis
Calypte anna
Manacus vitellinus
Acanthisitta chloris
Merops nubicus
Pseudopodoces humilis
Corvus brachyrhynchos
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Melopsittacus undulatus
Columba livia
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Cynoglossus semilaevis
Danio rerio
Stegastes partitus
Poecilia reticulata
Poecilia formosa

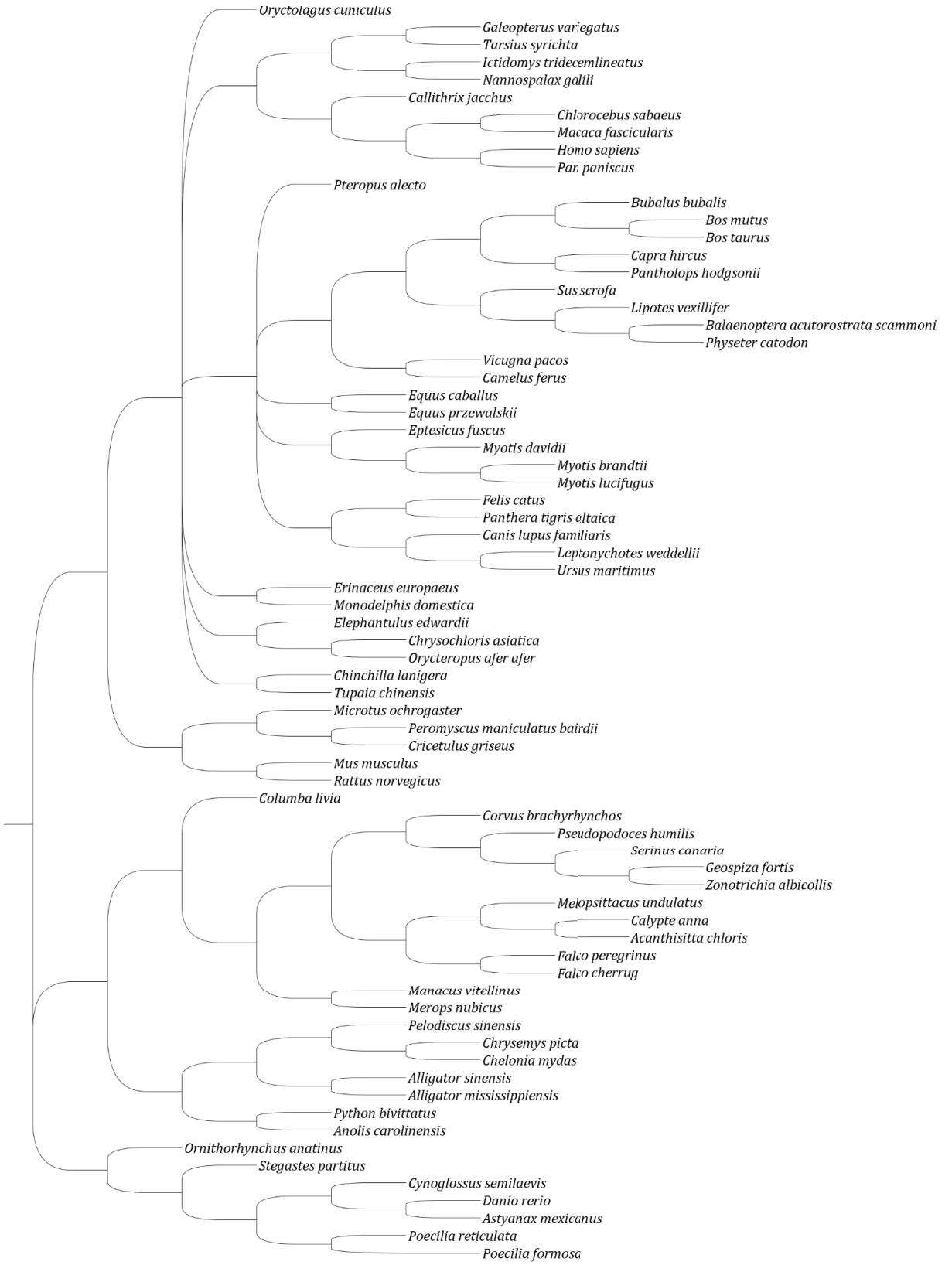

Erinaceus europaeus
Chinchilla lanigera
Nannospalax galili
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Oryctolagus cuniculus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Tupaia chinensis
Tarsius syrichta
Ictidomys tridecemlineatus
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Calypte anna
Acanthisitta chloris
Python bivittatus
Ornithorhynchus anatinus
Manacus vitellinus
Anolis carolinensis
Merops nubicus
Monodelphis domestica
Pseudopodoces humilis
Corvus brachyrhynchos
Geospiza fortis
Serinus canaria
Zonotrichia albicollis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Melopsittacus undulatus
Columba livia
Falco peregrinus
Falco cherrug
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Galeopterus variegatus
Equus caballus
Equus przewalskii
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Vicugna pacos
Camelus ferus
Sus scrofa
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Pteropus alecto
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus

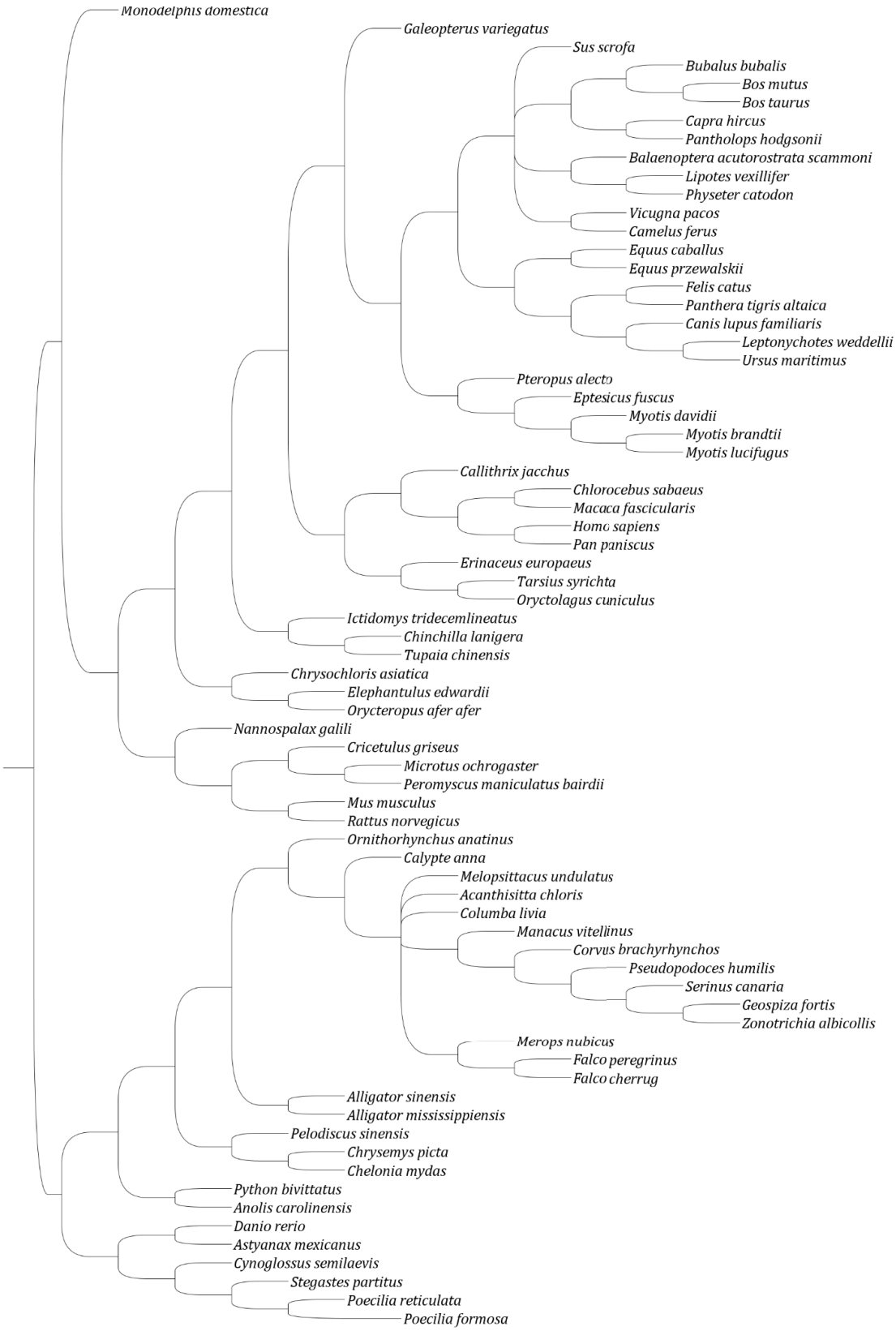

Galeopterus variegatus
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Equus caballus
Equus przewalskii
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Pteropus alecto
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Vicugna pacos
Camelus ferus
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Eptesicus fuscus
Myotis brandtii
Myotis davidii
Myotis lucifugus
Tupaia chinensis
Oryctolagus cuniculus
Chrysochloris asiatica
Orycteropus afer afer
Erinaceus europaeus
Tarsius syrichta
Elephantulus edwardii
Chinchilla lanigera
Ictidomys tridecemlineatus
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Ornithorhynchus anatinus
Geospiza fortis
Pseudopodoces humilis
Acanthisitta chloris
Serinus canaria
Manacus vitellinus
Zonotrichia albicollis
Corvus brachyrhynchos
Merops nubicus
Melopsittacus undulatus
Calypte anna
Columba livia
Falco peregrinus
Falco cherrug
Monodelphis domestica
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Danio rerio
Astyanax mexicanus
Python bivittatus
Anolis carolinensis
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

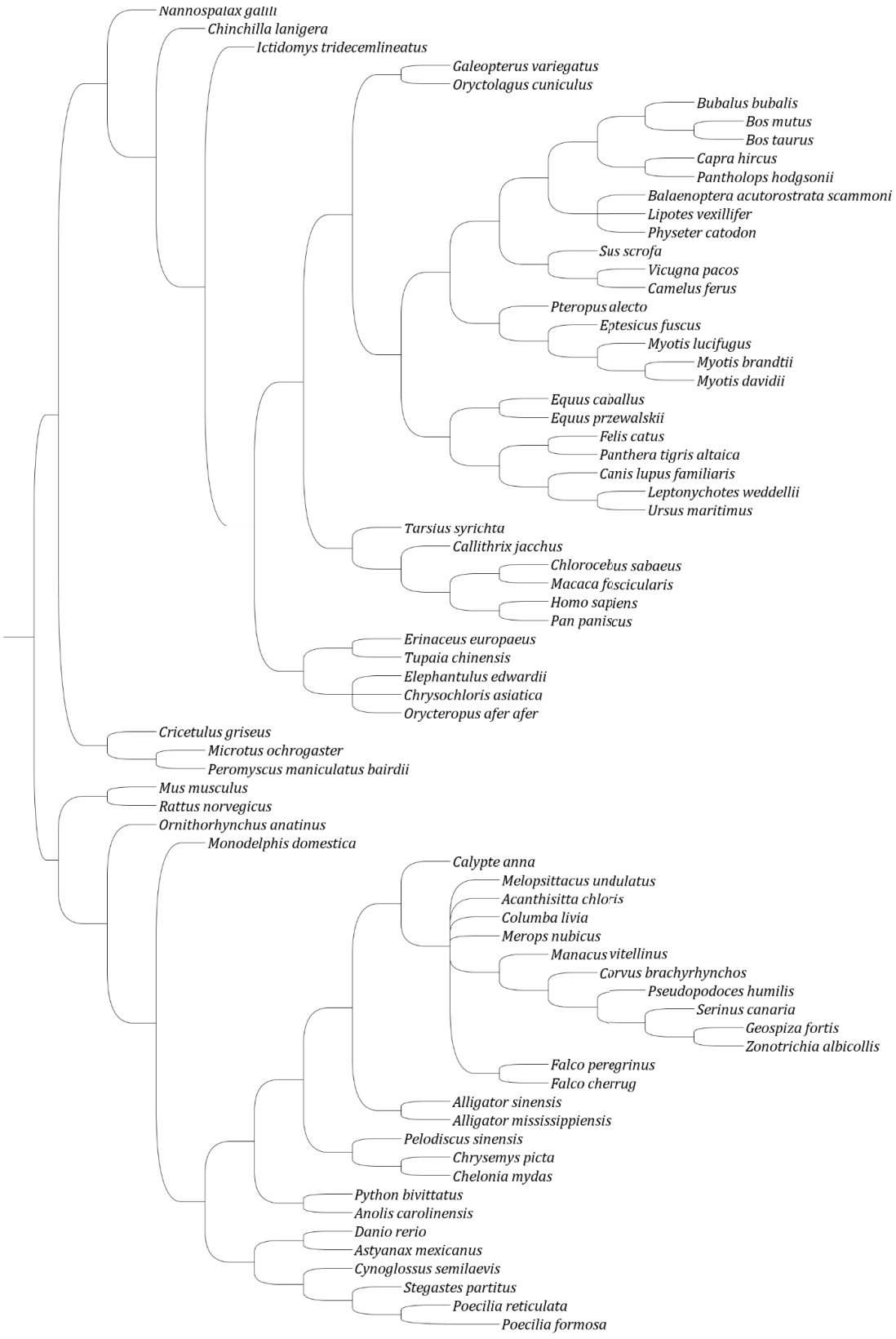

Python bivittatus
Anolis carolinensis
Nannospalax galili
Ictidomys tridecemlineatus
Galeopterus variegatus
Tarsius syrichta
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Sus scrofa
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Erinaceus europaeus
Chinchilla lanigera
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Tupaia chinensis
Oryctolagus cuniculus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Monodelphis domestica
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Melopsittacus undulatus
Columba livia
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Geospiza fortis
Serinus canaria
Zonotrichia albicollis
Calypte anna
Acanthisitta chloris
Merops nubicus
Falco peregrinus
Falco cherrug
Ornithorhynchus anatinus
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas

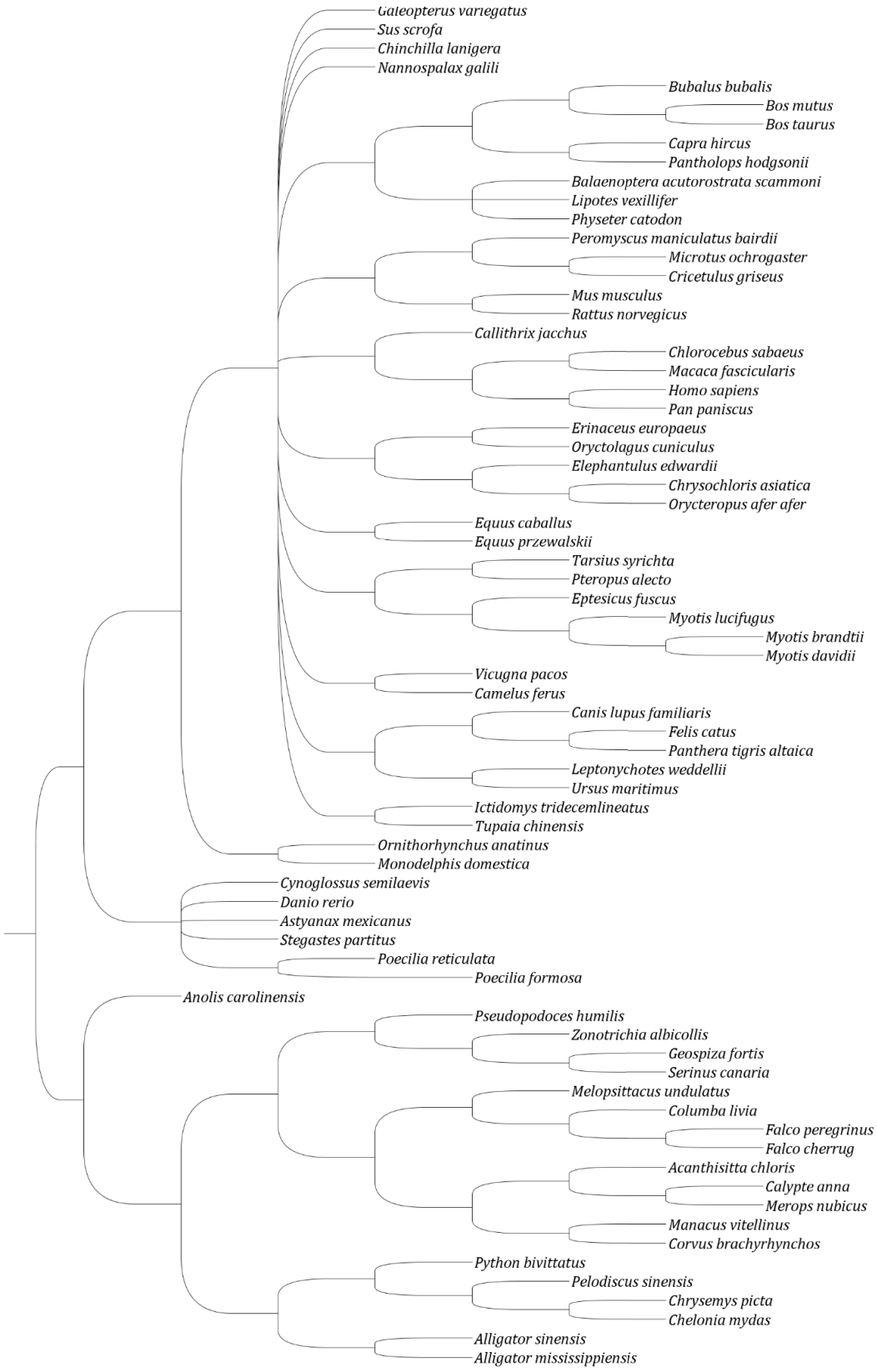

Ictidomys tridecemlineatus
Galeopterus variegatus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Vicugna pacos
Camelus ferus
Pteropus alecto
Oryctolagus cuniculus
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Tarsius syrichta
Tupaia chinensis
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Chinchilla lanigera
Erinaceus europaeus
Ornithorhynchus anatinus
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Calypte anna
Melopsittacus undulatus
Acanthisitta chloris
Merops nubicus
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Columba livia
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Monodelphis domestica
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa


Galeopterus variegatus
Ictidomys tridecemlineatus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Physeter catodon
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Sus scrofa
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Tupaia chinensis
Pteropus alecto
Erinaceus europaeus
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Oryctolagus cuniculus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Nannospalax galili
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Chinchilla lanigera
Monodelphis domestica
Melopsittacus undulatus
Calypte anna
Acanthisitta chloris
Columba livia
Manacus vitellinus
Merops nubicus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Ornithorhynchus anatinus
Cynoglossus semilaevis
Stegastes partitus
Danio rerio
Astyanax mexicanus
Poecilia reticulata
Poecilia formosa

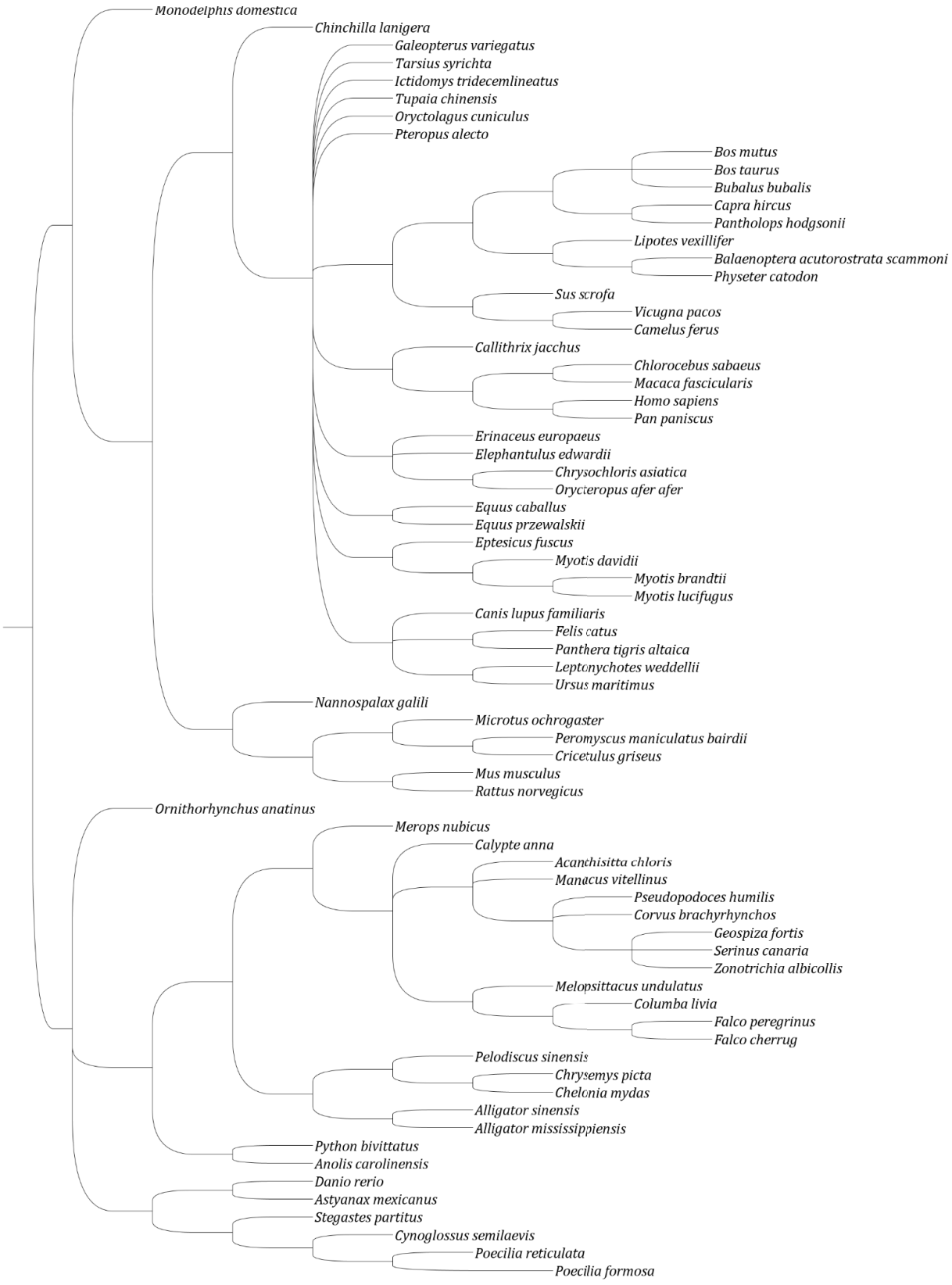

Oryctolagus cuniculus
Galeopterus variegatus
Tarsius syrichta
Ictidomys tridecemlineatus
Nannospalax galili
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Pteropus alecto
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Sus scrofa
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Erinaceus europaeus
Monodelphis domestica
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Chinchilla lanigera
Tupaia chinensis
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Columba livia
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Melopsittacus undulatus
Calypte anna
Acanthisitta chloris
Falco peregrinus
Falco cherrug
Manacus vitellinus
Merops nubicus
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Ornithorhynchus anatinus
Stegastes partitus
Cynoglossus semilaevis
Danio rerio
Astyanax mexicanus
Poecilia reticulata
Poecilia formosa

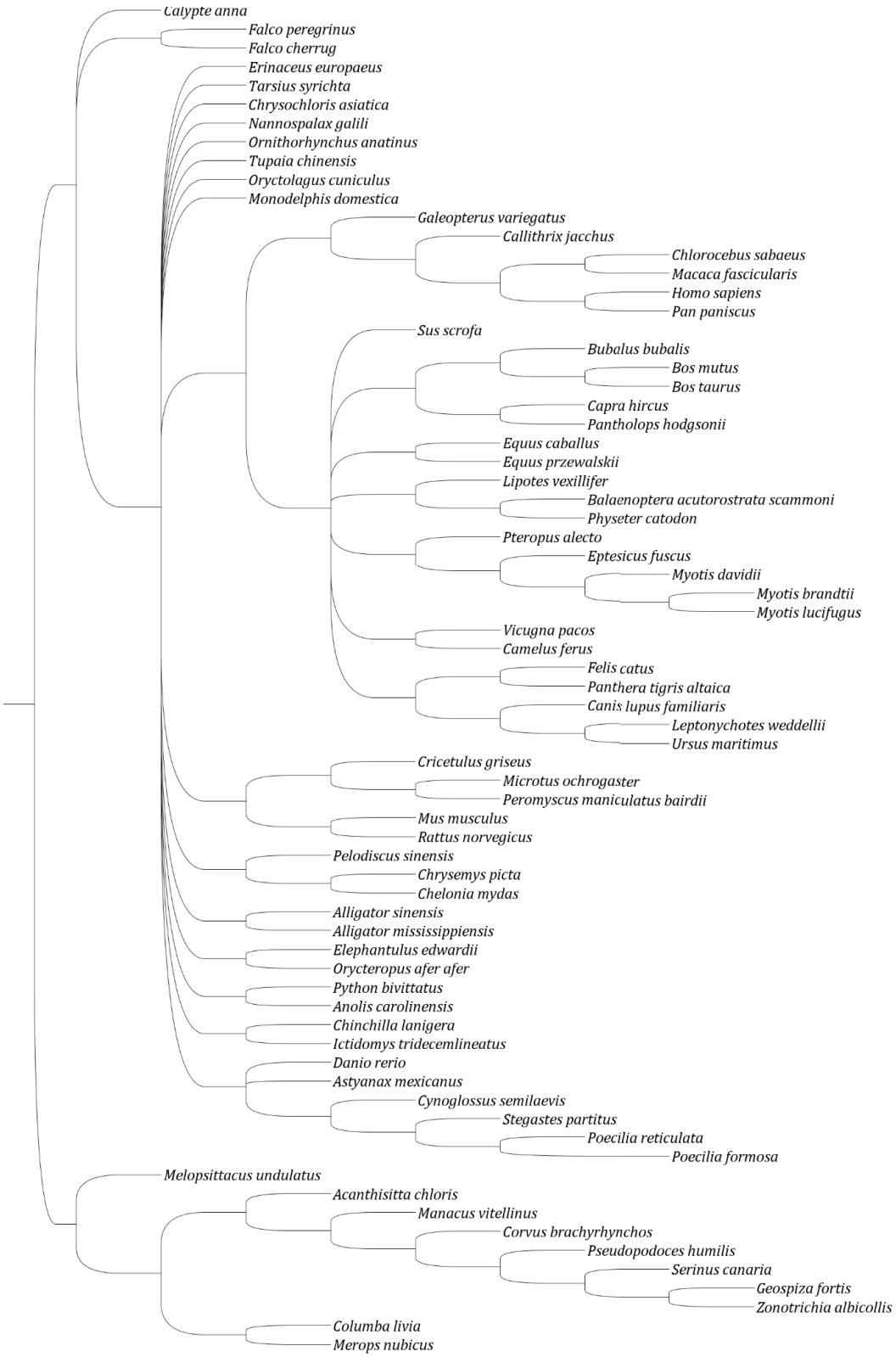

Monodelphis domestica
Galeopterus variegatus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Erinaceus europaeus
Tarsius syrichta
Oryctolagus cuniculus
Ictidomys tridecemlineatus
Chinchilla lanigera
Tupaia chinensis
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Ornithorhynchus anatinus
Calypte anna
Melopsittacus undulatus
Acanthisitta chloris
Columba livia
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Merops nubicus
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

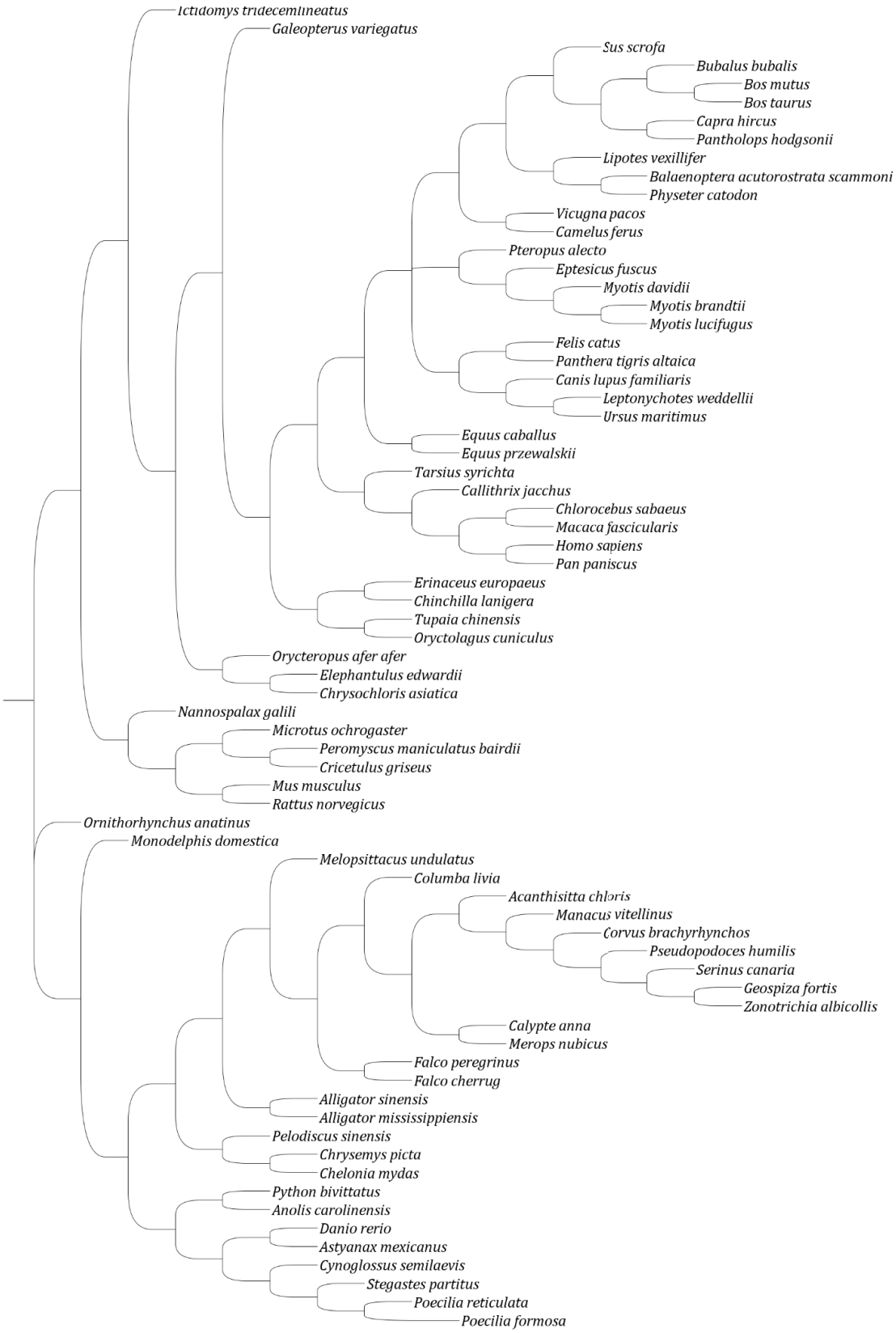

Nannospalax galili
Chinchilla lanigera
Ictidomys tridecemlineatus
Galeopterus variegatus
Oryctolagus cuniculus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Erinaceus europaeus
Tupaia chinensis
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Ornithorhynchus anatinus
Monodelphis domestica
Calypte anna
Melopsittacus undulatus
Acanthisitta chloris
Columba livia
Merops nubicus
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

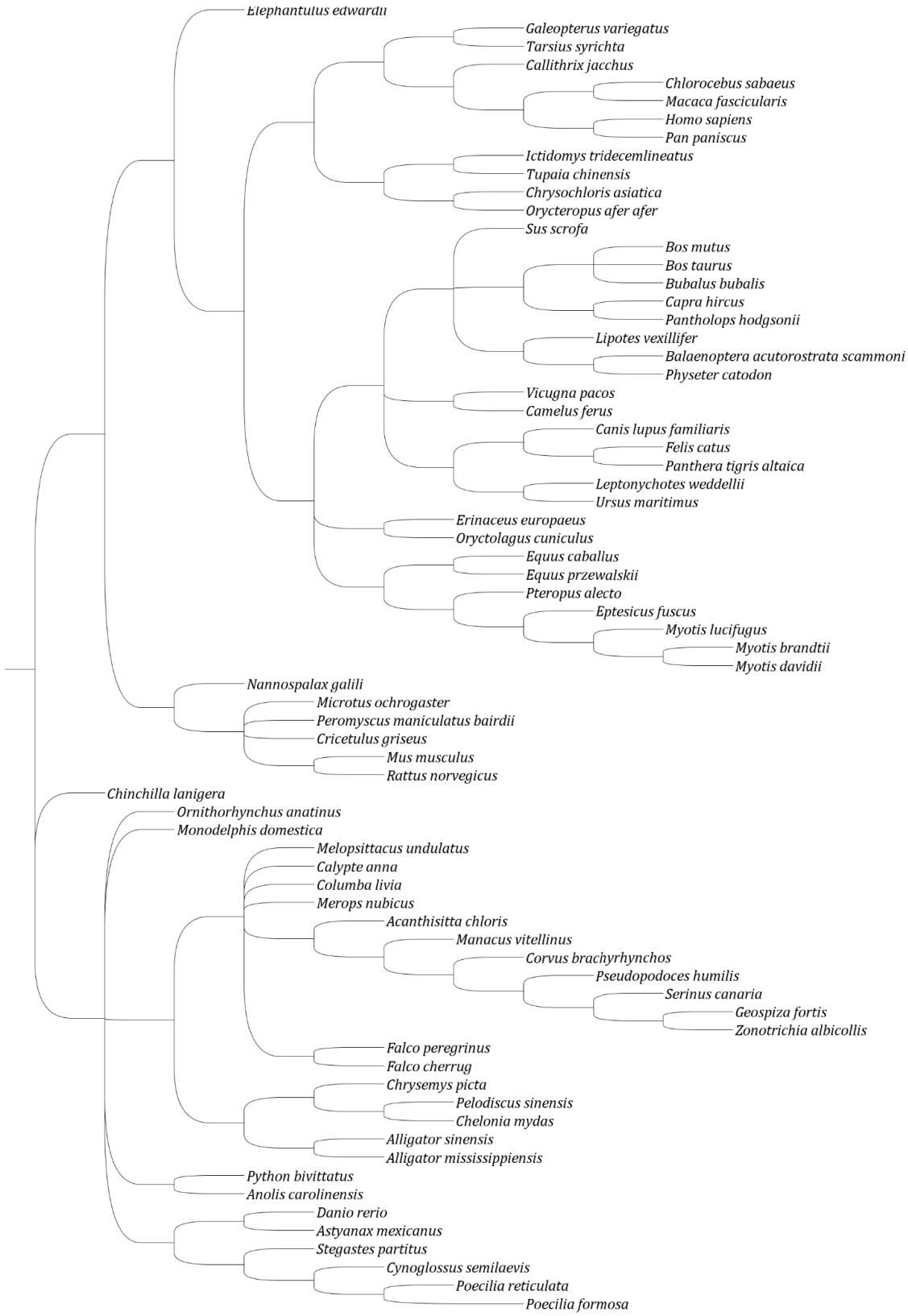

Galeopterus variegatus
Sus scrofa
Chinchilla lanigera
Nannospalax galili
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Erinaceus europaeus
Oryctolagus cuniculus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Equus caballus
Equus przewalskii
Tarsius syrichta
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Vicugna pacos
Camelus ferus
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Ictidomys tridecemlineatus
Tupaia chinensis
Ornithorhynchus anatinus
Monodelphis domestica
Cynoglossus semilaevis
Danio rerio
Astyanax mexicanus
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Anolis carolinensis
Pseudopodoces humilis
Zonotrichia albicollis
Geospiza fortis
Serinus canaria
Melopsittacus undulatus
Columba livia
Falco peregrinus
Falco cherrug
Acanthisitta chloris
Calypte anna
Merops nubicus
Manacus vitellinus
Corvus brachyrhynchos
Python bivittatus
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis

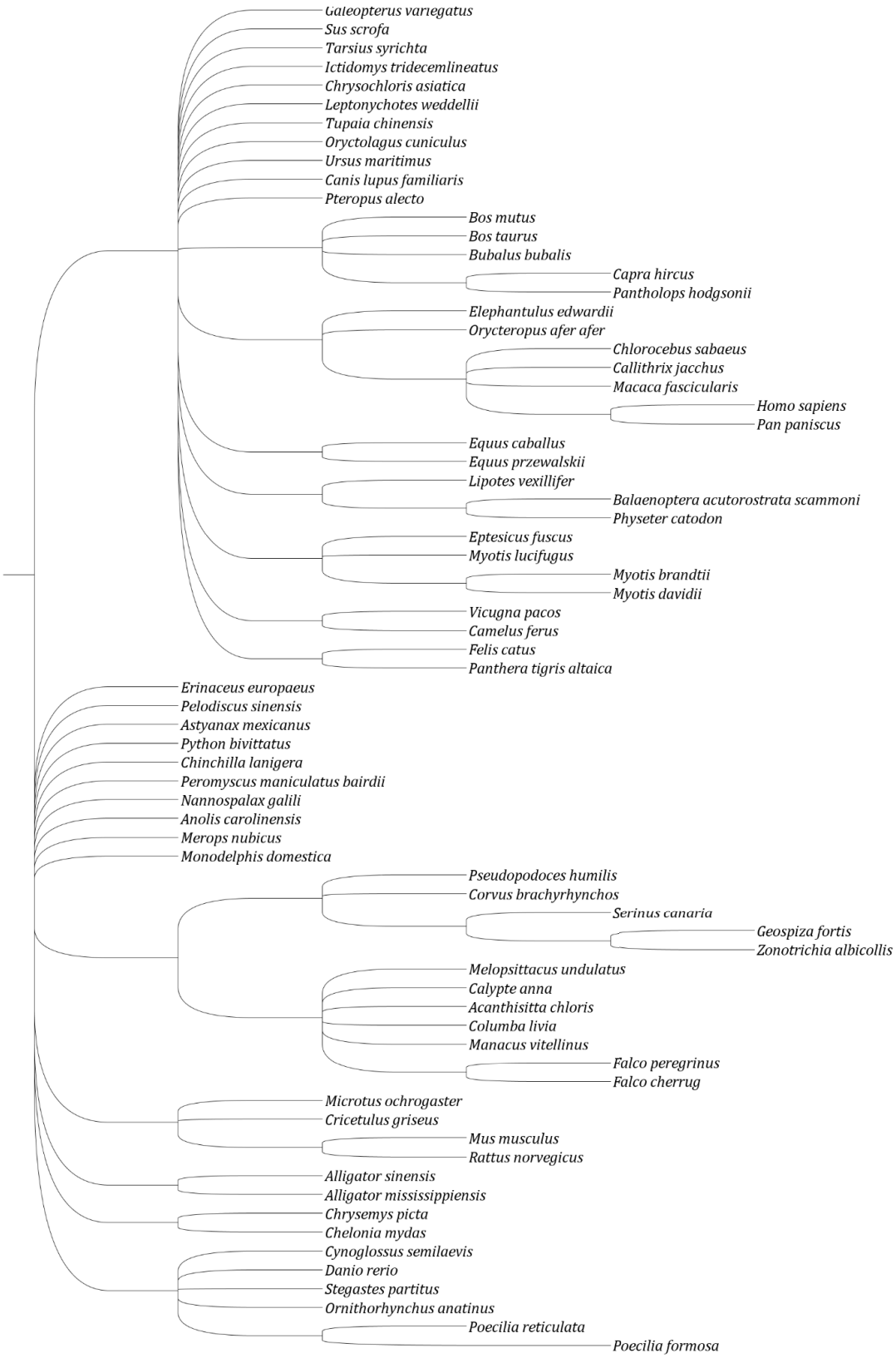
AGG

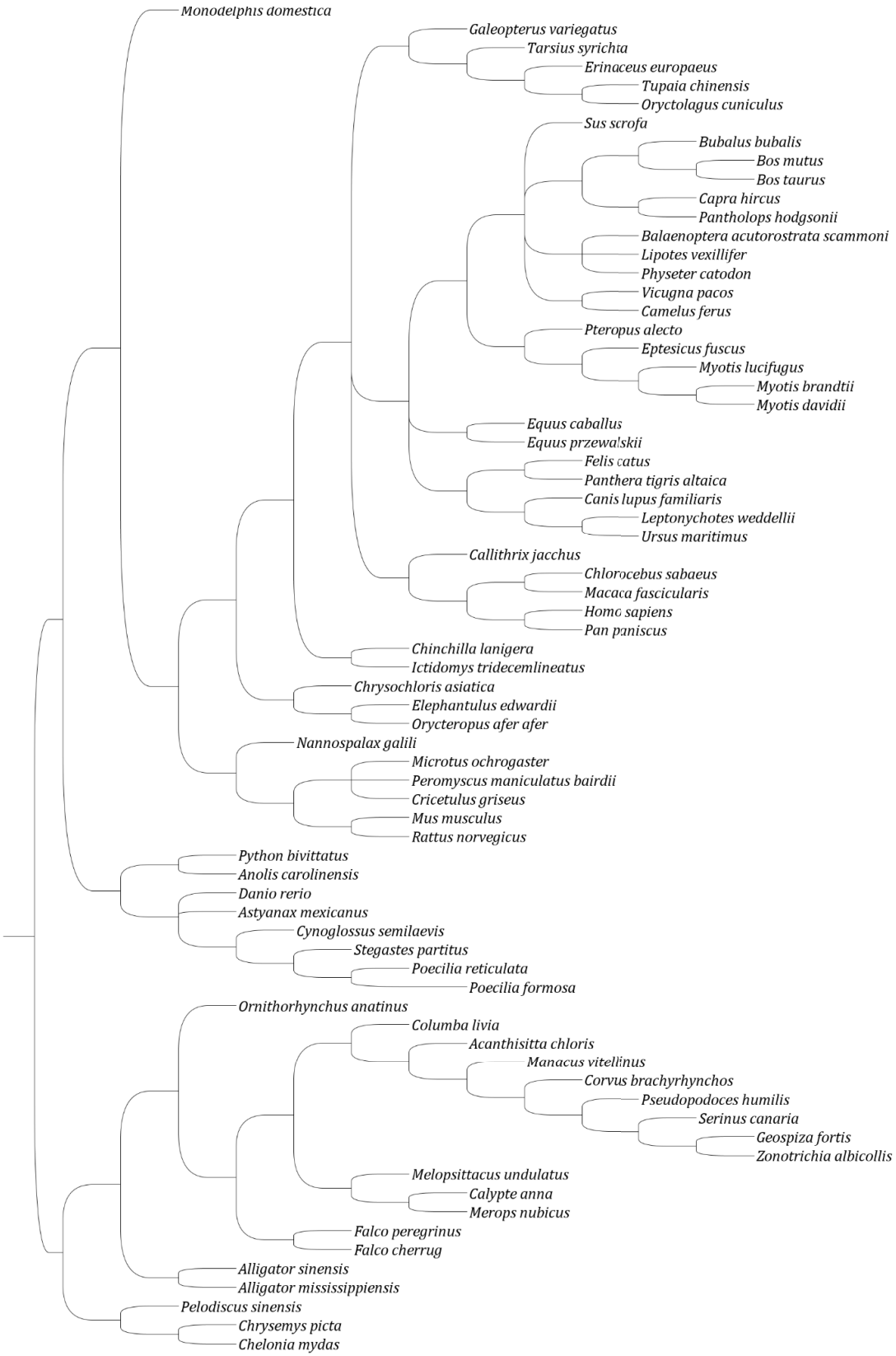

Monodelphis domestica
Chinchilla lanigera
Galeopterus variegatus
Tarsius syrichta
Ictidomys tridecemlineatus
Tupaia chinensis
Oryctolagus cuniculus
Pteropus alecto
Bos mutus
Bos taurus
Bubalus bubalis
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Erinaceus europaeus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Equus caballus
Equus przewalskii
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Ornithorhynchus anatinus
Merops nubicus
Calypte anna
Acanthisitta chloris
Manacus vitellinus
Pseudopodoces humilis
Corvus brachyrhynchos
Geospiza fortis
Serinus canaria
Zonotrichia albicollis
Melopsittacus undulatus
Columba livia
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Stegastes partitus
Cynoglossus semilaevis
Poecilia reticulata
Poecilia formosa

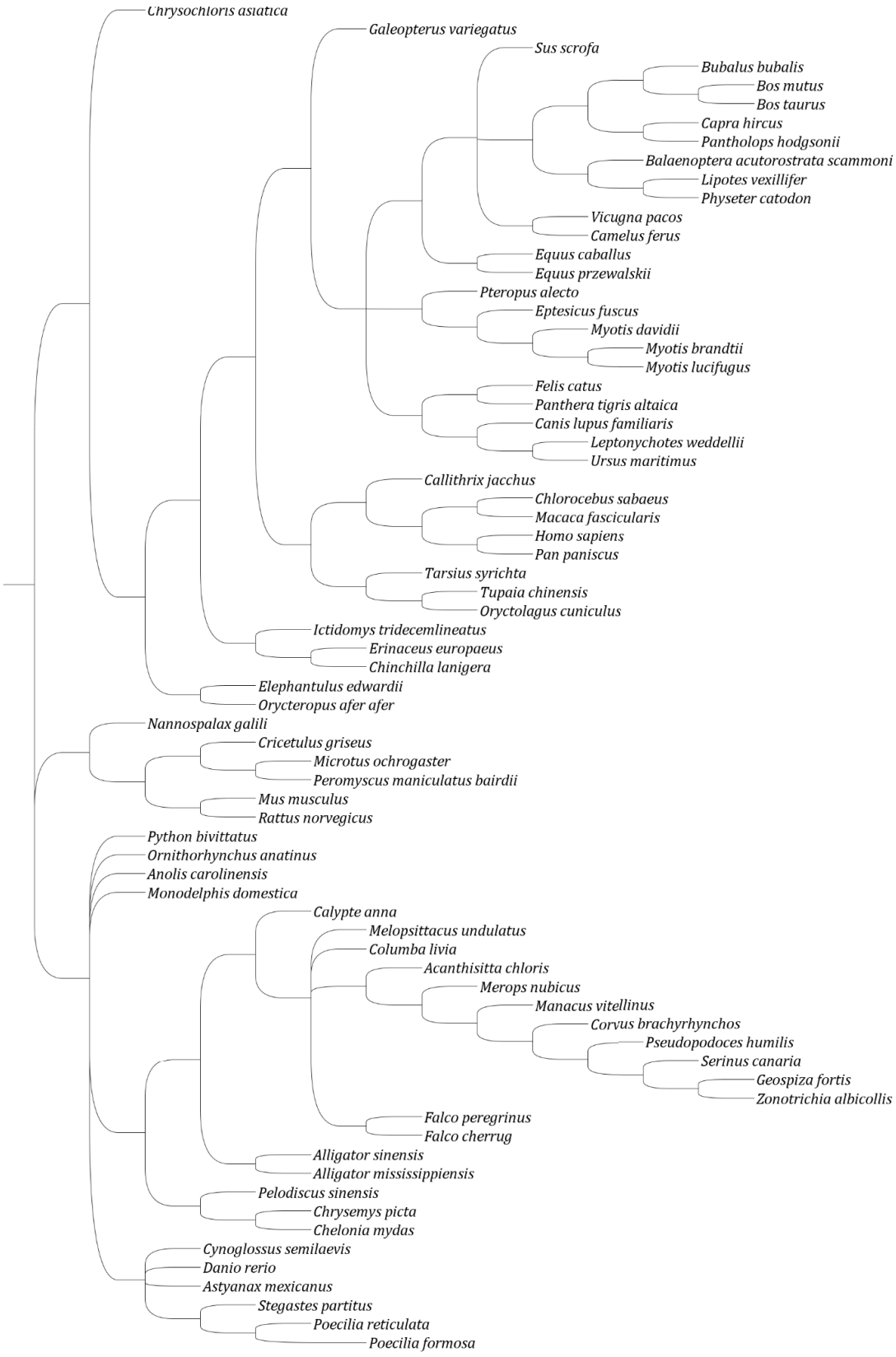

Calypte anna
Falco peregrinus
Falco cherrug
Erinaceus europaeus
Tarsius syrichta
Chrysochloris asiatica
Nannospalax galili
Ornithorhynchus anatinus
Tupaia chinensis
Oryctolagus cuniculus
Monodelphis domestica
Galeopterus variegatus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Equus caballus
Equus przewalskii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Vicugna pacos
Camelus ferus
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Elephantulus edwardii
Orycteropus afer afer
Python bivittatus
Anolis carolinensis
Chinchilla lanigera
Ictidomys tridecemlineatus
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Melopsittacus undulatus
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Columba livia
Merops nubicus

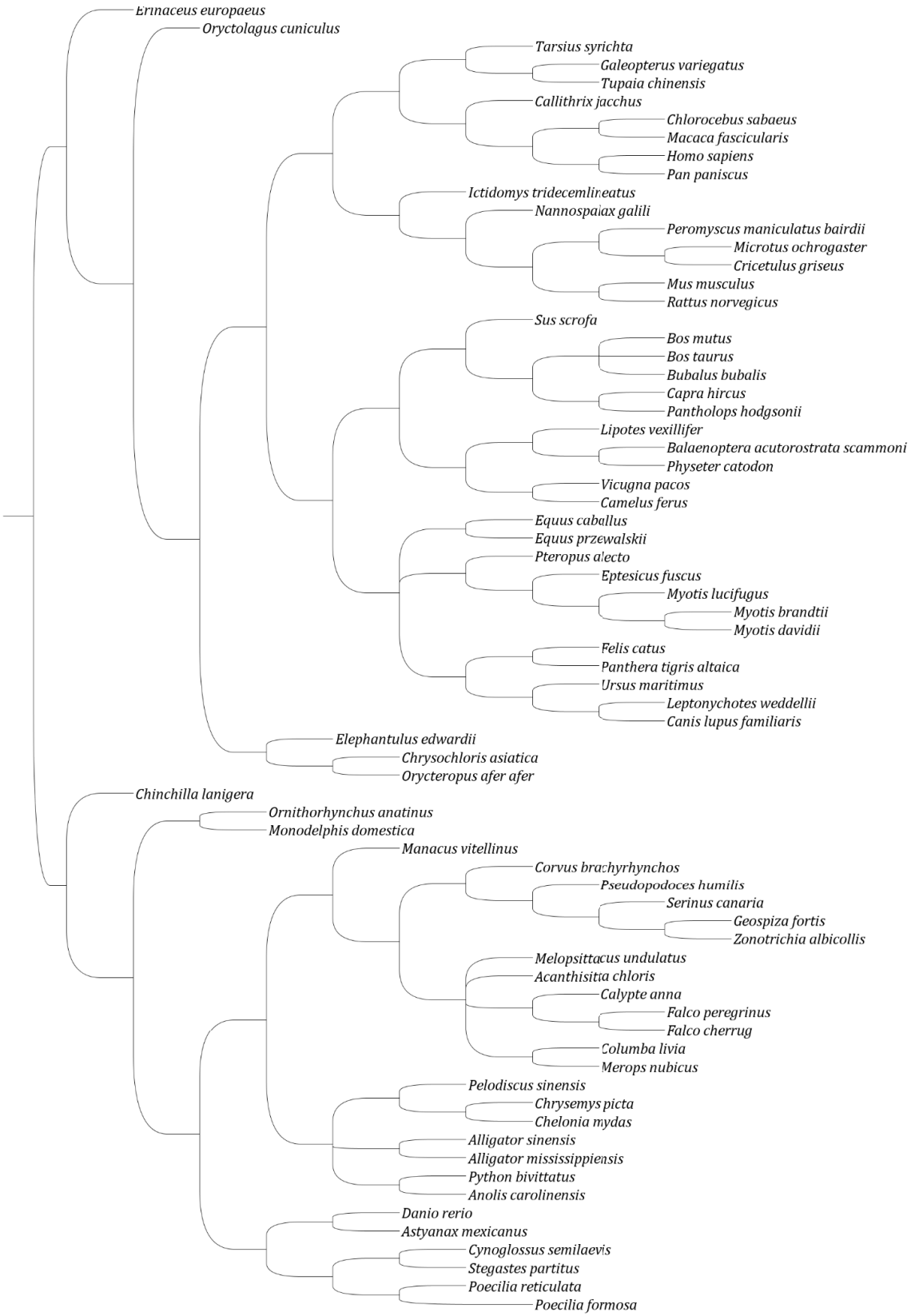

Ictidomys tridecemlineatus
Galeopterus variegatus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Equus caballus
Equus przewalskii
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Erinaceus europaeus
Chinchilla lanigera
Tupaia chinensis
Oryctolagus cuniculus
Orycteropus afer afer
Elephantulus edwardii
Chrysochloris asiatica
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Ornithorhynchus anatinus
Monodelphis domestica
Melopsittacus undulatus
Columba livia
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Calypte anna
Merops nubicus
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa


Elephantulus edwardii
Galeopterus variegatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Ictidomys tridecemlineatus
Tupaia chinensis
Chrysochloris asiatica
Orycteropus afer afer
Sus scrofa
Bos mutus
Bos taurus
Bubalus bubalis
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Erinaceus europaeus
Oryctolagus cuniculus
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Chinchilla lanigera
Ornithorhynchus anatinus
Monodelphis domestica
Melopsittacus undulatus
Calypte anna
Columba livia
Merops nubicus
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Chrysemys picta
Pelodiscus sinensis
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Stegastes partitus
Cynoglossus semilaevis
Poecilia reticulata
Poecilia formosa

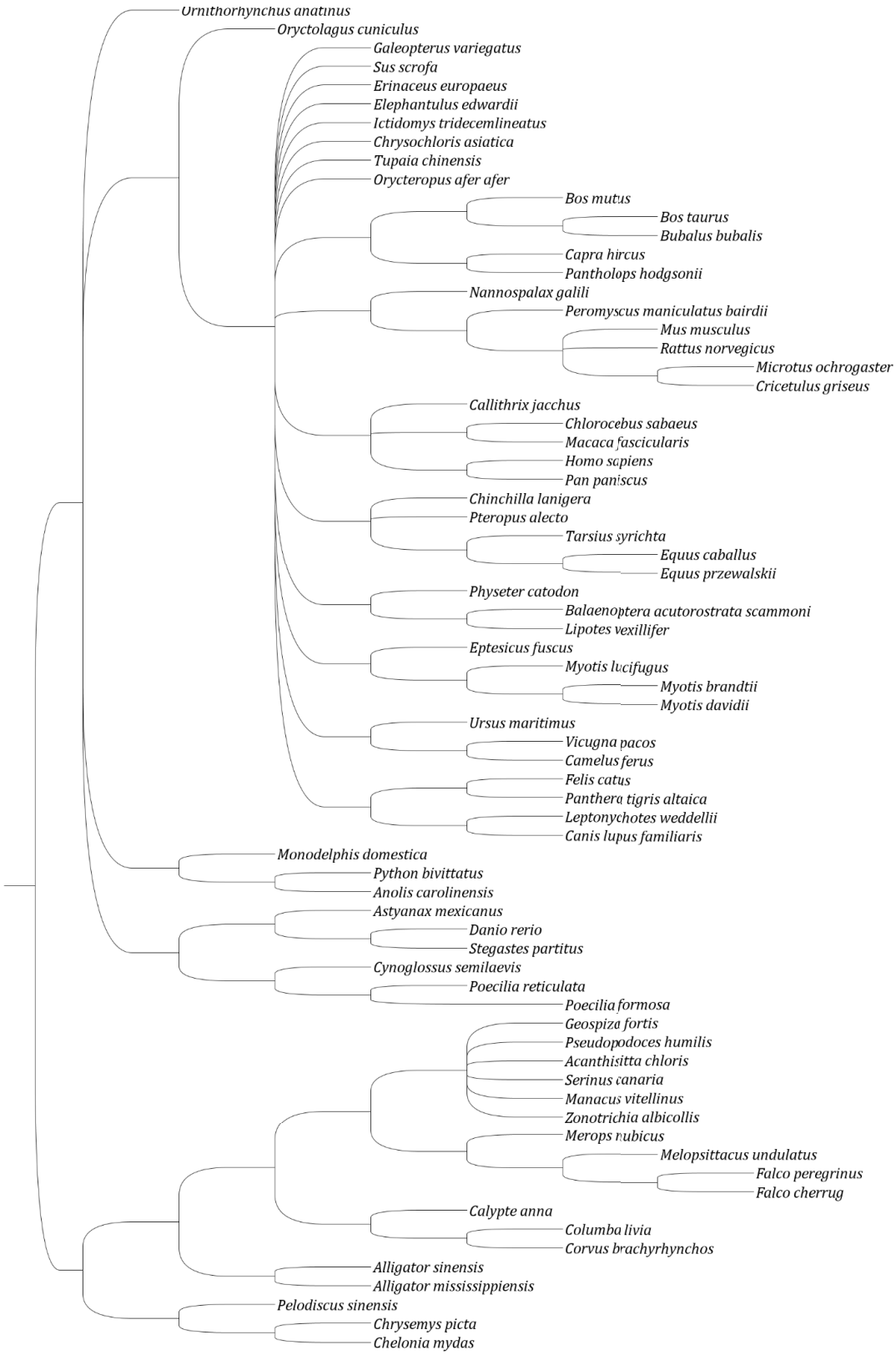

Galeopterus variegatus
Sus scrofa
Tarsius syrichta
Ictidomys tridecemlineatus
Chrysochloris asiatica
Leptonychotes weddellii
Tupaia chinensis
Oryctolagus cuniculus
Ursus maritimus
Canis lupus familiaris
Pteropus alecto
Bos mutus
Bos taurus
Bubalus bubalis
Capra hircus
Pantholops hodgsonii
Elephantulus edwardii
Orycteropus afer afer
Chlorocebus sabaeus
Callithrix jacchus
Macaca fascicularis
Homo sapiens
Pan paniscus
Equus caballus
Equus przewalskii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Vicugna pacos
Camelus ferus
Felis catus
Panthera tigris altaica
Erinaceus europaeus
Pelodiscus sinensis
Astyanax mexicanus
Python bivittatus
Chinchilla lanigera
Peromyscus maniculatus bairdii
Nannospalax galili
Anolis carolinensis
Merops nubicus
Monodelphis domestica
Pseudopodoces humilis
Corvus brachyrhynchos
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Melopsittacus undulatus
Calypte anna
Acanthisitta chloris
Columba livia
Manacus vitellinus
Falco peregrinus
Falco cherrug
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Alligator sinensis
Alligator mississippiensis
Chrysemys picta
Chelonia mydas
Cynoglossus semilaevis
Danio rerio
Stegastes partitus
Ornithorhynchus anatinus
Poecilia reticulata
Poecilia formosa

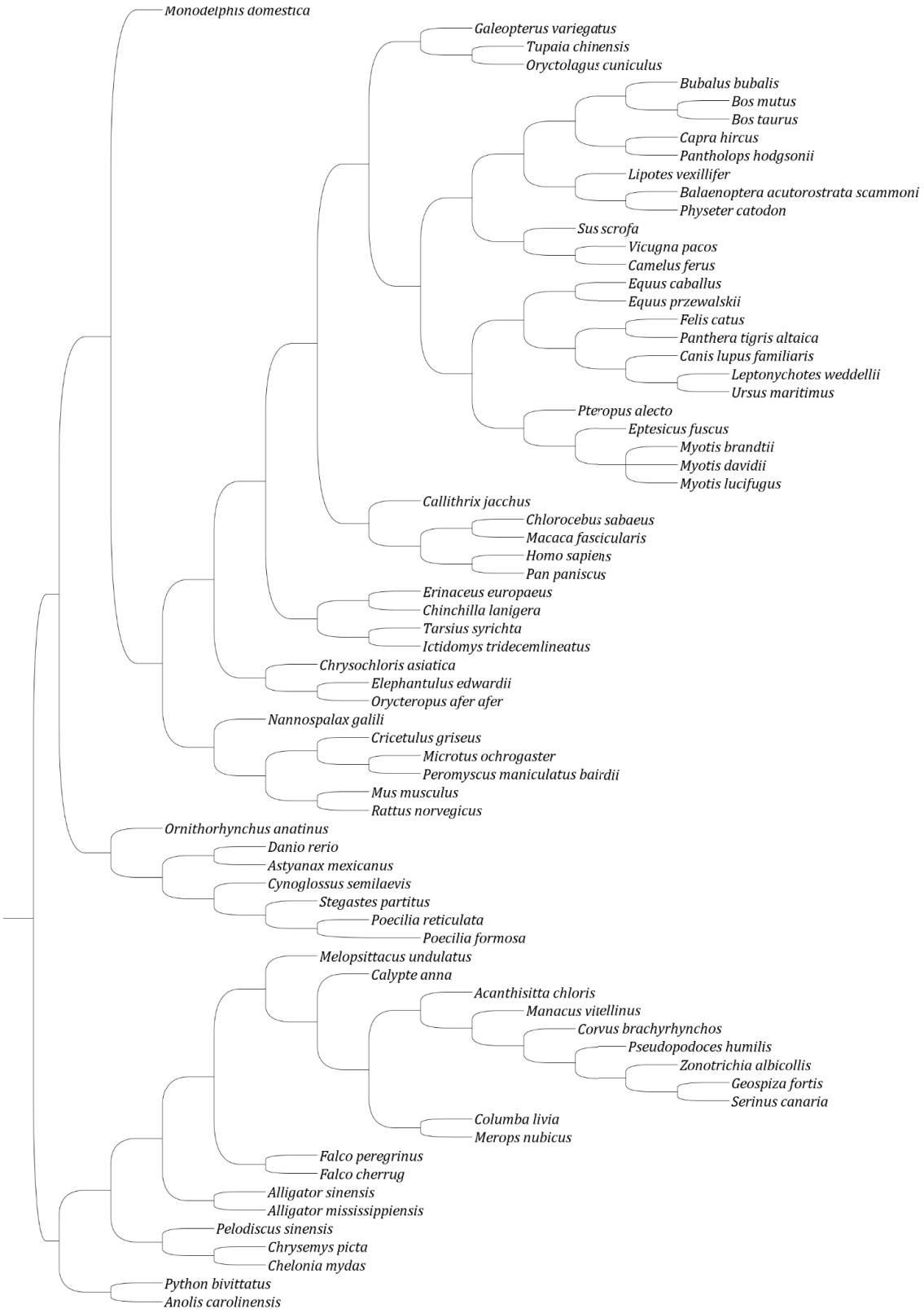

Monodelphis domestica
Galeopterus variegatus
Tarsius syrichta
Erinaceus europaeus
Tupaia chinensis
Oryctolagus cuniculus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Vicugna pacos
Camelus ferus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Chinchilla lanigera
Ictidomys tridecemlineatus
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Ornithorhynchus anatinus
Columba livia
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Melopsittacus undulatus
Calypte anna
Merops nubicus
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas

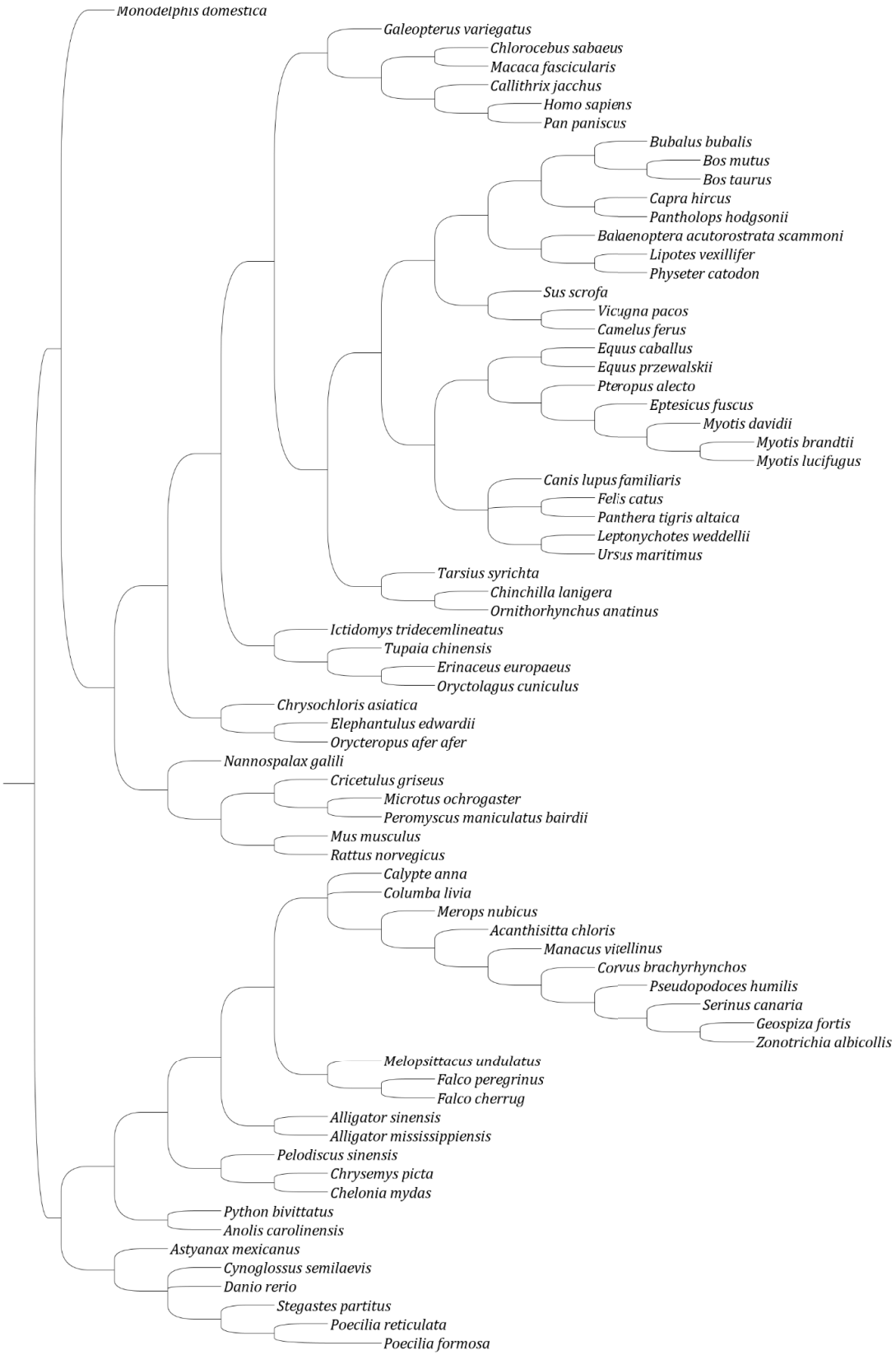

Chrysochloris asiatica
Galeopterus variegatus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Tarsius syrichta
Tupaia chinensis
Oryctolagus cuniculus
Ictidomys tridecemlineatus
Erinaceus europaeus
Chinchilla lanigera
Elephantulus edwardii
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Python bivittatus
Ornithorhynchus anatinus
Anolis carolinensis
Monodelphis domestica
Calypte anna
Melopsittacus undulatus
Columba livia
Acanthisitta chloris
Merops nubicus
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Cynoglossus semilaevis
Danio rerio
Astyanax mexicanus
Stegastes partitus
Poecilia reticulata
Poecilia formosa

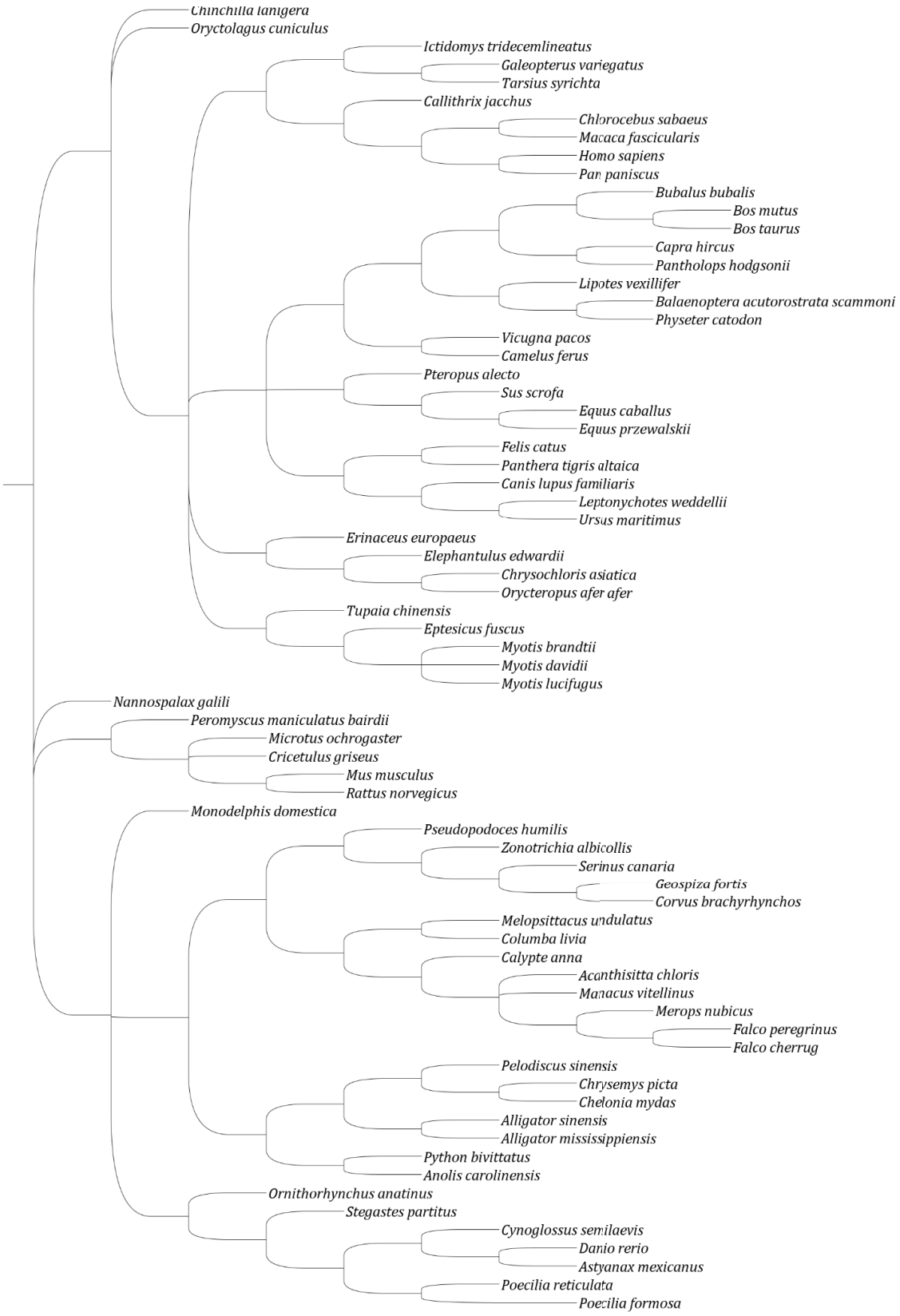

Erinaceus europaeus
Oryctolagus cuniculus
Tarsius syrichta
Galeopterus variegatus
Tupaia chinensis
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Ictidomys tridecemlineatus
Nannospalax galili
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Sus scrofa
Bos mutus
Bos taurus
Bubalus bubalis
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Felis catus
Panthera tigris altaica
Ursus maritimus
Leptonychotes weddellii
Canis lupus familiaris
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Chinchilla lanigera
Ornithorhynchus anatinus
Monodelphis domestica
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Melopsittacus undulatus
Acanthisitta chloris
Calypte anna
Falco peregrinus
Falco cherrug
Columba livia
Merops nubicus
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

CAG

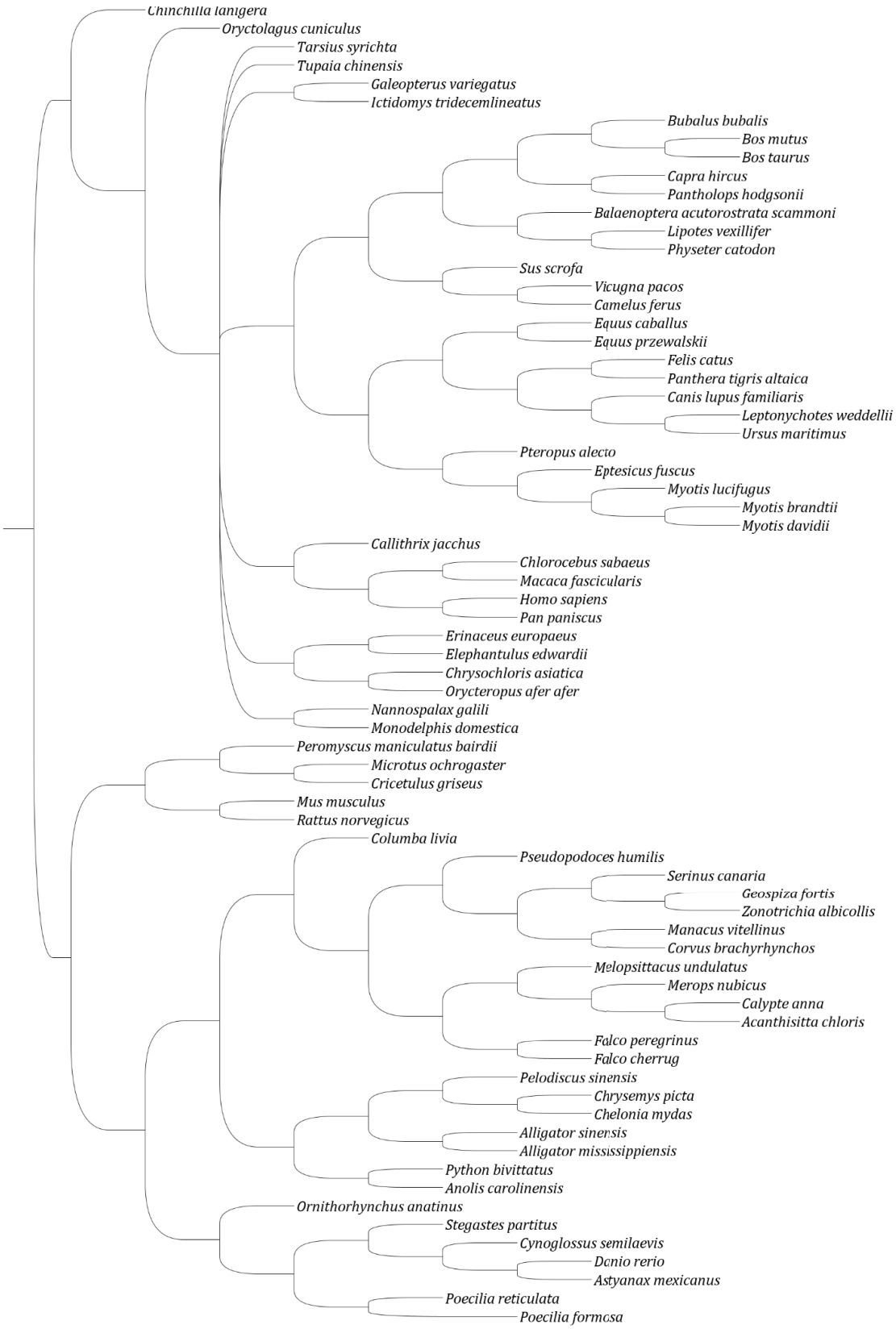

Ornithorhynchus anatinus
Oryctolagus cuniculus
Galeopterus variegatus
Sus scrofa
Erinaceus europaeus
Elephantulus edwardii
Ictidomys tridecemlineatus
Chrysochloris asiatica
Tupaia chinensis
Orycteropus afer afer
Bos mutus
Bos taurus
Bubalus bubalis
Capra hircus
Pantholops hodgsonii
Nannospalax galili
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Microtus ochrogaster
Cricetulus griseus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Chinchilla lanigera
Pteropus alecto
Tarsius syrichta
Equus caballus
Equus przewalskii
Physeter catodon
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Ursus maritimus
Vicugna pacos
Camelus ferus
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Canis lupus familiaris
Monodelphis domestica
Python bivittatus
Anolis carolinensis
Astyanax mexicanus
Danio rerio
Stegastes partitus
Cynoglossus semilaevis
Poecilia reticulata
Poecilia formosa
Geospiza fortis
Pseudopodoces humilis
Acanthisitta chloris
Serinus canaria
Manacus vitellinus
Zonotrichia albicollis
Merops nubicus
Melopsittacus undulatus
Falco peregrinus
Falco cherrug
Calypte anna
Columba livia
Corvus brachyrhynchos
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas

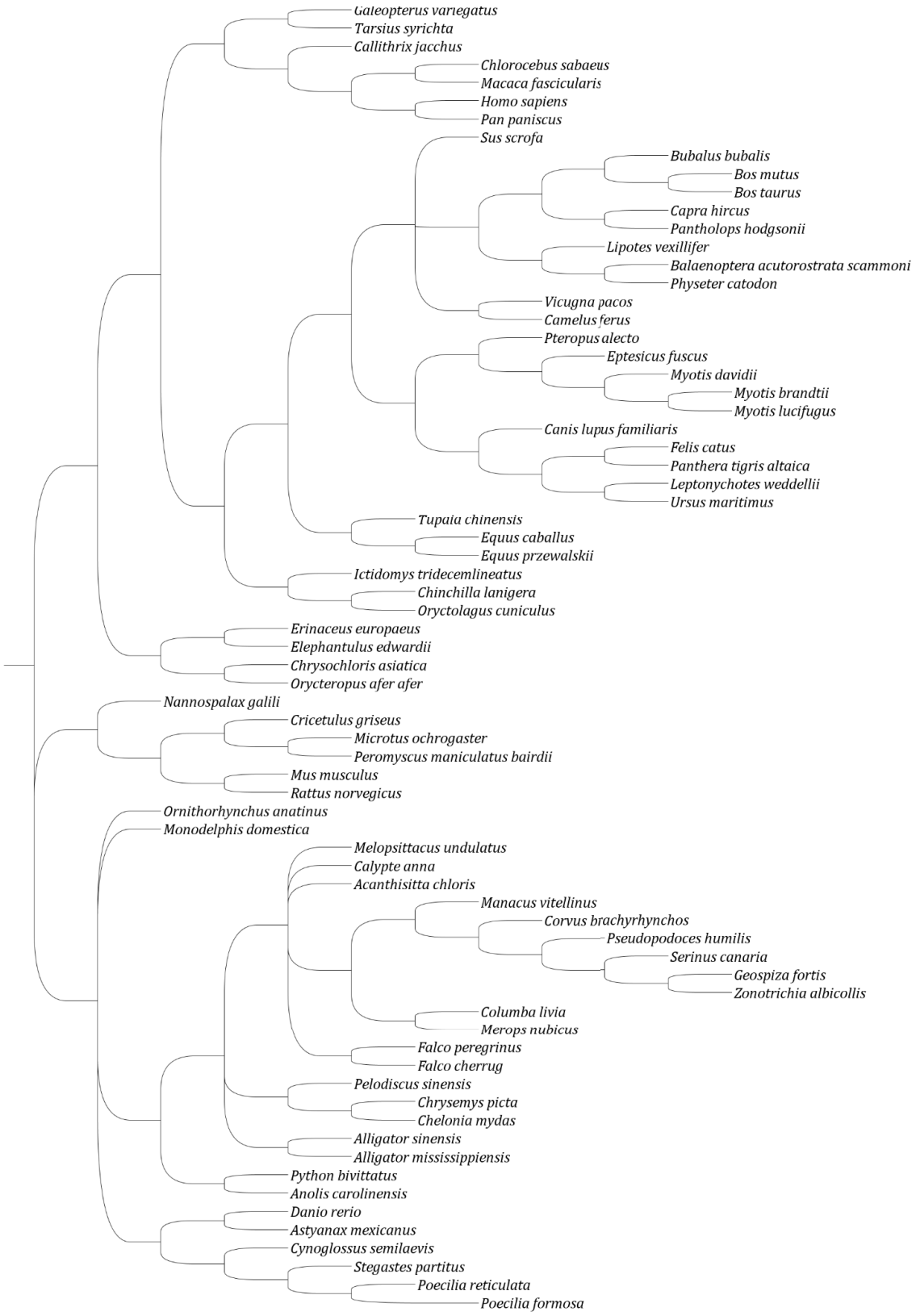

Monodelphis domestica
Galeopterus variegatus
Tupaia chinensis
Oryctolagus cuniculus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis brandtii
Myotis davidii
Myotis lucifugus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Erinaceus europaeus
Chinchilla lanigera
Tarsius syrichta
Ictidomys tridecemlineatus
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Ornithorhynchus anatinus
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Melopsittacus undulatus
Calypte anna
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Zonotrichia albicollis
Geospiza fortis
Serinus canaria
Columba livia
Merops nubicus
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Python bivittatus
Anolis carolinensis

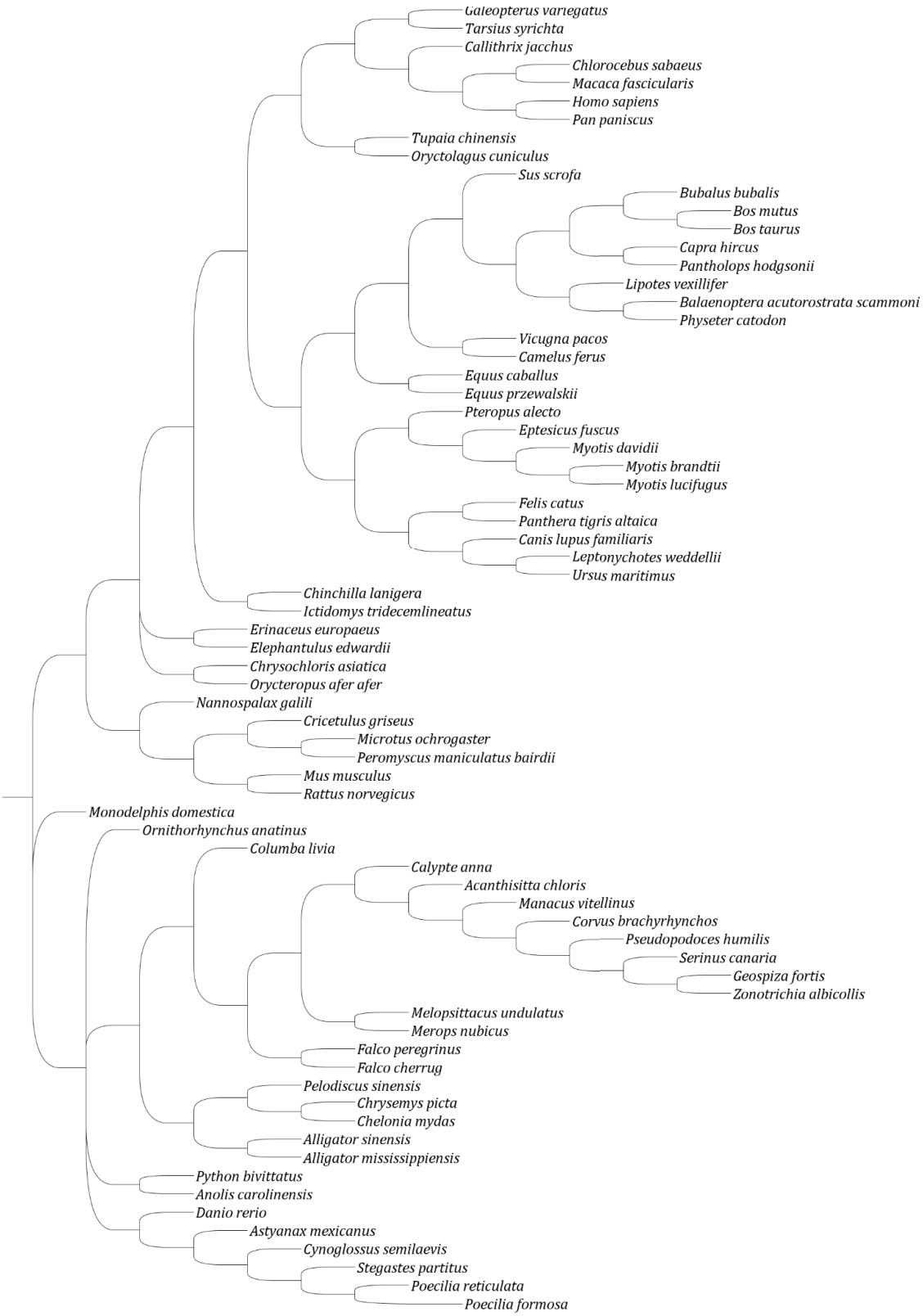

Monodelphis domestica
Galeopterus variegatus
Chlorocebus sabaeus
Macaca fascicularis
Callithrix jacchus
Homo sapiens
Pan paniscus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Tarsius syrichta
Chinchilla lanigera
Ornithorhynchus anatinus
Ictidomys tridecemlineatus
Tupaia chinensis
Erinaceus europaeus
Oryctolagus cuniculus
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Calypte anna
Columba livia
Merops nubicus
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Melopsittacus undulatus
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Python bivittatus
Anolis carolinensis
Astyanax mexicanus
Cynoglossus semilaevis
Danio rerio
Stegastes partitus
Poecilia reticulata
Poecilia formosa

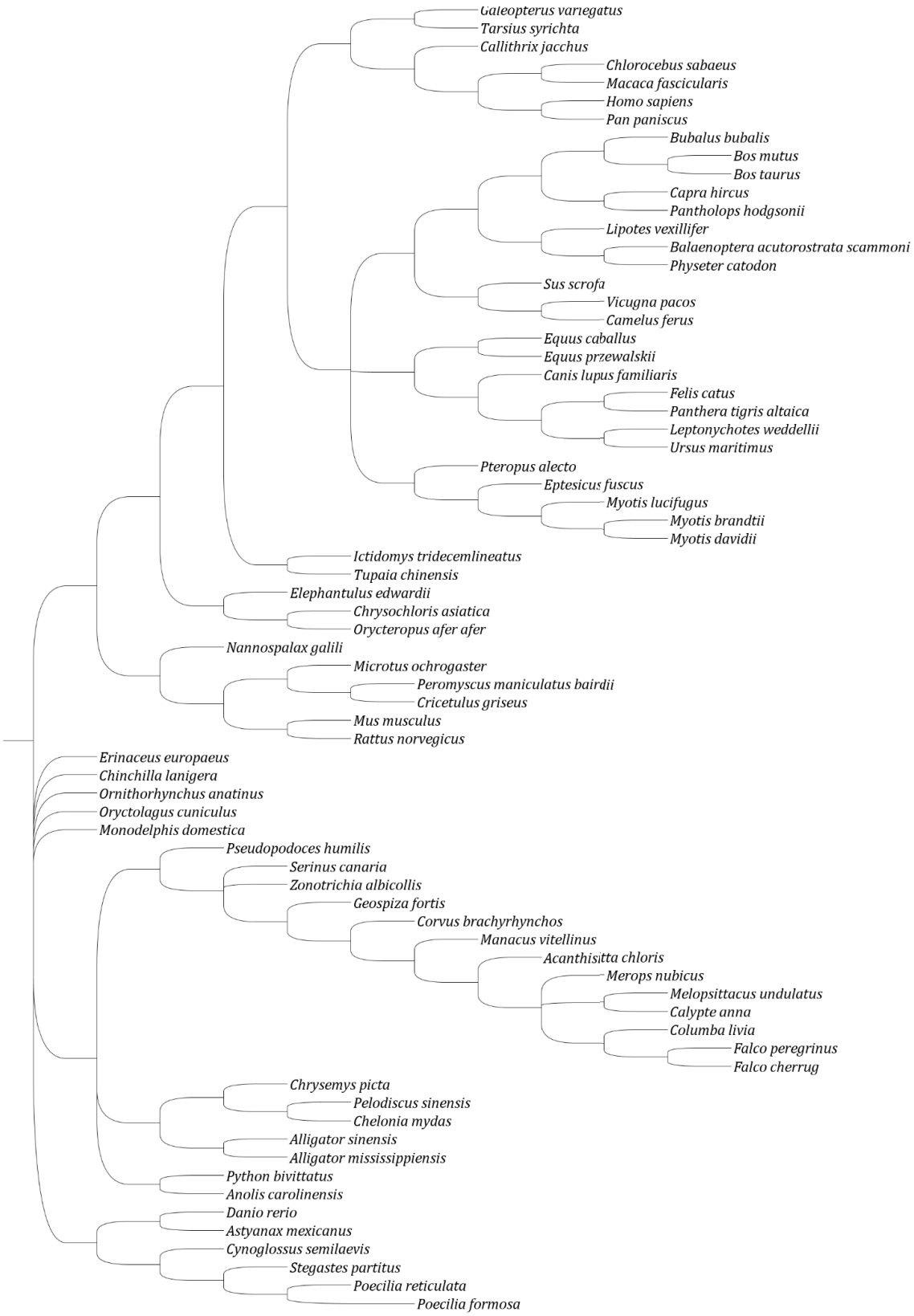

Chinchilla lanigera
Oryctolagus cuniculus
Ictidomys tridecemlineatus
Galeopterus variegatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Pteropus alecto
Sus scrofa
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Erinaceus europaeus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Tupaia chinensis
Eptesicus fuscus
Myotis brandtii
Myotis davidii
Myotis lucifugus
Nannospalax galili
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Monodelphis domestica
Pseudopodoces humilis
Zonotrichia albicollis
Serinus canaria
Geospiza fortis
Corvus brachyrhynchos
Melopsittacus undulatus
Columba livia
Calypte anna
Acanthisitta chloris
Manacus vitellinus
Merops nubicus
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Ornithorhynchus anatinus
Stegastes partitus
Cynoglossus semilaevis
Danio rerio
Astyanax mexicanus
Poecilia reticulata
Poecilia formosa

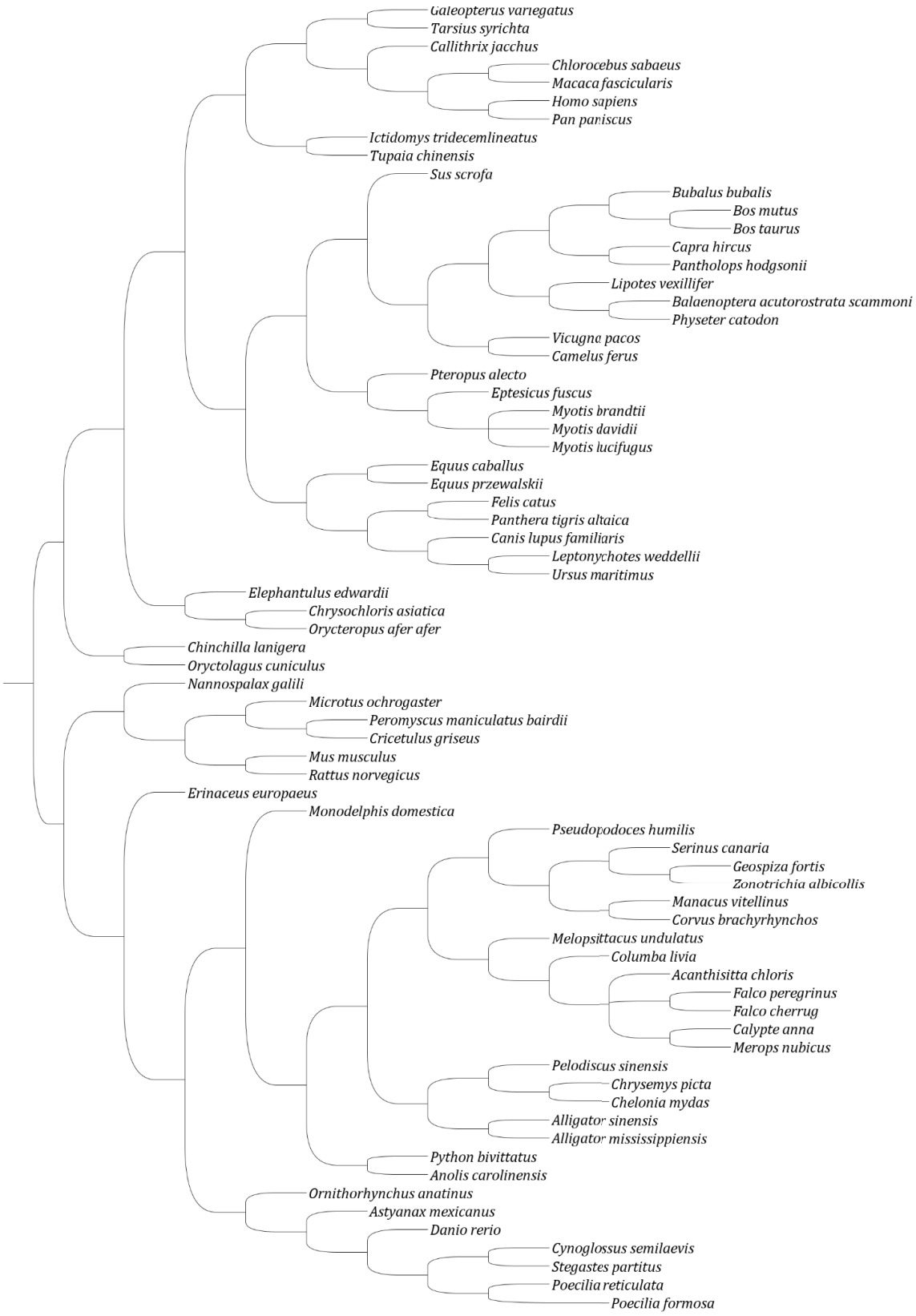
CCG

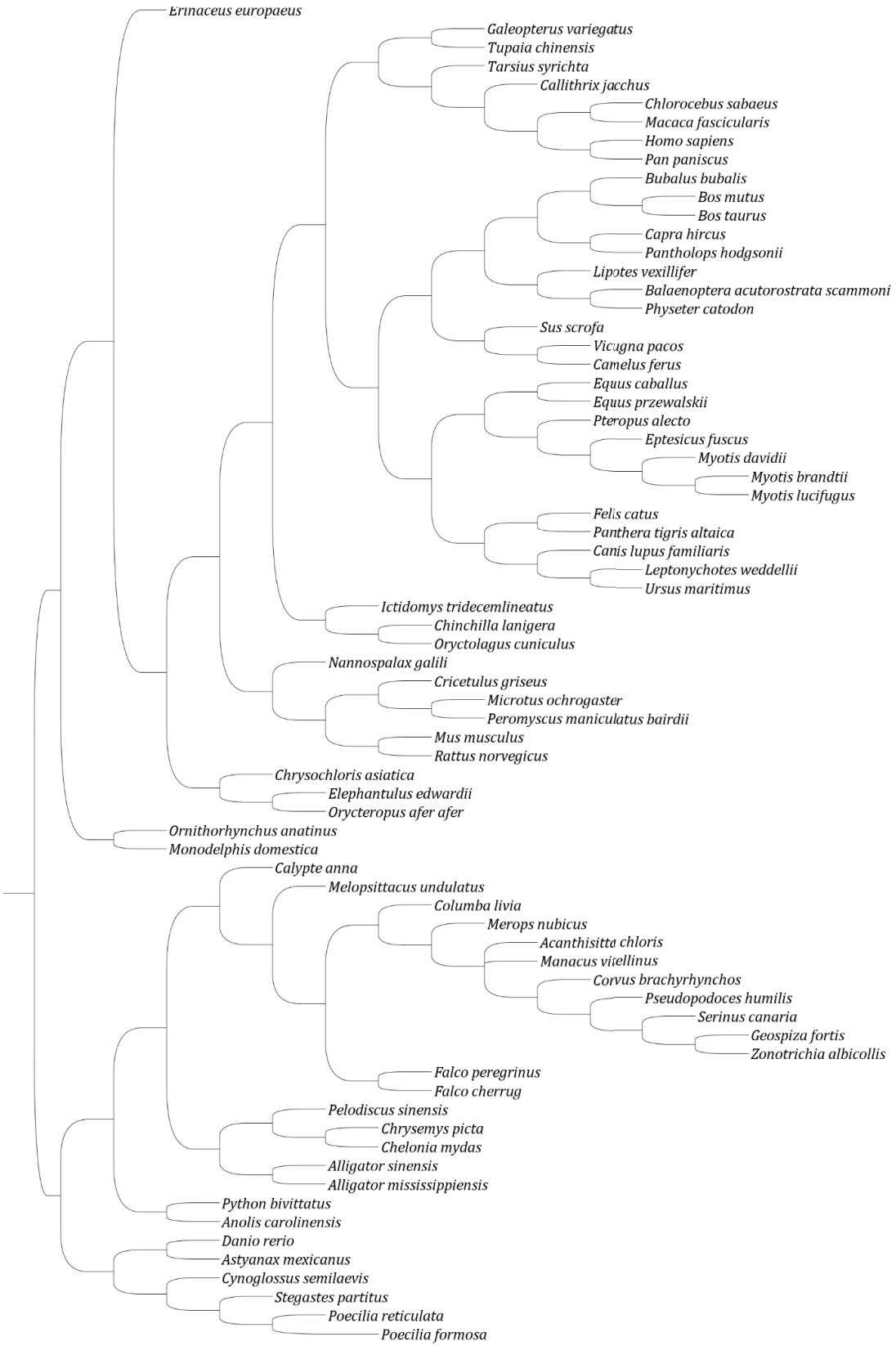

Chinchilla lanigera
Oryctolagus cuniculus
Tarsius syrichta
Tupaia chinensis
Galeopterus variegatus
Ictidomys tridecemlineatus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Erinaceus europaeus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Nannospalax galili
Monodelphis domestica
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Columba livia
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Manacus vitellinus
Corvus brachyrhynchos
Melopsittacus undulatus
Merops nubicus
Calypte anna
Acanthisitta chloris
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Ornithorhynchus anatinus
Stegastes partitus
Cynoglossus semilaevis
Danio rerio
Astyanax mexicanus
Poecilia reticulata
Poecilia formosa

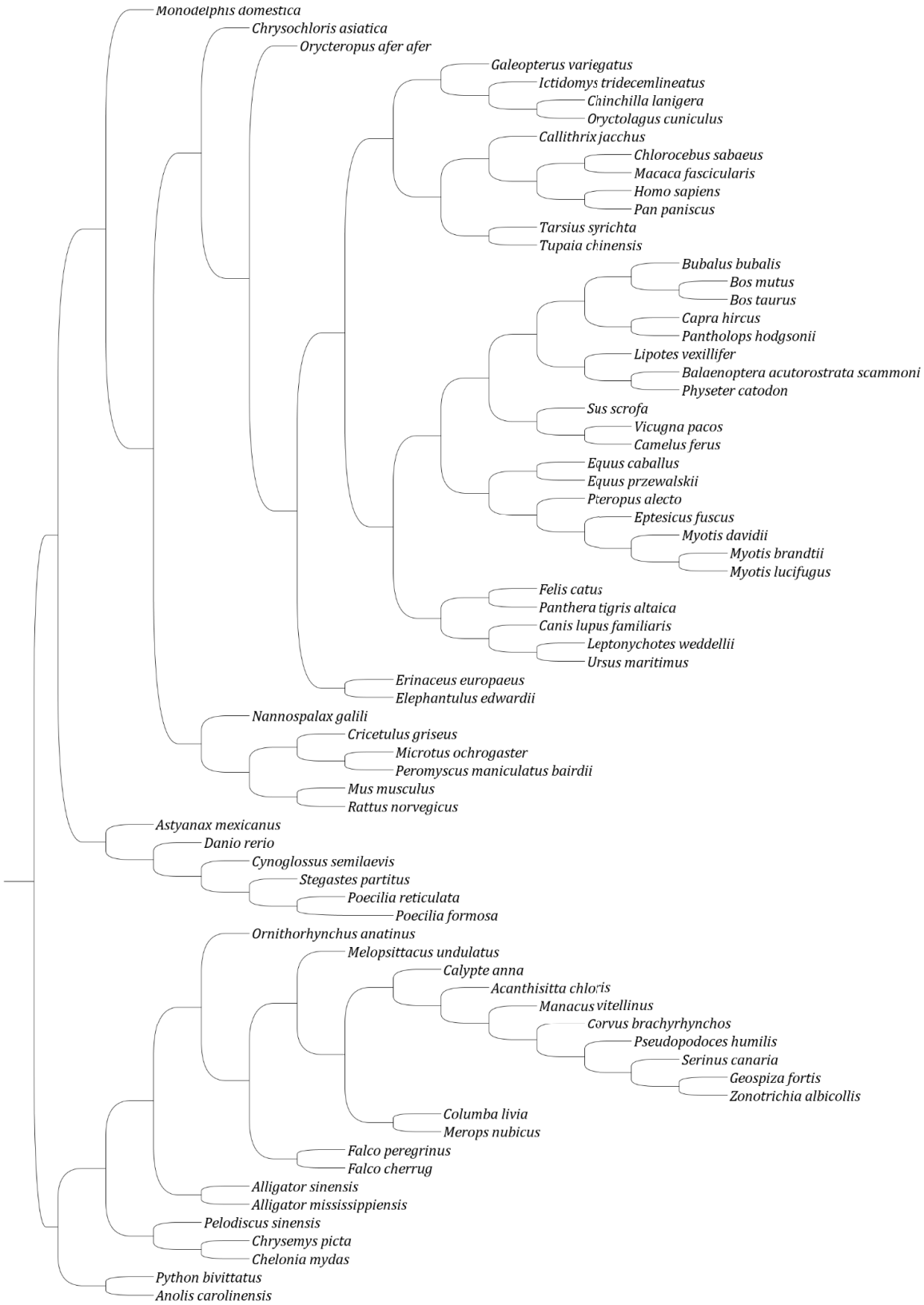

Galeopterus variegatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Tupaia chinensis
Equus caballus
Equus przewalskii
Ictidomys tridecemlineatus
Chinchilla lanigera
Oryctolagus cuniculus
Erinaceus europaeus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Ornithorhynchus anatinus
Monodelphis domestica
Melopsittacus undulatus
Calypte anna
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Columba livia
Merops nubicus
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

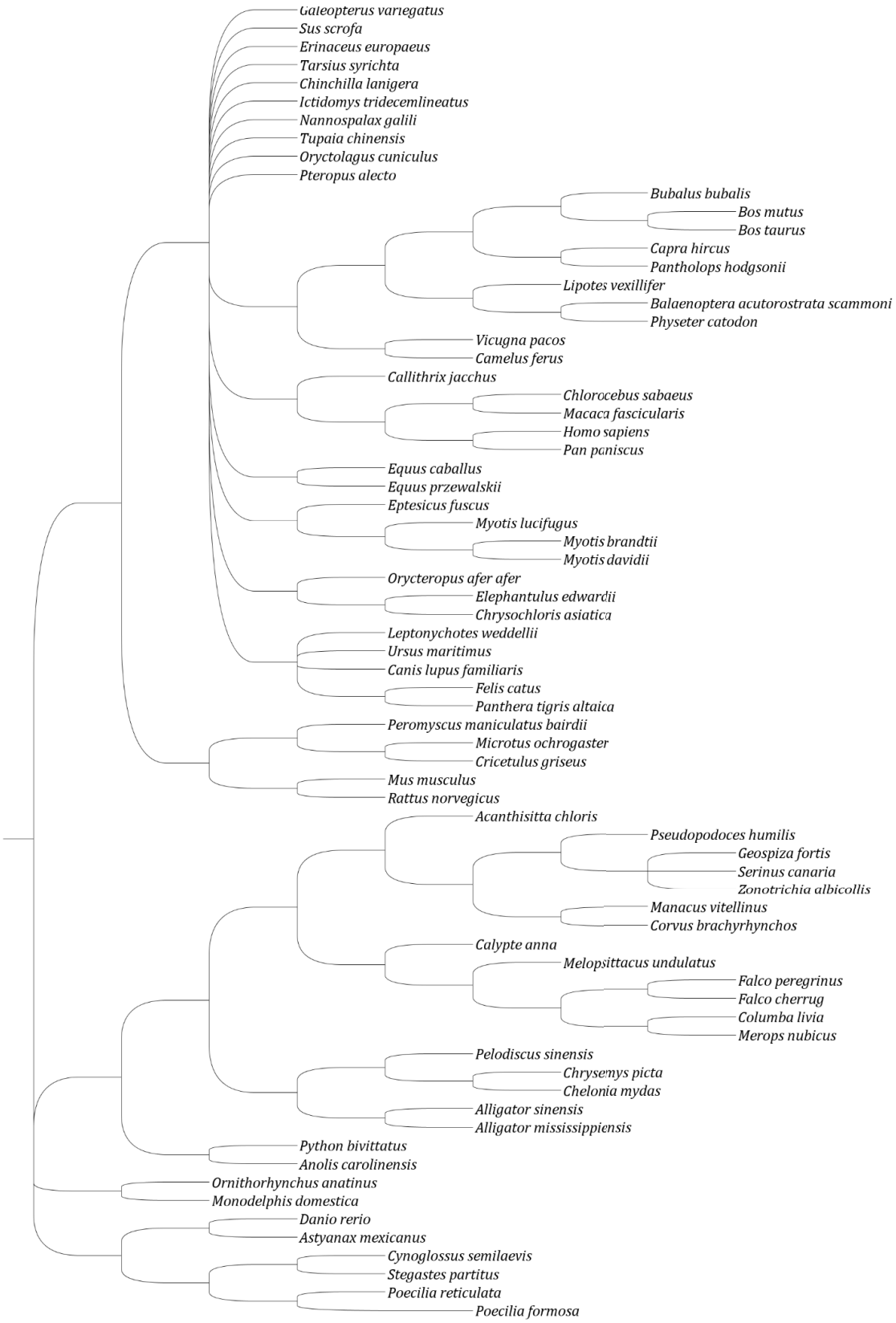

Galeopterus variegatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Tupaia chinensis
Oryctolagus cuniculus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Chinchilla lanigera
Ictidomys tridecemlineatus
Erinaceus europaeus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Monodelphis domestica
Ornithorhynchus anatinus
Columba livia
Calypte anna
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Melopsittacus undulatus
Merops nubicus
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa


Galeopterus variegatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Ictidomys tridecemlineatus
Tupaia chinensis
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Erinaceus europaeus
Chinchilla lanigera
Ornithorhynchus anatinus
Oryctolagus cuniculus
Monodelphis domestica
Pseudopodoces humilis
Serinus canaria
Zonotrichia albicollis
Geospiza fortis
Corvus brachyrhynchos
Manacus vitellinus
Acanthisitta chloris
Merops nubicus
Melopsittacus undulatus
Calypte anna
Columba livia
Falco peregrinus
Falco cherrug
Chrysemys picta
Pelodiscus sinensis
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

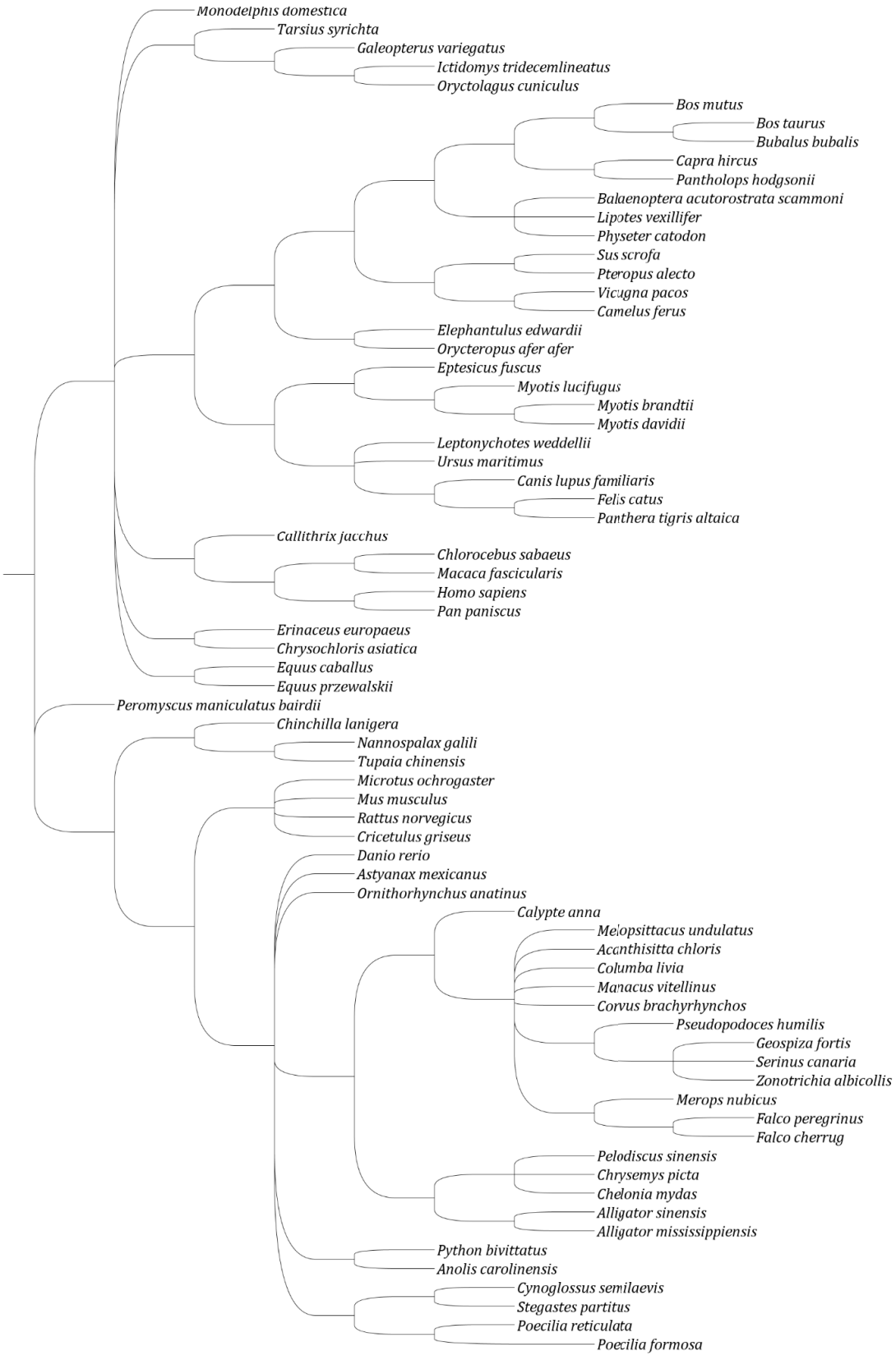

Galeopterus variegatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Ictidomys tridecemlineatus
Tupaia chinensis
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Pteropus alecto
Eptesicus fuscus
Myotis brandtii
Myotis davidii
Myotis lucifugus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Chinchilla lanigera
Oryctolagus cuniculus
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Erinaceus europaeus
Monodelphis domestica
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Manacus vitellinus
Corvus brachyrhynchos
Melopsittacus undulatus
Columba livia
Acanthisitta chloris
Falco peregrinus
Falco cherrug
Calypte anna
Merops nubicus
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Ornithorhynchus anatinus
Astyanax mexicanus
Danio rerio
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

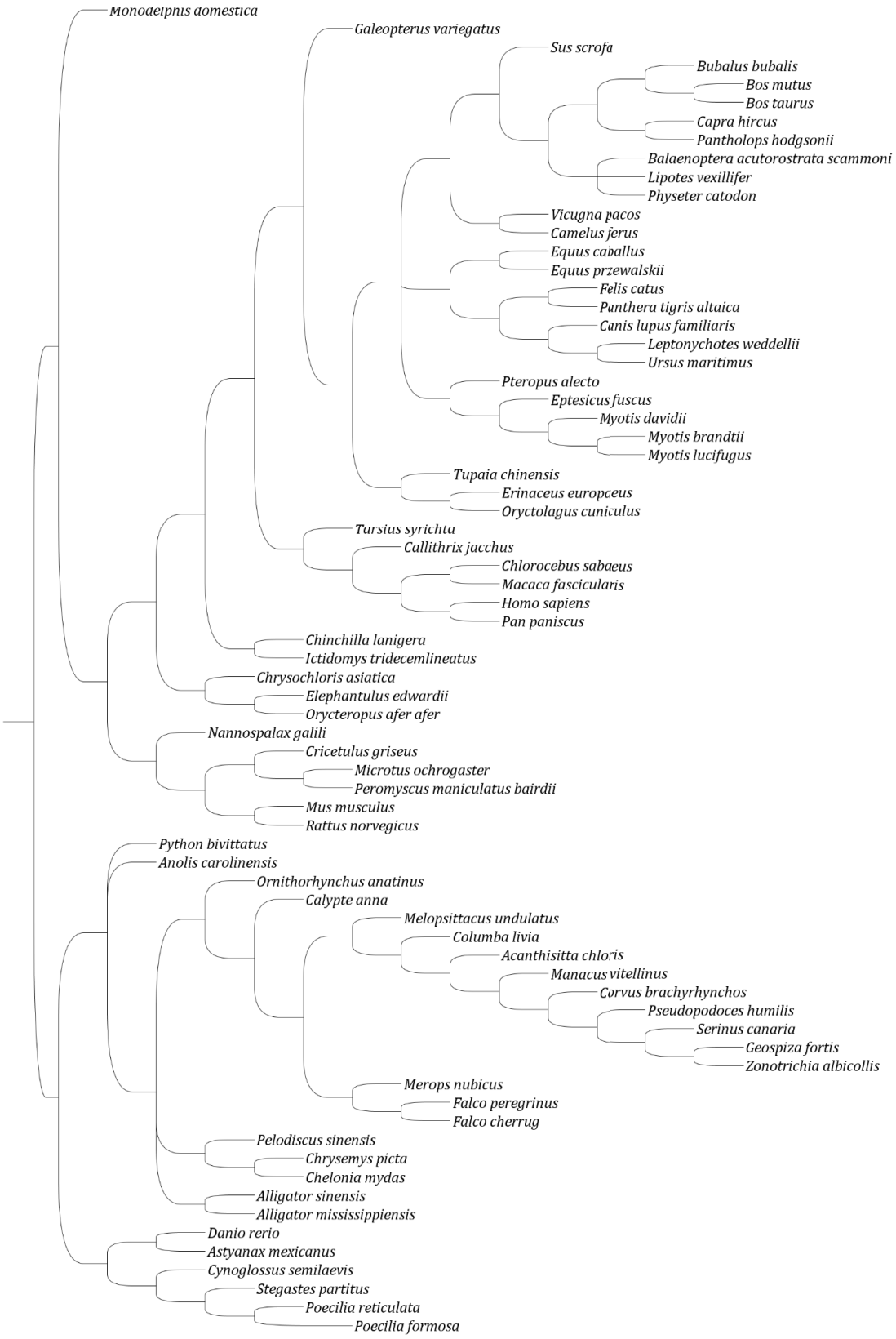

Erinaceus europaeus
Galeopterus variegatus
Tupaia chinensis
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Ictidomys tridecemlineatus
Chinchilla lanigera
Oryctolagus cuniculus
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Ornithorhynchus anatinus
Monodelphis domestica
Calypte anna
Melopsittacus undulatus
Columba livia
Merops nubicus
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

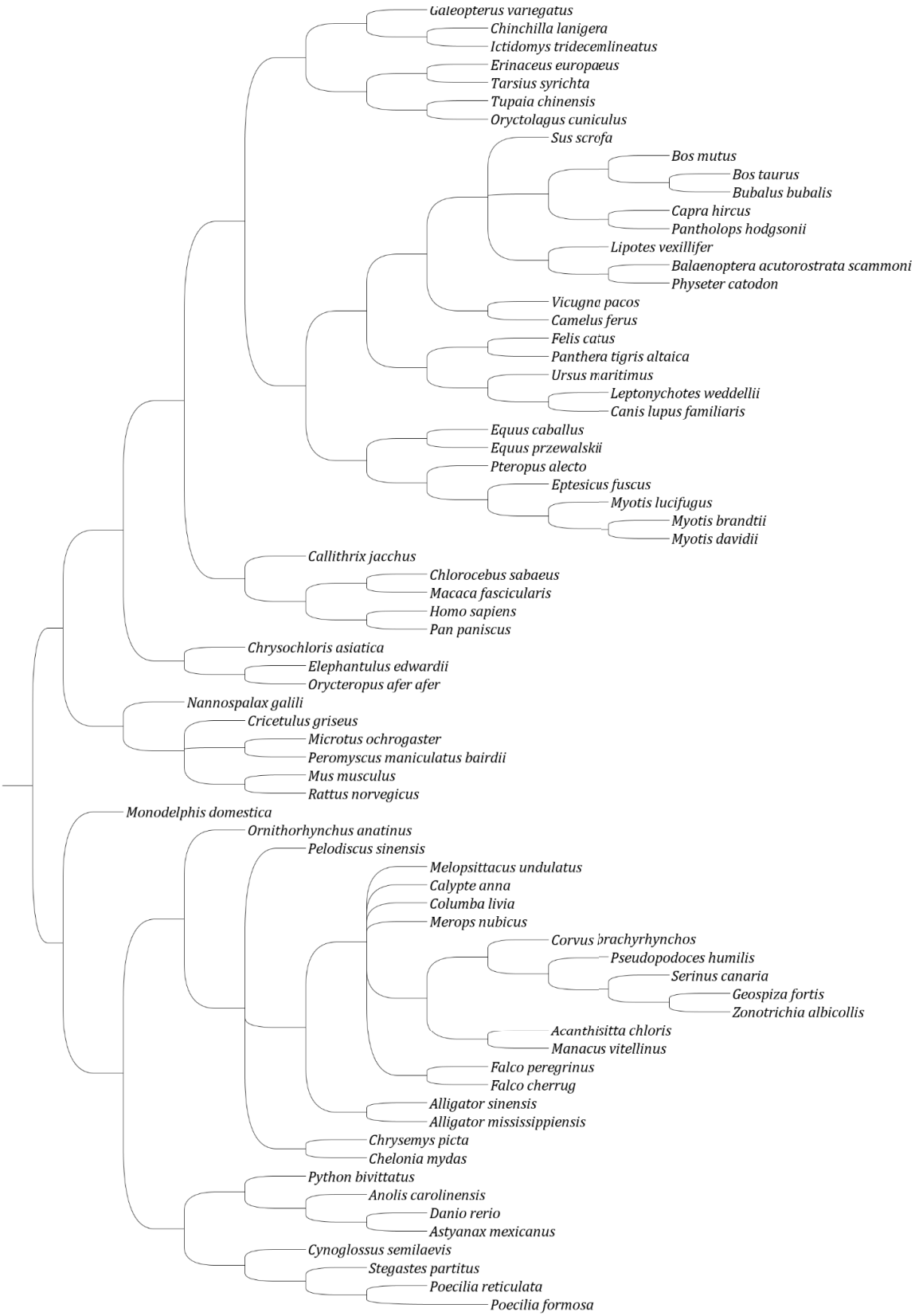

Monodelphis domestica
Chrysochloris asiatica
Orycteropus afer afer
Galeopterus variegatus
Ictidomys tridecemlineatus
Chinchilla lanigera
Oryctolagus cuniculus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Tarsius syrichta
Tupaia chinensis
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Erinaceus europaeus
Elephantulus edwardii
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Astyanax mexicanus
Danio rerio
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Ornithorhynchus anatinus
Melopsittacus undulatus
Calypte anna
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Columba livia
Merops nubicus
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Python bivittatus
Anolis carolinensis

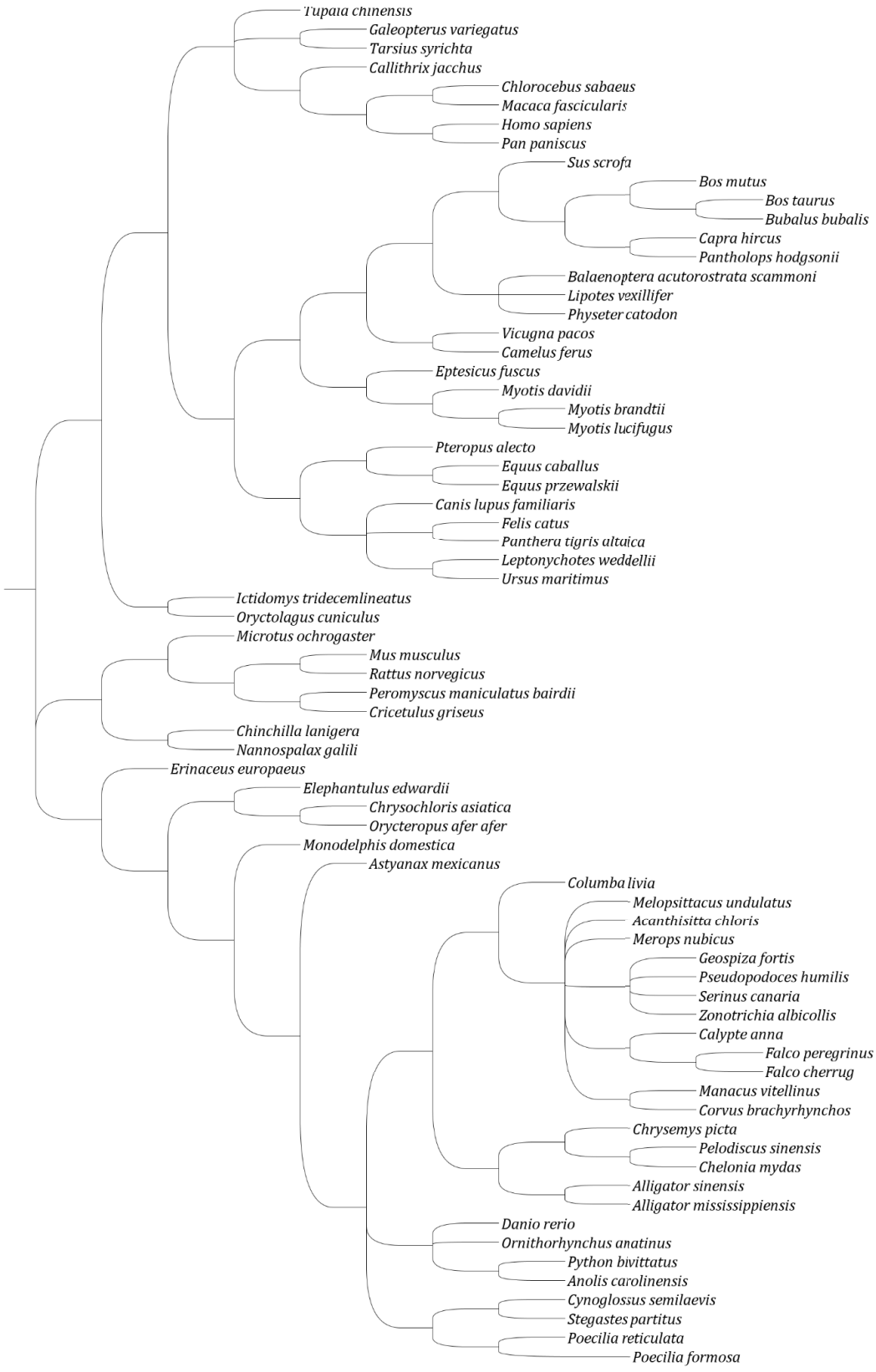

Galeopterus variegatus
Sus scrofa
Erinaceus europaeus
Tarsius syrichta
Chinchilla lanigera
Ictidomys tridecemlineatus
Nannospalax galili
Tupaia chinensis
Oryctolagus cuniculus
Pteropus alecto
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Equus caballus
Equus przewalskii
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Orycteropus afer afer
Elephantulus edwardii
Chrysochloris asiatica
Leptonychotes weddellii
Ursus maritimus
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Acanthisitta chloris
Pseudopodoces humilis
Geospiza fortis
Serinus canaria
Zonotrichia albicollis
Manacus vitellinus
Corvus brachyrhynchos
Calypte anna
Melopsittacus undulatus
Falco peregrinus
Falco cherrug
Columba livia
Merops nubicus
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Ornithorhynchus anatinus
Monodelphis domestica
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

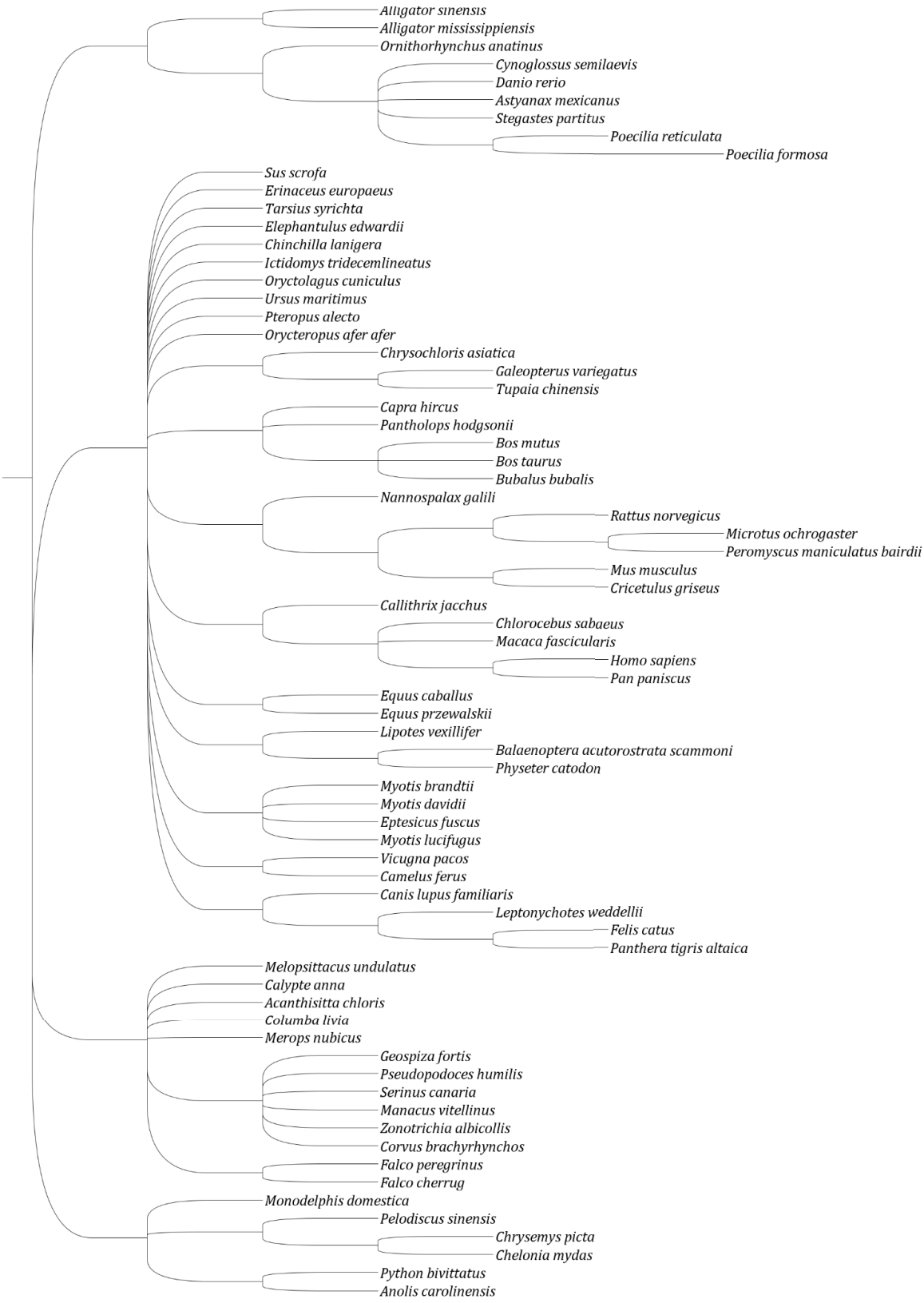
CTG


Monodelphis domestica
Tarsius syrichta
Galeopterus variegatus
Ictidomys tridecemlineatus
Oryctolagus cuniculus
Bos mutus
Bos taurus
Bubalus bubalis
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Sus scrofa
Pteropus alecto
Vicugna pacos
Camelus ferus
Elephantulus edwardii
Orycteropus afer afer
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Leptonychotes weddellii
Ursus maritimus
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Erinaceus europaeus
Chrysochloris asiatica
Equus caballus
Equus przewalskii
Peromyscus maniculatus bairdii
Chinchilla lanigera
Nannospalax galili
Tupaia chinensis
Microtus ochrogaster
Mus musculus
Rattus norvegicus
Cricetulus griseus
Danio rerio
Astyanax mexicanus
Ornithorhynchus anatinus
Calypte anna
Melopsittacus undulatus
Acanthisitta chloris
Columba livia
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Geospiza fortis
Serinus canaria
Zonotrichia albicollis
Merops nubicus
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

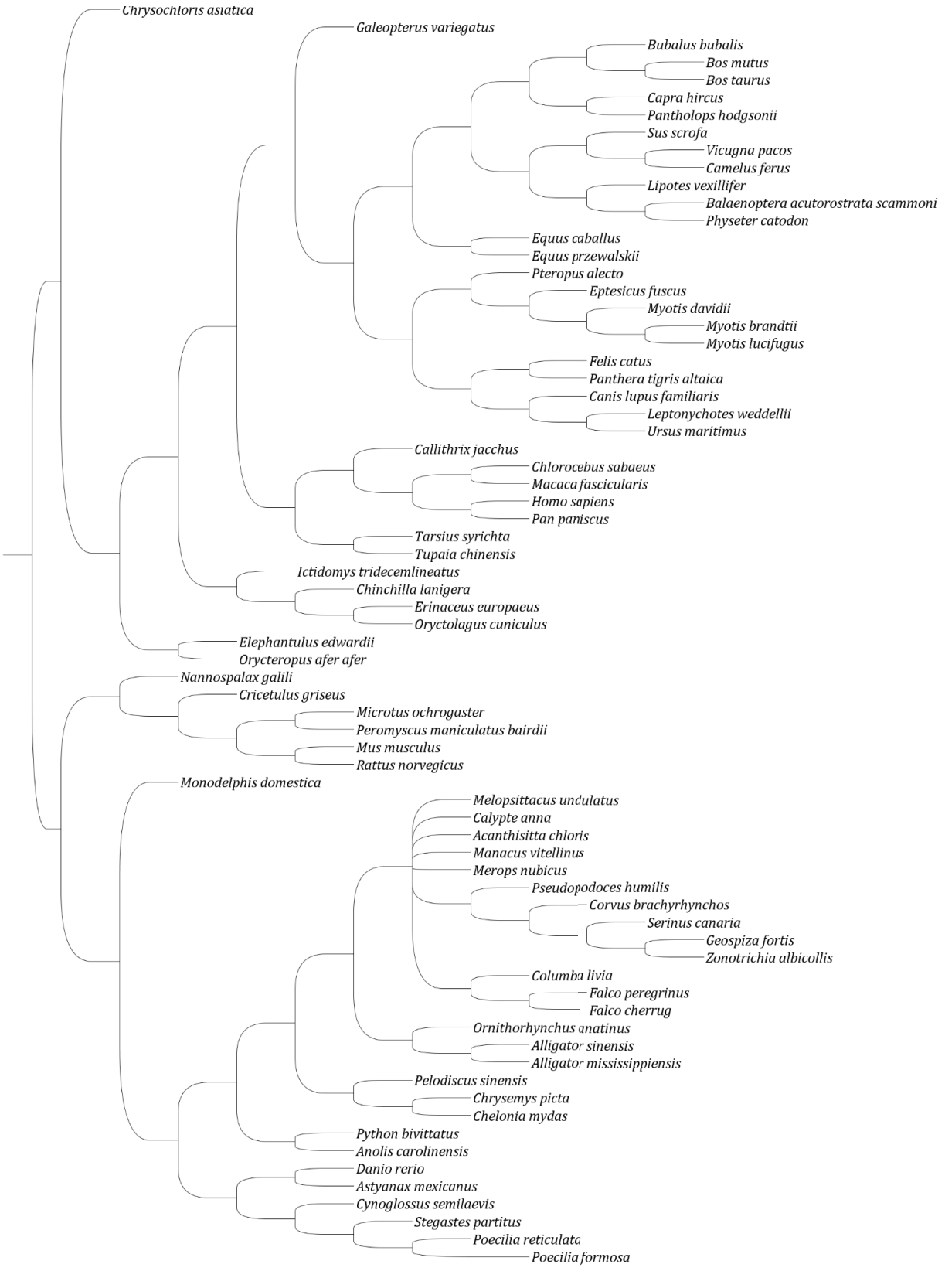

Monodelphis domestica
Galeopterus variegatus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Tupaia chinensis
Erinaceus europaeus
Oryctolagus cuniculus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Chinchilla lanigera
Ictidomys tridecemlineatus
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Python bivittatus
Anolis carolinensis
Ornithorhynchus anatinus
Calypte anna
Melopsittacus undulatus
Columba livia
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Merops nubicus
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

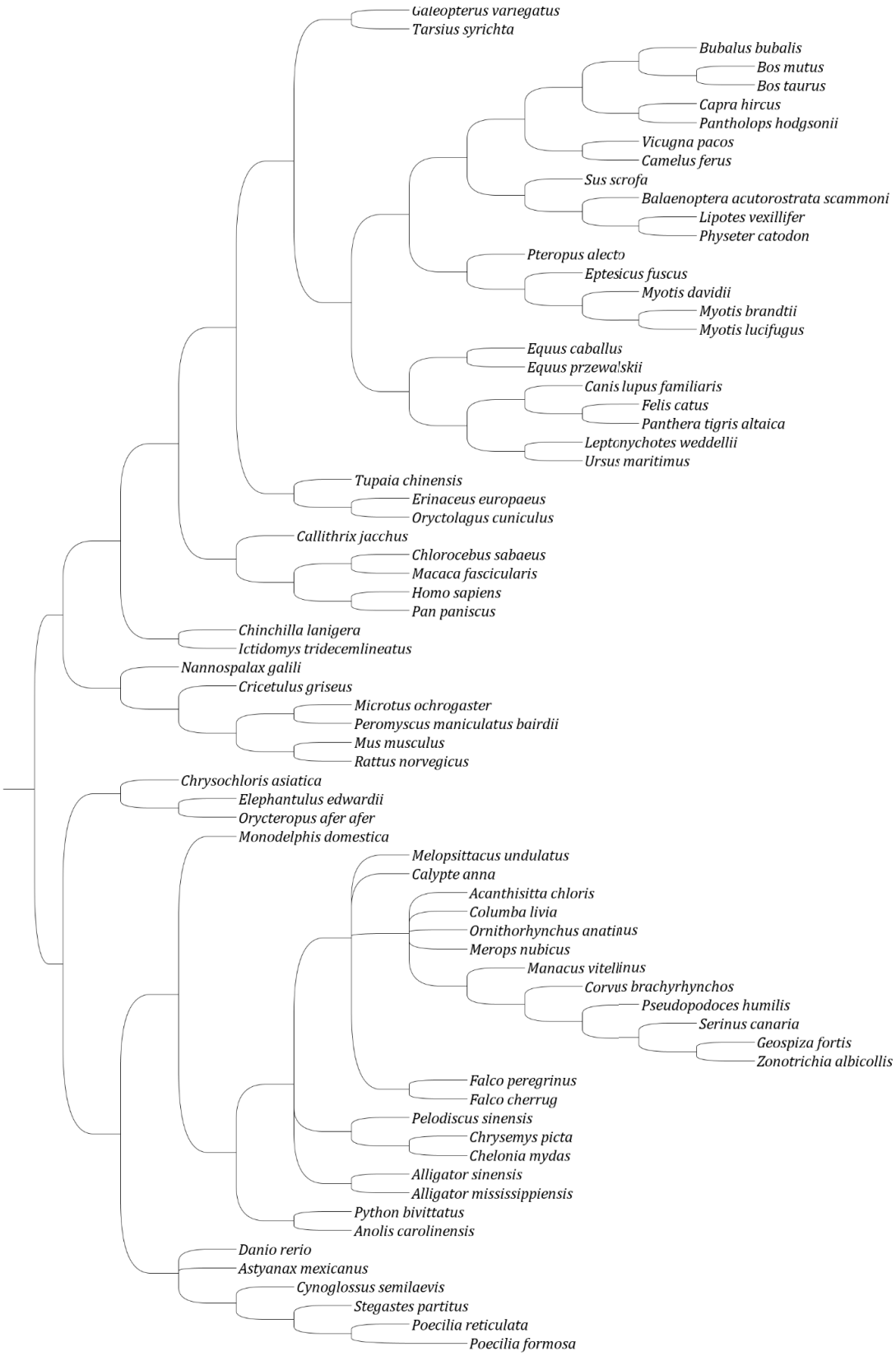

Galeopterus variegatus
Chinchilla lanigera
Ictidomys tridecemlineatus
Erinaceus europaeus
Tarsius syrichta
Tupaia chinensis
Oryctolagus cuniculus
Sus scrofa
Bos mutus
Bos taurus
Bubalus bubalis
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Felis catus
Panthera tigris altaica
Ursus maritimus
Leptonychotes weddellii
Canis lupus familiaris
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Monodelphis domestica
Ornithorhynchus anatinus
Pelodiscus sinensis
Melopsittacus undulatus
Calypte anna
Columba livia
Merops nubicus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Acanthisitta chloris
Manacus vitellinus
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Chrysemys picta
Chelonia mydas
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

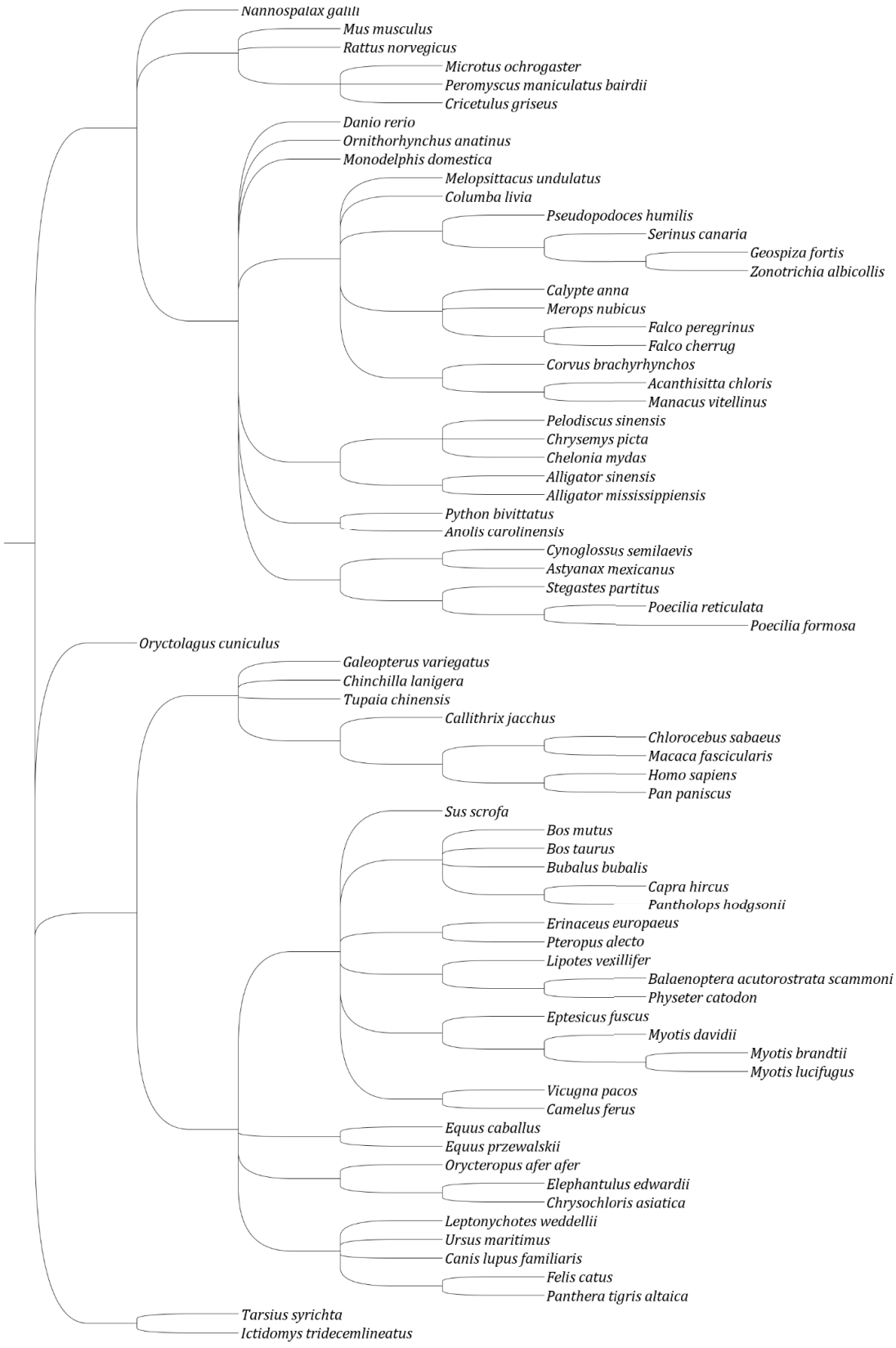

Tupaia chinensis
Galeopterus variegatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Sus scrofa
Bos mutus
Bos taurus
Bubalus bubalis
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Vicugna pacos
Camelus ferus
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Pteropus alecto
Equus caballus
Equus przewalskii
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Ictidomys tridecemlineatus
Oryctolagus cuniculus
Microtus ochrogaster
Mus musculus
Rattus norvegicus
Peromyscus maniculatus bairdii
Cricetulus griseus
Chinchilla lanigera
Nannospalax galili
Erinaceus europaeus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Monodelphis domestica
Astyanax mexicanus
Columba livia
Melopsittacus undulatus
Acanthisitta chloris
Merops nubicus
Geospiza fortis
Pseudopodoces humilis
Serinus canaria
Zonotrichia albicollis
Calypte anna
Falco peregrinus
Falco cherrug
Manacus vitellinus
Corvus brachyrhynchos
Chrysemys picta
Pelodiscus sinensis
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Danio rerio
Ornithorhynchus anatinus
Python bivittatus
Anolis carolinensis
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

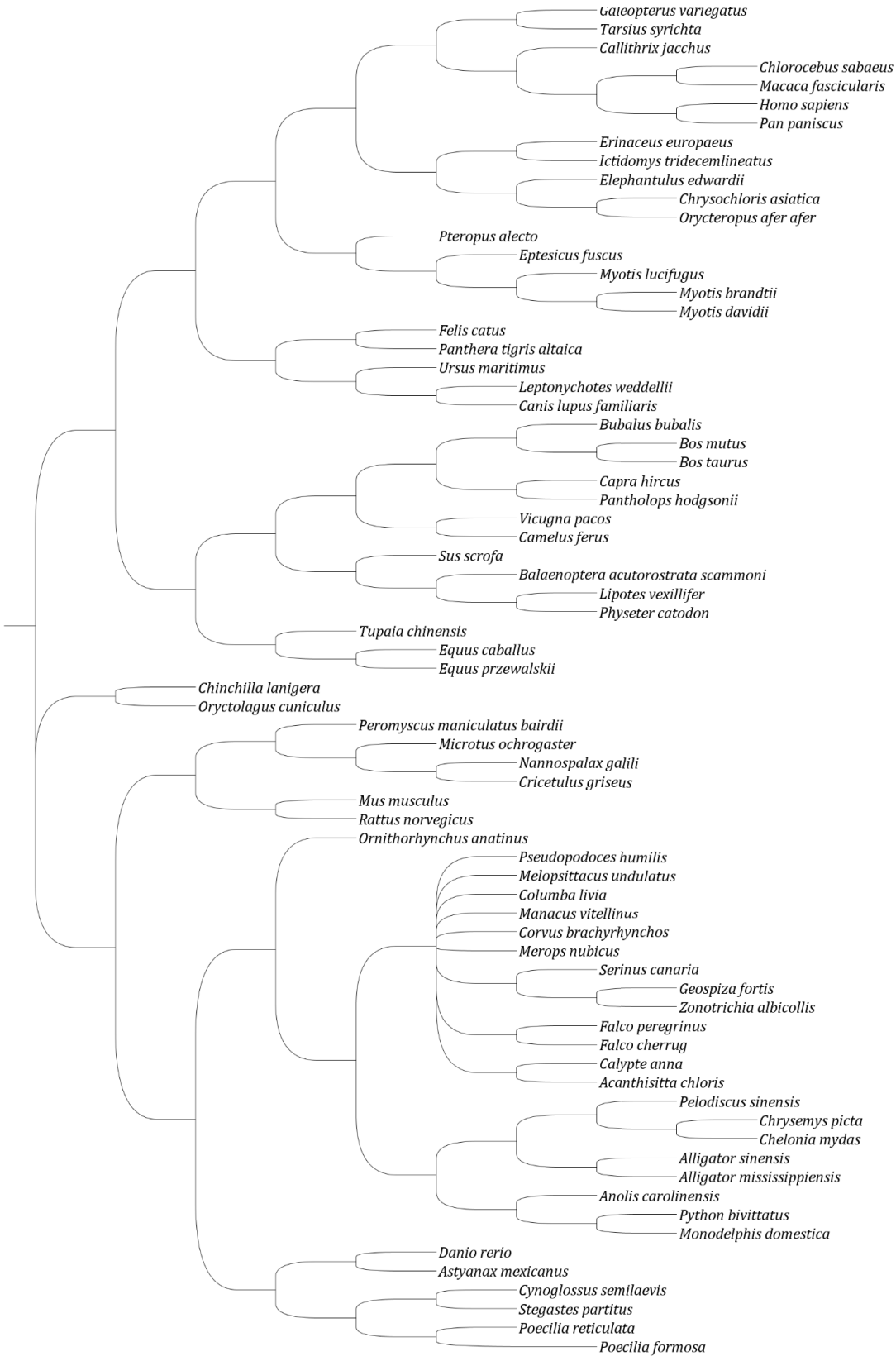

Alligator sinensis
Alligator mississippiensis
Ornithorhynchus anatinus
Cynoglossus semilaevis
Danio rerio
Astyanax mexicanus
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Sus scrofa
Erinaceus europaeus
Tarsius syrichta
Elephantulus edwardii
Chinchilla lanigera
Ictidomys tridecemlineatus
Oryctolagus cuniculus
Ursus maritimus
Pteropus alecto
Orycteropus afer afer
Chrysochloris asiatica
Galeopterus variegatus
Tupaia chinensis
Capra hircus
Panthelops hodgsonii
Bos mutus
Bos taurus
Bubalus bubalis
Nannospalax galili
Rattus norvegicus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Cricetulus griseus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Equus caballus
Equus przewalskii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Myotis brandtii
Myotis davidii
Eptesicus fuscus
Myotis lucifugus
Vicugna pacos
Camelus ferus
Canis lupus familiaris
Leptonychotes weddellii
Felis catus
Panthera tigris altaica
Melopsittacus undulatus
Calypte anna
Acanthisitta chloris
Columba livia
Merops nubicus
Geospiza fortis
Pseudopodoces humilis
Serinus canaria
Manacus vitellinus
Zonotrichia albicollis
Corvus brachyrhynchos
Falco peregrinus
Falco cherrug
Monodelphis domestica
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Python bivittatus
Anolis carolinensis

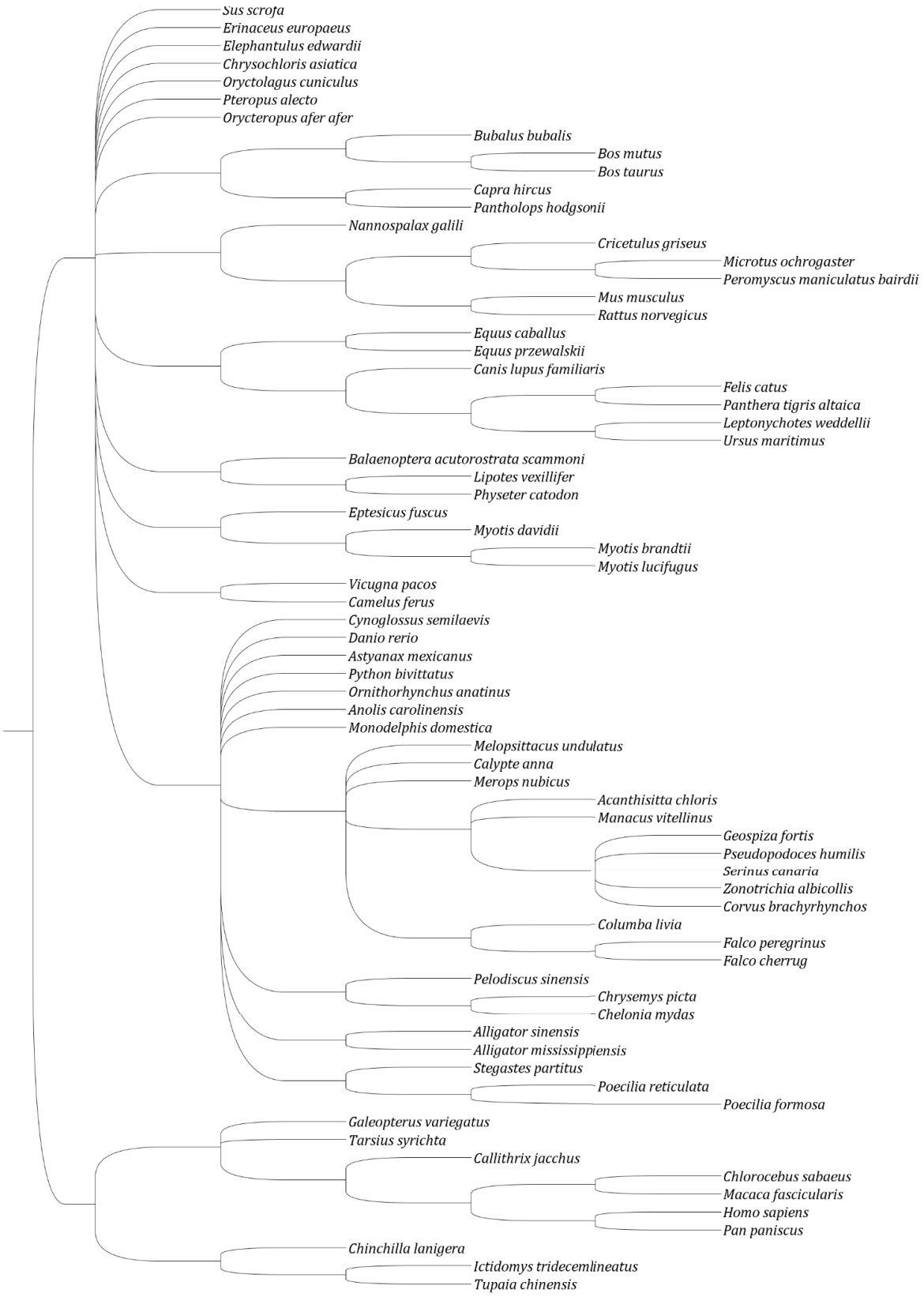
GAT


Chrysochloris asiatica
Galeopterus variegatus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Sus scrofa
Vicugna pacos
Camelus ferus
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Tarsius syrichta
Tupaia chinensis
Ictidomys tridecemlineatus
Chinchilla lanigera
Erinaceus europaeus
Oryctolagus cuniculus
Elephantulus edwardii
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Monodelphis domestica
Melopsittacus undulatus
Calypte anna
Acanthisitta chloris
Manacus vitellinus
Merops nubicus
Pseudopodoces humilis
Corvus brachyrhynchos
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Columba livia
Falco peregrinus
Falco cherrug
Ornithorhynchus anatinus
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

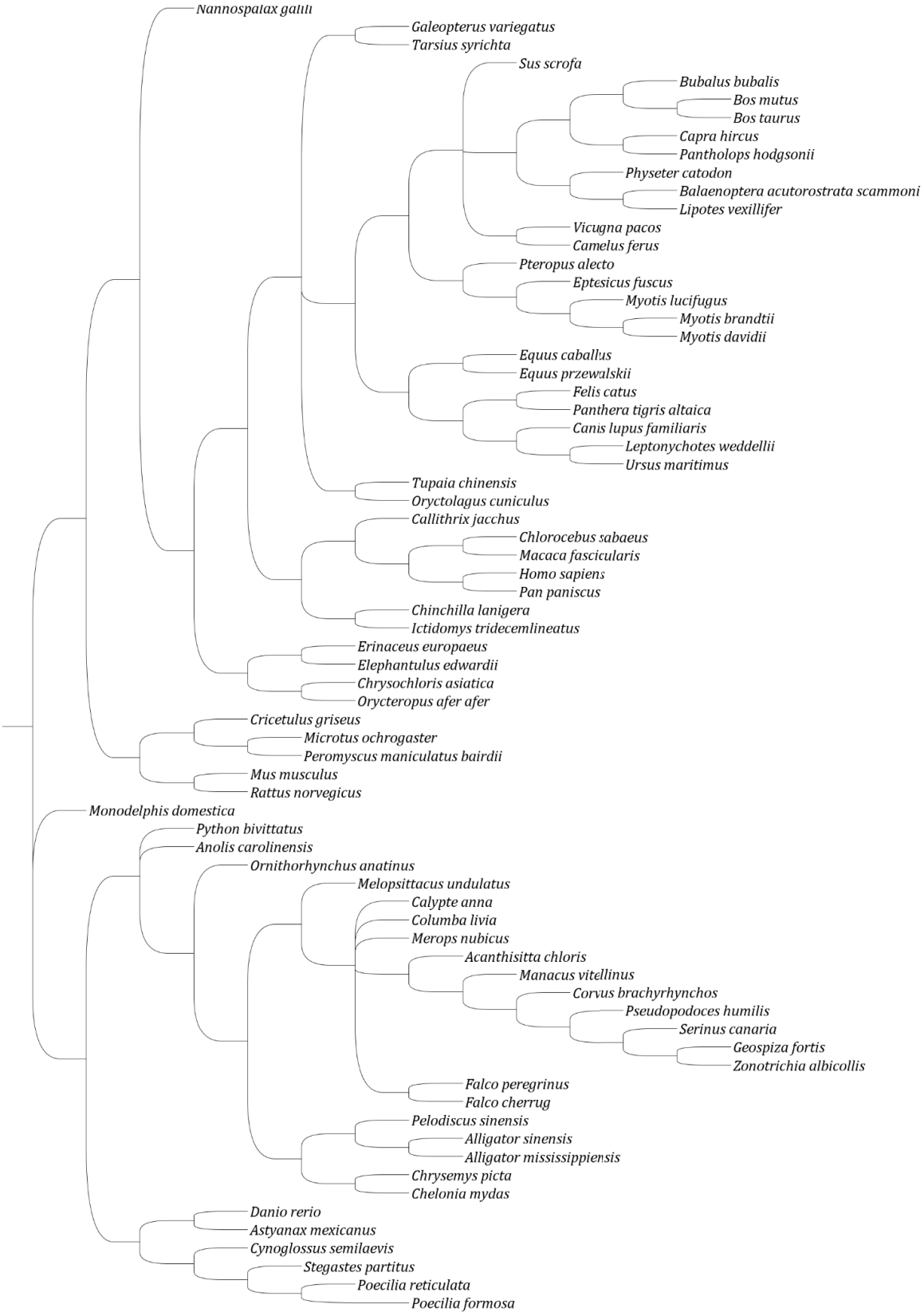

Galeopterus variegatus
Tarsius syrichta
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Vicugna pacos
Camelus ferus
Sus scrofa
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Equus caballus
Equus przewalskii
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Tupaia chinensis
Erinaceus europaeus
Oryctolagus cuniculus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Chinchilla lanigera
Ictidomys tridecemlineatus
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Monodelphis domestica
Melopsittacus undulatus
Calypte anna
Acanthisitta chloris
Columba livia
Ornithorhynchus anatinus
Merops nubicus
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

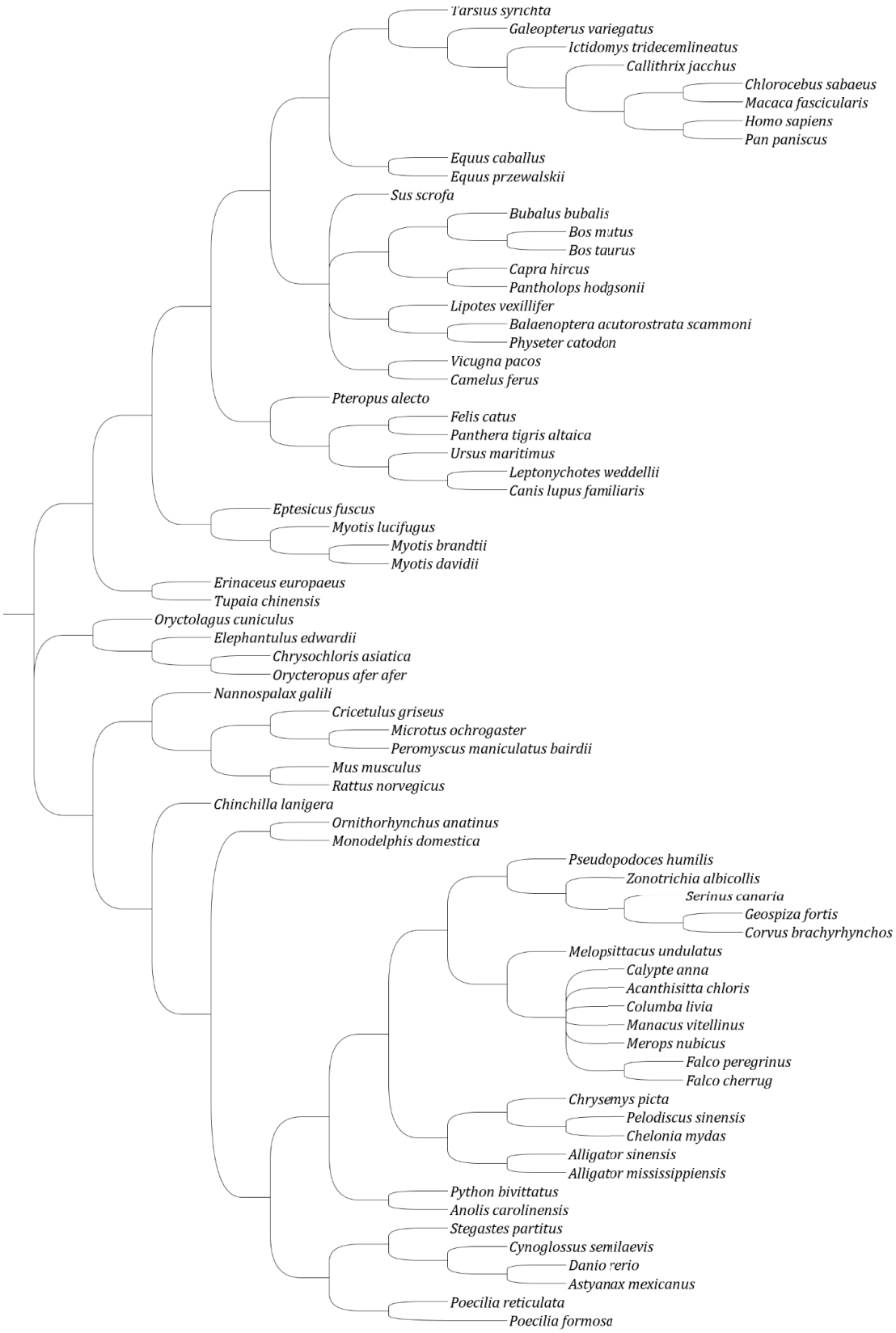

Nannospalax galili
Mus musculus
Rattus norvegicus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Danio rerio
Ornithorhynchus anatinus
Monodelphis domestica
Melopsittacus undulatus
Columba livia
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Calypte anna
Merops nubicus
Falco peregrinus
Falco cherrug
Corvus brachyrhynchos
Acanthisitta chloris
Manacus vitellinus
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Cynoglossus semilaevis
Astyanax mexicanus
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Oryctolagus cuniculus
Galeopterus variegatus
Chinchilla lanigera
Tupaia chinensis
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Sus scrofa
Bos mutus
Bos taurus
Bubalus bubalis
Capra hircus
Pantholops hodgsonii
Erinaceus europaeus
Pteropus alecto
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Orycteropus afer afer
Elephantulus edwardii
Chrysochloris asiatica
Leptonychotes weddellii
Ursus maritimus
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Tarsius syrichta
Ictidomys tridecemlineatus

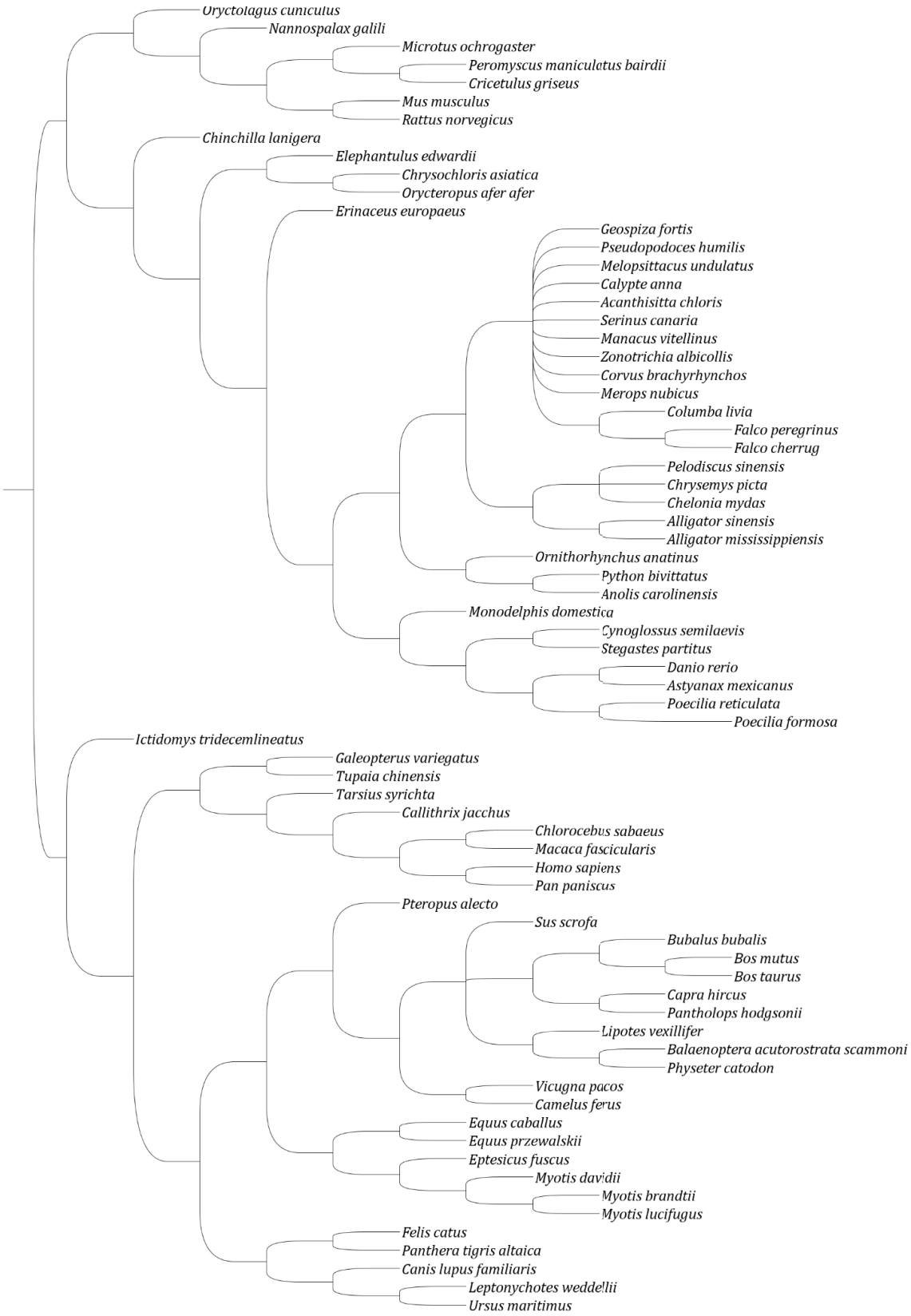

Galeopterus variegatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Erinaceus europaeus
Ictidomys tridecemlineatus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Felis catus
Panthera tigris altaica
Ursus maritimus
Leptonychotes weddellii
Canis lupus familiaris
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Vicugna pacos
Camelus ferus
Sus scrofa
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Tupaia chinensis
Equus caballus
Equus przewalskii
Chinchilla lanigera
Oryctolagus cuniculus
Peromyscus maniculatus bairdii
Microtus ochrogaster
Nannospalax galili
Cricetulus griseus
Mus musculus
Rattus norvegicus
Ornithorhynchus anatinus
Pseudopodoces humilis
Melopsittacus undulatus
Columba livia
Manacus vitellinus
Corvus brachyrhynchos
Merops nubicus
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Calypte anna
Acanthisitta chloris
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Anolis carolinensis
Python bivittatus
Monodelphis domestica
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

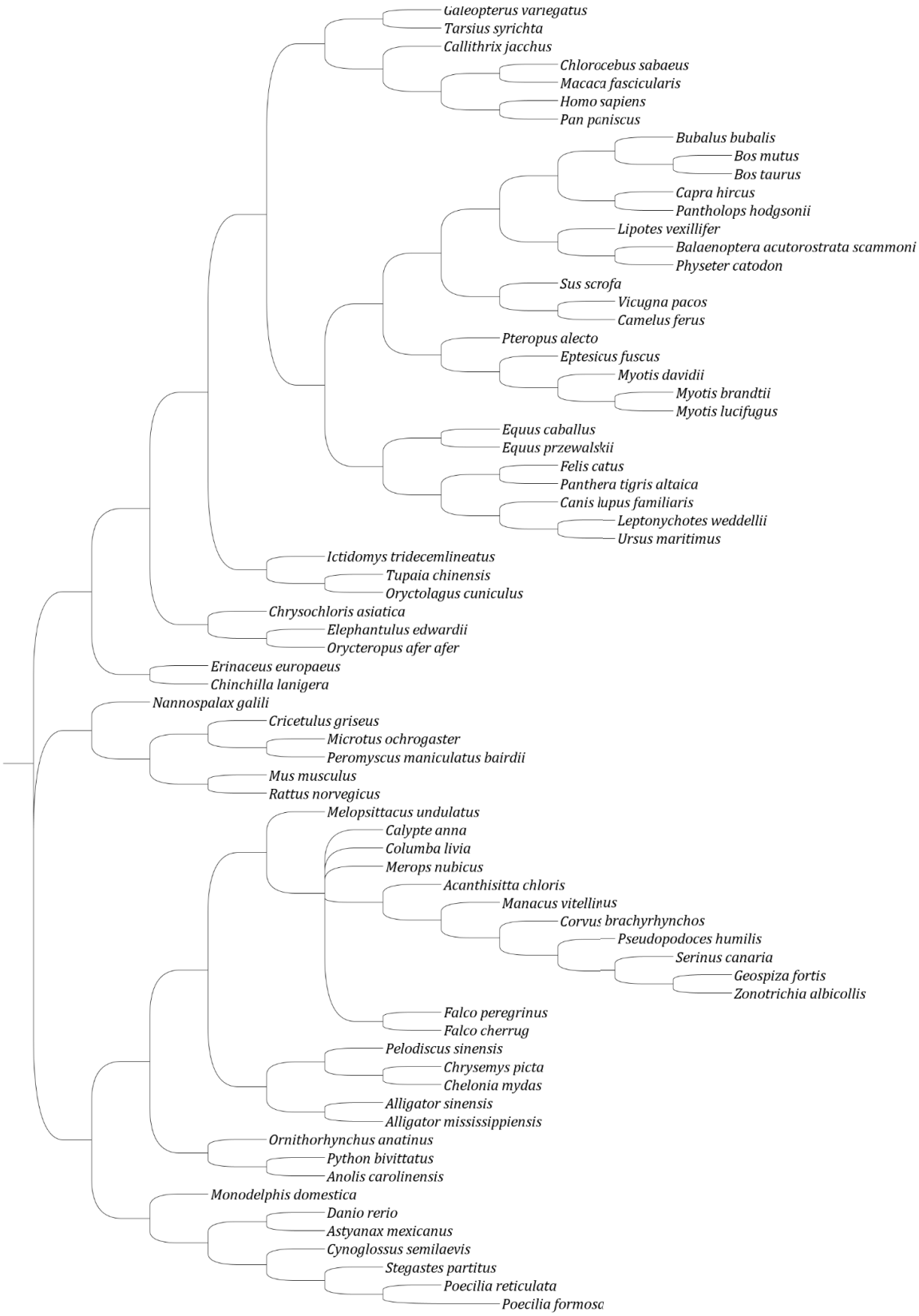

Sus scrofa
Erinaceus europaeus
Elephantulus edwardii
Chrysochloris asiatica
Oryctolagus cuniculus
Pteropus alecto
Orycteropus afer afer
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Equus caballus
Equus przewalskii
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Vicugna pacos
Camelus ferus
Cynoglossus semilaevis
Danio rerio
Astyanax mexicanus
Python bivittatus
Ornithorhynchus anatinus
Anolis carolinensis
Monodelphis domestica
Melopsittacus undulatus
Calypte anna
Merops nubicus
Acanthisitta chloris
Manacus vitellinus
Geospiza fortis
Pseudopodoces humilis
Serinus canaria
Zonotrichia albicollis
Corvus brachyrhynchos
Columba livia
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Galeopterus variegatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Chinchilla lanigera
Ictidomys tridecemlineatus
Tupaia chinensis

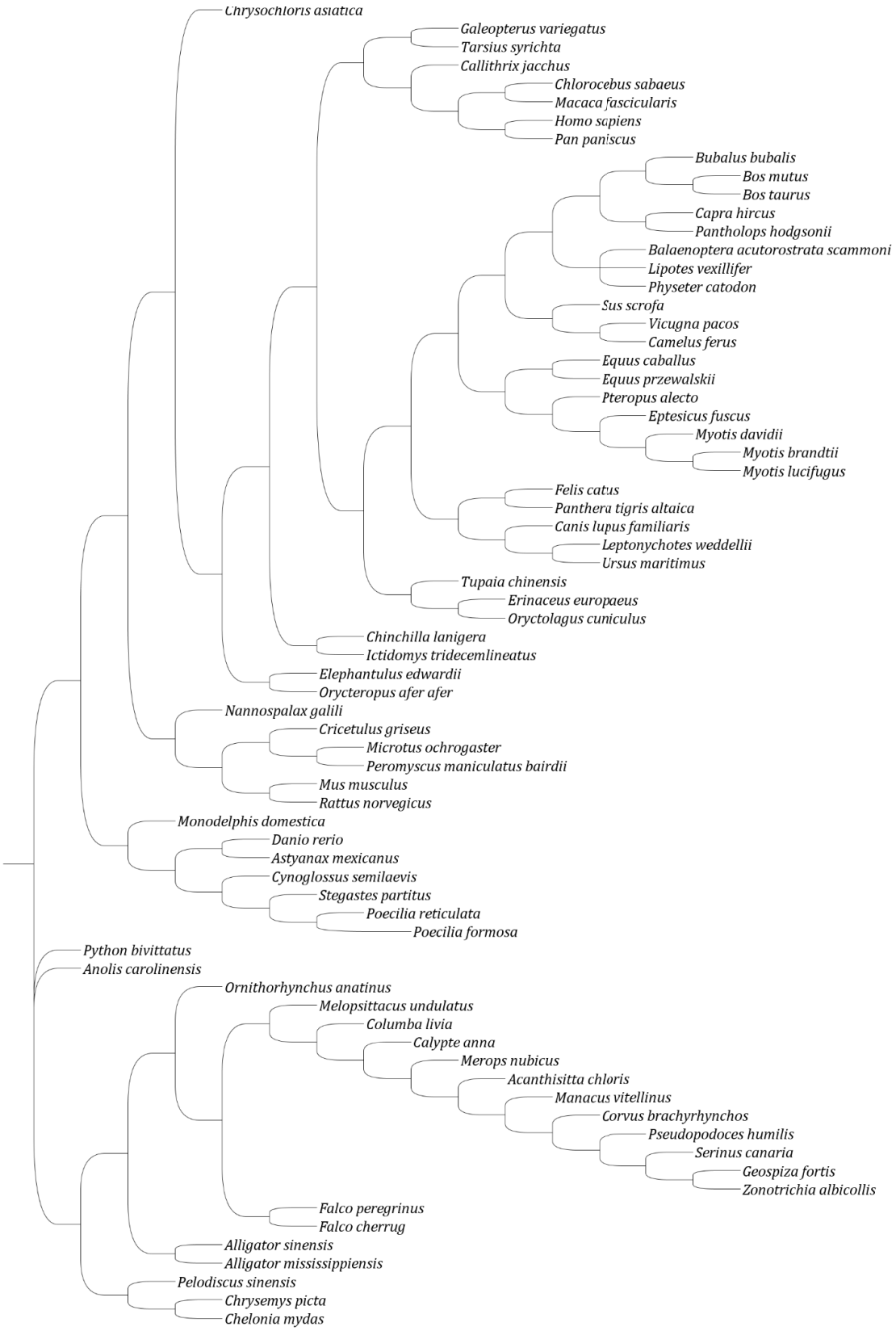
GGA

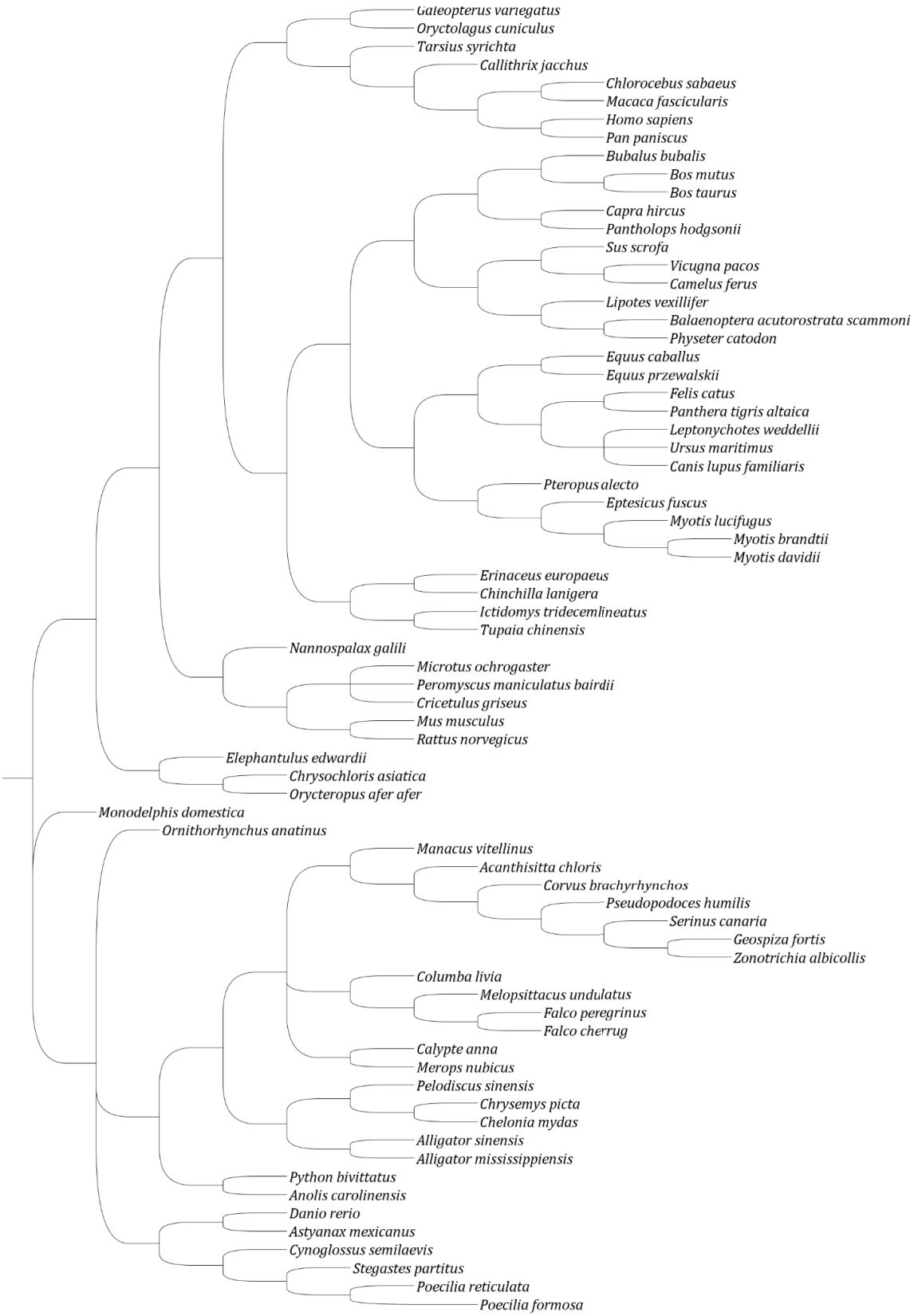

Nannospalax galili
Galeopterus variegatus
Tarsius syrichta
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Physeter catodon
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Vicugna pacos
Camelus ferus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Tupaia chinensis
Oryctolagus cuniculus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Chinchilla lanigera
Ictidomys tridecemlineatus
Erinaceus europaeus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Monodelphis domestica
Python bivittatus
Anolis carolinensis
Ornithorhynchus anatinus
Melopsittacus undulatus
Calypte anna
Columba livia
Merops nubicus
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Alligator sinensis
Alligator mississippiensis
Chrysemys picta
Chelonia mydas
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

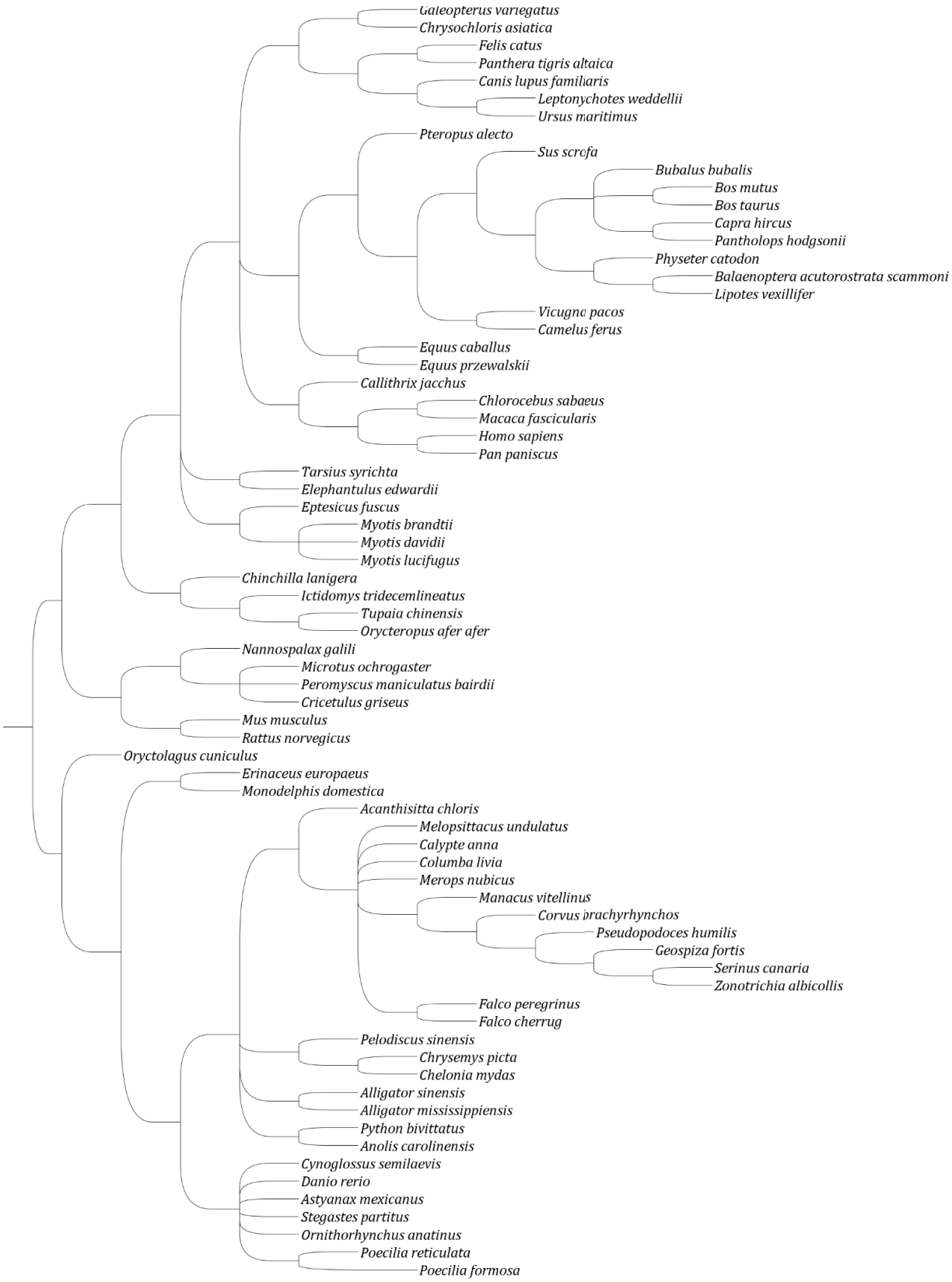

Tarsius syrichta
Galeopterus variegatus
Ictidomys tridecemlineatus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Equus caballus
Equus przewalskii
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Pteropus alecto
Felis catus
Panthera tigris altaica
Ursus maritimus
Leptonychotes weddellii
Canis lupus familiaris
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Erinaceus europaeus
Tupaia chinensis
Oryctolagus cuniculus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Chinchilla lanigera
Ornithorhynchus anatinus
Monodelphis domestica
Pseudopodoces humilis
Zonotrichia albicollis
Serinus canaria
Geospiza fortis
Corvus brachyrhynchos
Melopsittacus undulatus
Calypte anna
Acanthisitta chloris
Columba livia
Manacus vitellinus
Merops nubicus
Falco peregrinus
Falco cherrug
Chrysemys picta
Pelodiscus sinensis
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Stegastes partitus
Cynoglossus semilaevis
Danio rerio
Astyanax mexicanus
Poecilia reticulata
Poecilia formosa

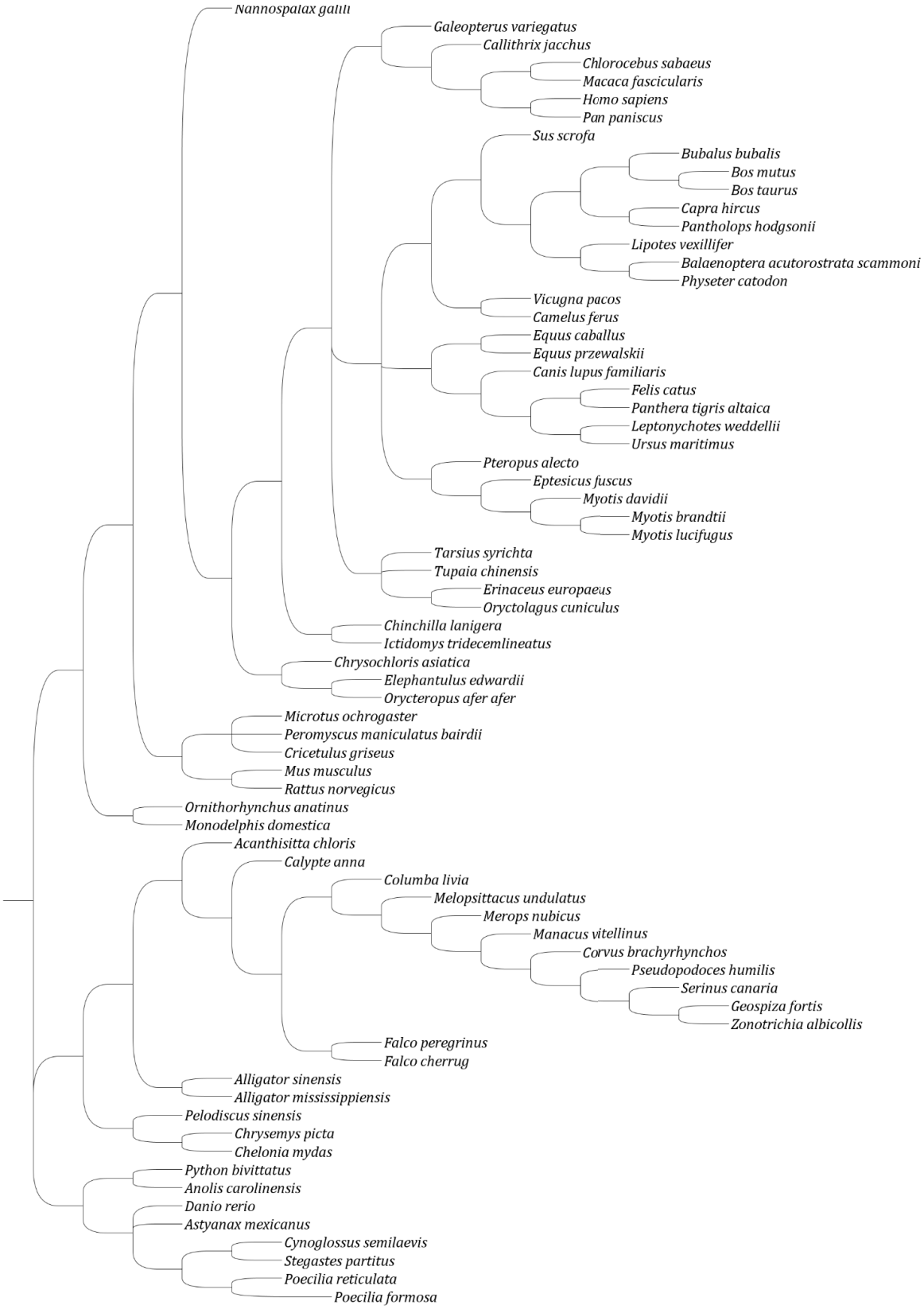

Oryctolagus cuniculus
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Chinchilla lanigera
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Erinaceus europaeus
Geospiza fortis
Pseudopodoces humilis
Melopsittacus undulatus
Calypte anna
Acanthisitta chloris
Serinus canaria
Manacus vitellinus
Zonotrichia albicollis
Corvus brachyrhynchos
Merops nubicus
Columba livia
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Ornithorhynchus anatinus
Python bivittatus
Anolis carolinensis
Monodelphis domestica
Cynoglossus semilaevis
Stegastes partitus
Danio rerio
Astyanax mexicanus
Poecilia reticulata
Poecilia formosa
Ictidomys tridecemlineatus
Galeopterus variegatus
Tupaia chinensis
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Pteropus alecto
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus

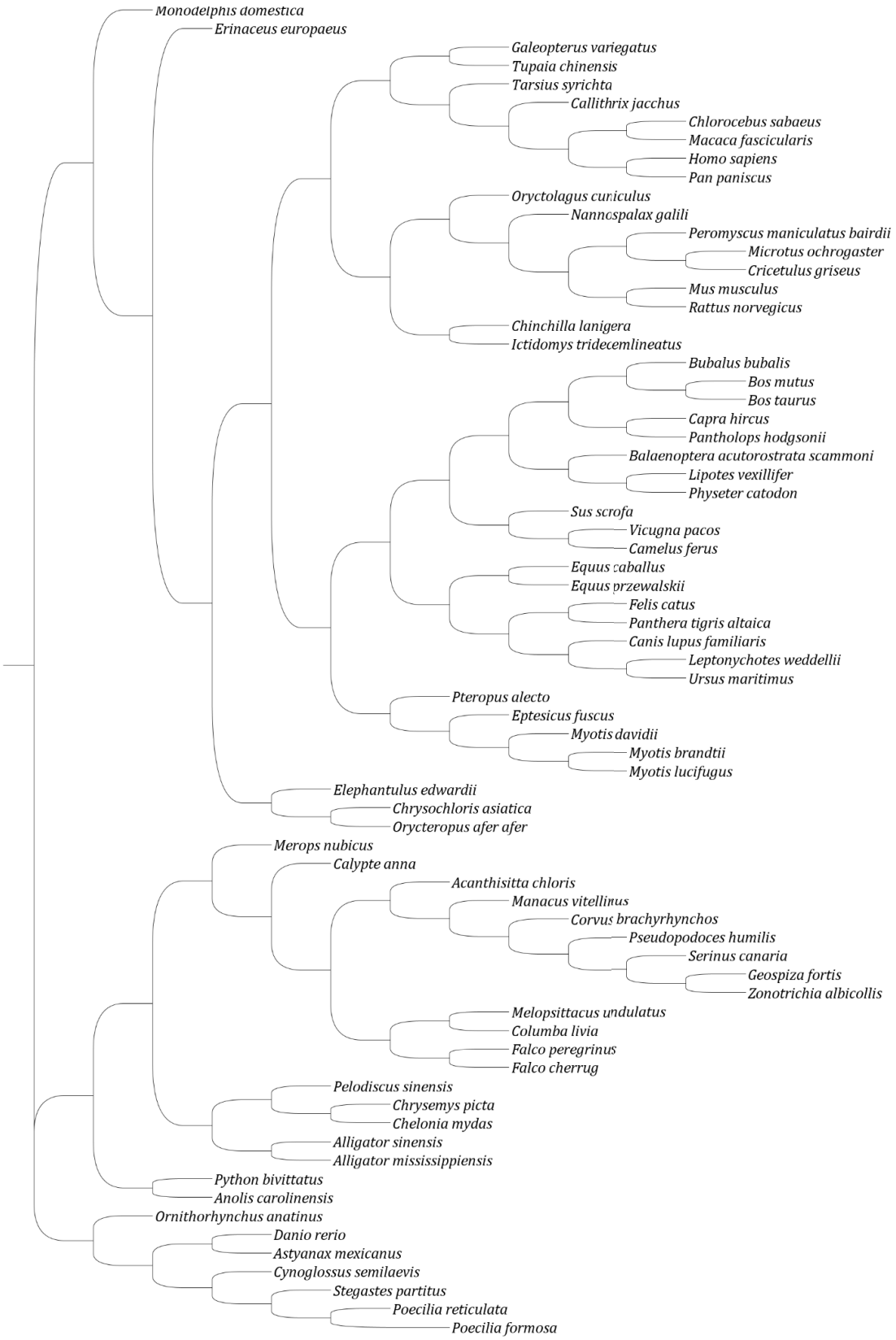

Galeopterus variegatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Ictidomys tridecemlineatus
Tupaia chinensis
Oryctolagus cuniculus
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Erinaceus europaeus
Chinchilla lanigera
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Melopsittacus undulatus
Calypte anna
Columba livia
Merops nubicus
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Ornithorhynchus anatinus
Python bivittatus
Anolis carolinensis
Monodelphis domestica
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

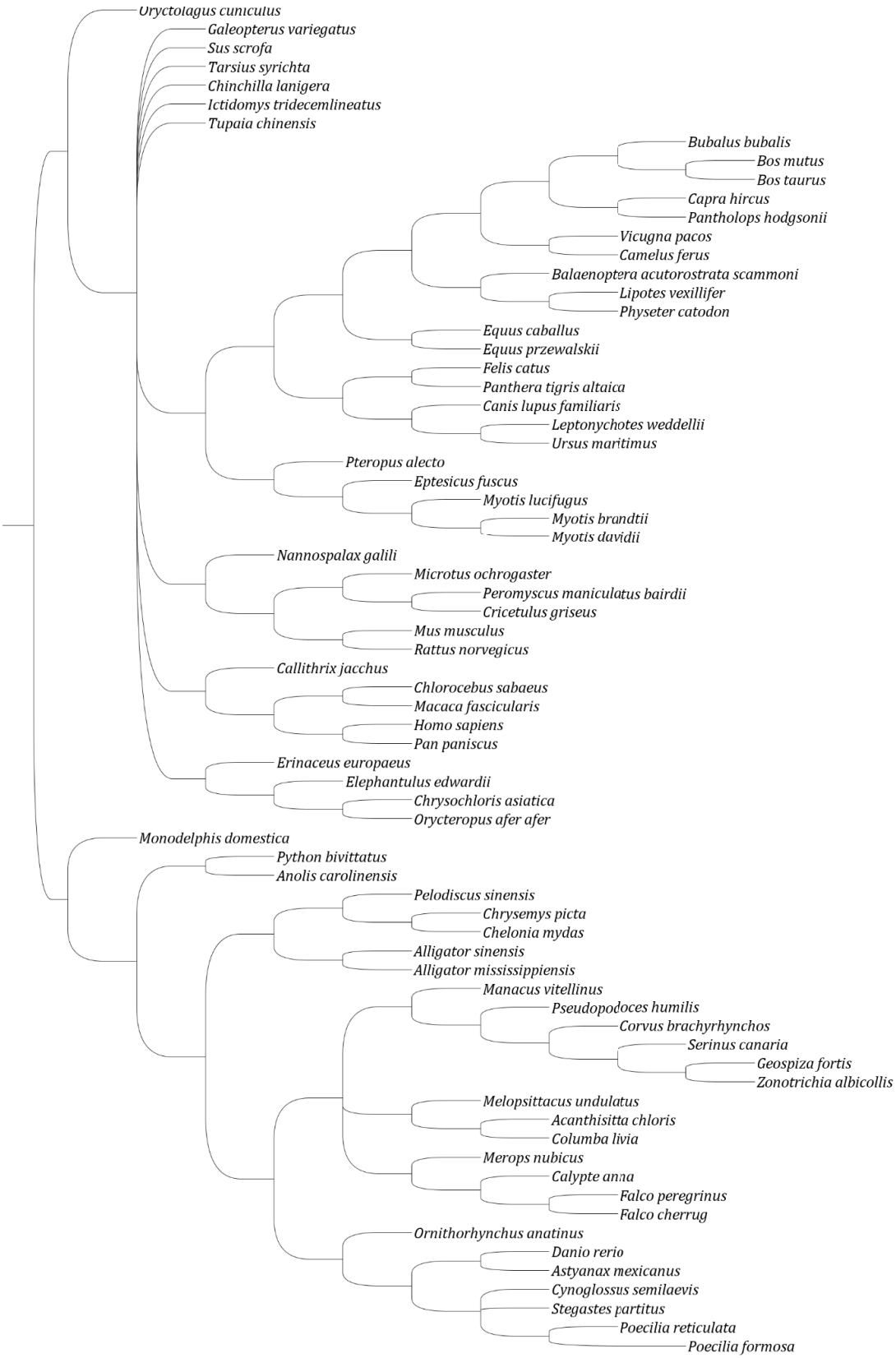

Chrysochloris asiatica
Galeopterus variegatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Tupaia chinensis
Erinaceus europaeus
Oryctolagus cuniculus
Chinchilla lanigera
Ictidomys tridecemlineatus
Elephantulus edwardii
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Monodelphis domestica
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Python bivittatus
Anolis carolinensis
Ornithorhynchus anatinus
Melopsittacus undulatus
Columba livia
Calypte anna
Merops nubicus
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas

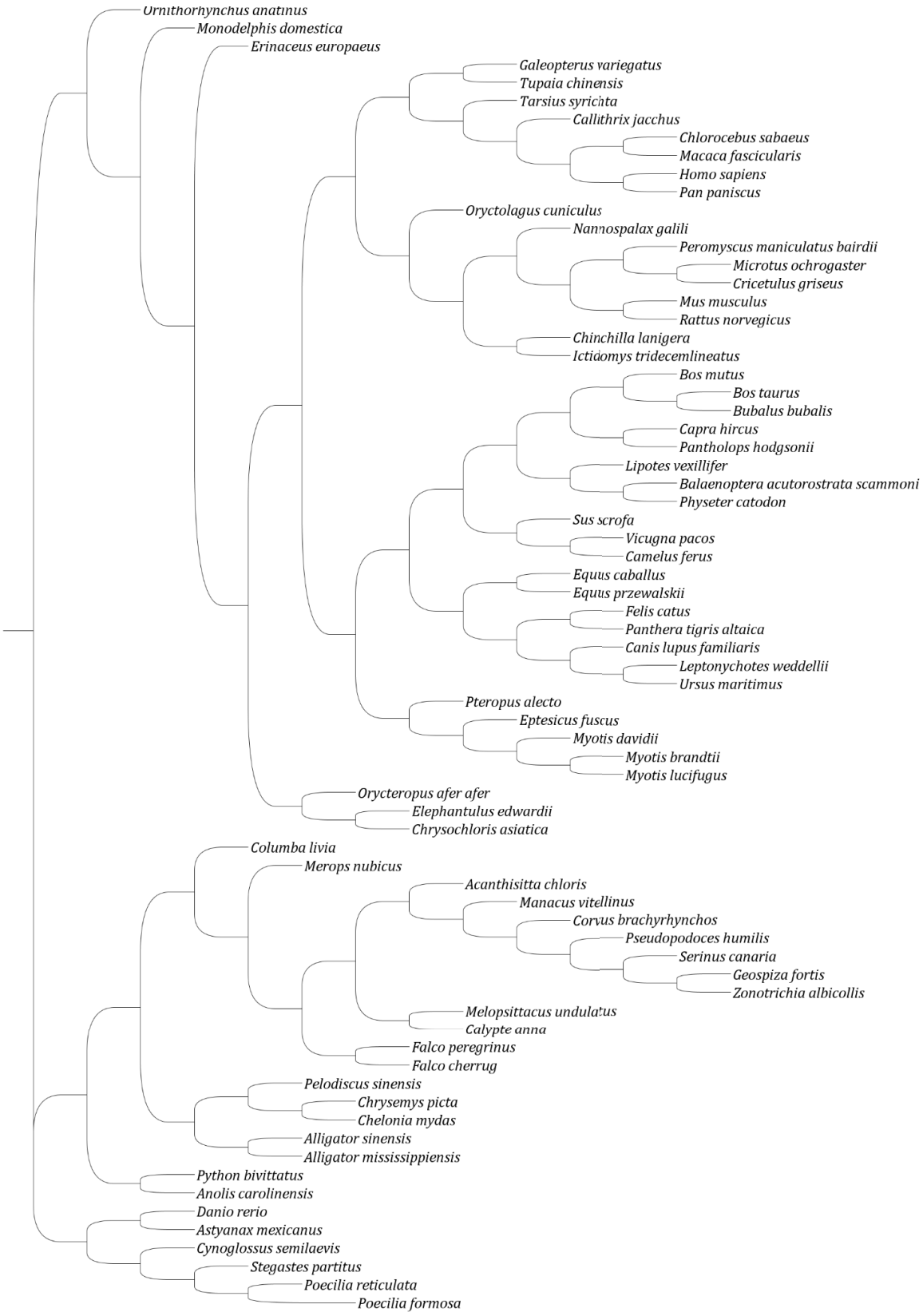

Galeopterus variegatus
Oryctolagus cuniculus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Sus scrofa
Vicugna pacos
Camelus ferus
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Canis lupus familiaris
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Erinaceus europaeus
Chinchilla lanigera
Ictidomys tridecemlineatus
Tupaia chinensis
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Monodelphis domestica
Ornithorhynchus anatinus
Manacus vitellinus
Acanthisitta chloris
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Columba livia
Melopsittacus undulatus
Falco peregrinus
Falco cherrug
Calypte anna
Merops nubicus
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

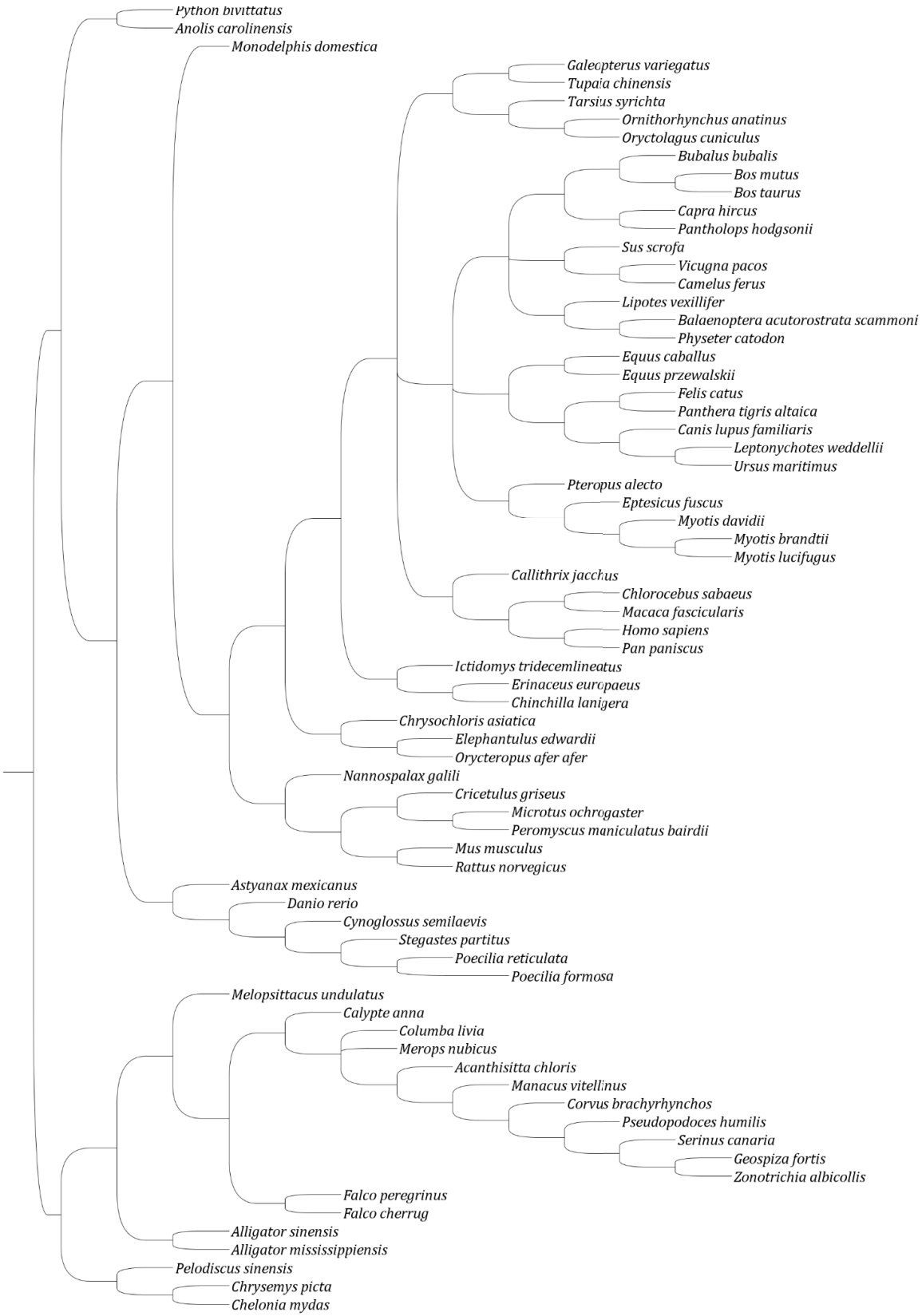

Galeopterus variegatus
Chrysochloris asiatica
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Physeter catodon
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Tarsius syrichta
Elephantulus edwardii
Eptesicus fuscus
Myotis brandtii
Myotis davidii
Myotis lucifugus
Chinchilla lanigera
Ictidomys tridecemlineatus
Tupaia chinensis
Orycteropus afer afer
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Oryctolagus cuniculus
Erinaceus europaeus
Monodelphis domestica
Acanthisitta chloris
Melopsittacus undulatus
Calypte anna
Columba livia
Merops nubicus
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Geospiza fortis
Serinus canaria
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Cynoglossus semilaevis
Danio rerio
Astyanax mexicanus
Stegastes partitus
Ornithorhynchus anatinus
Poecilia reticulata
Poecilia formosa

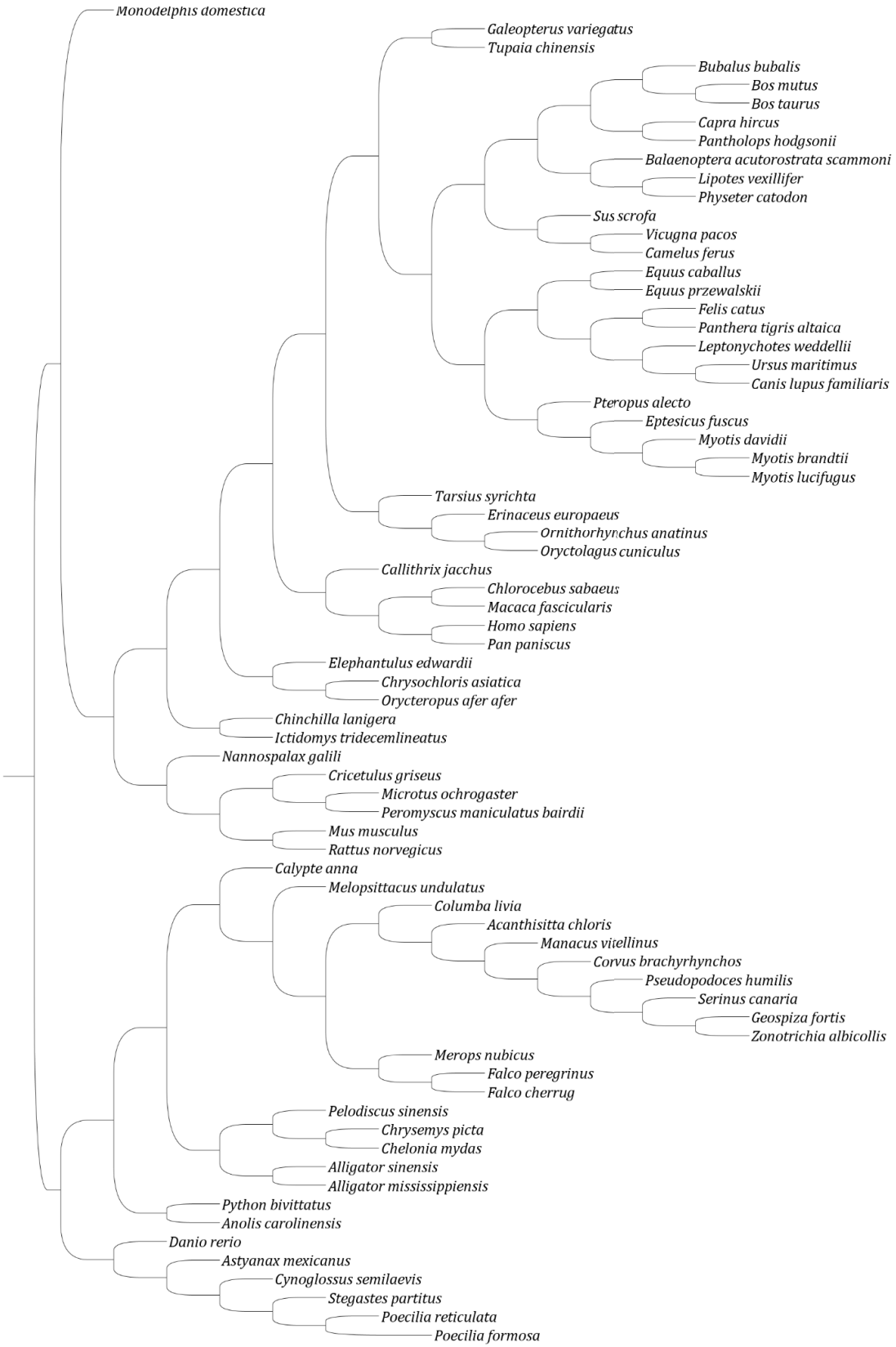

Nannospalax galili
Galeopterus variegatus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Tarsius syrichta
Tupaia chinensis
Erinaceus europaeus
Oryctolagus cuniculus
Chinchilla lanigera
Ictidomys tridecemlineatus
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Ornithorhynchus anatinus
Monodelphis domestica
Acanthisitta chloris
Calypte anna
Columba livia
Melopsittacus undulatus
Merops nubicus
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

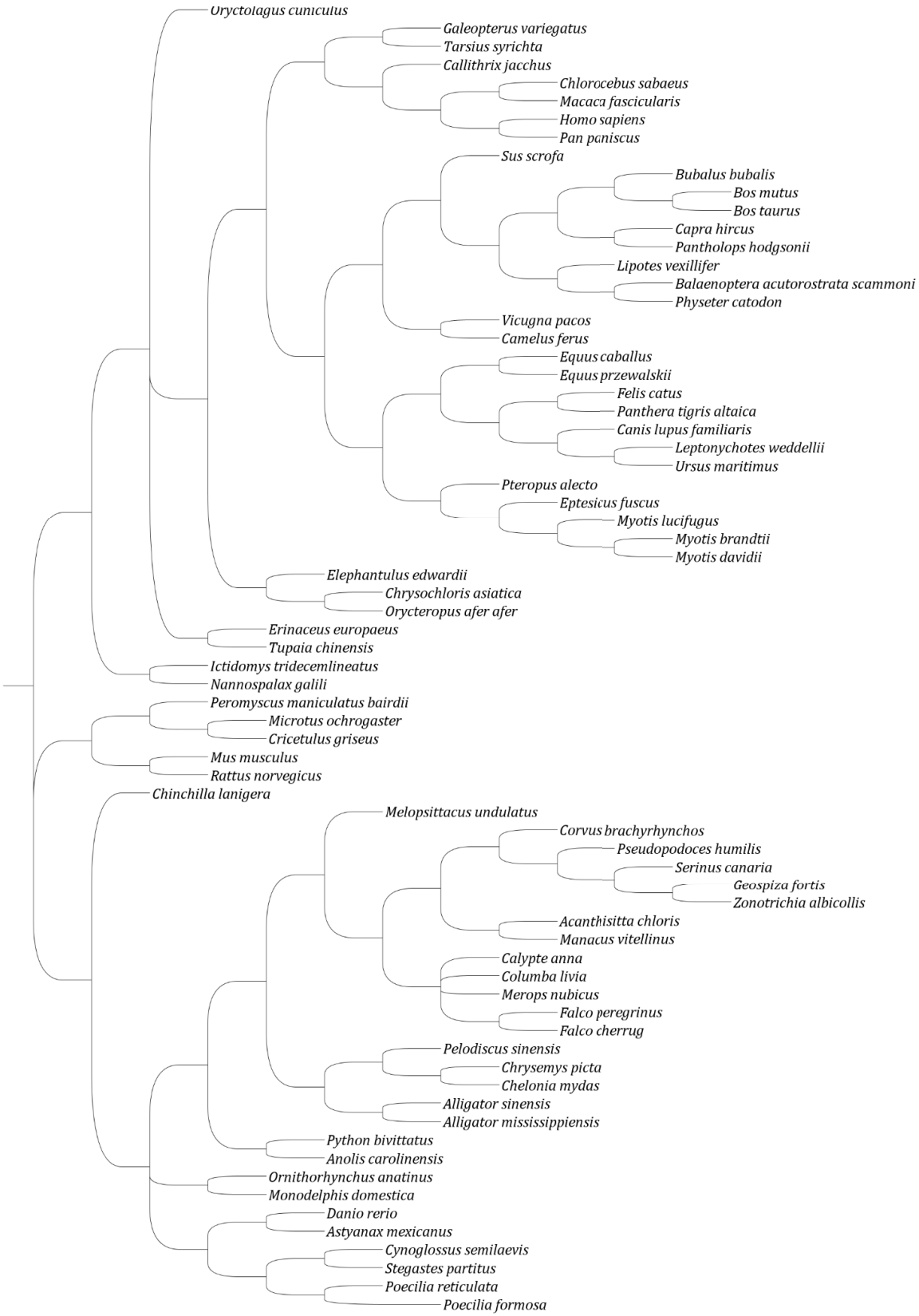

Monodelphis domestica
Erinaceus europaeus
Galeopterus variegatus
Tupaia chinensis
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Oryctolagus cuniculus
Nannospalax galili
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Chinchilla lanigera
Ictidomys tridecemlineatus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Merops nubicus
Calypte anna
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Melopsittacus undulatus
Columba livia
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Ornithorhynchus anatinus
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

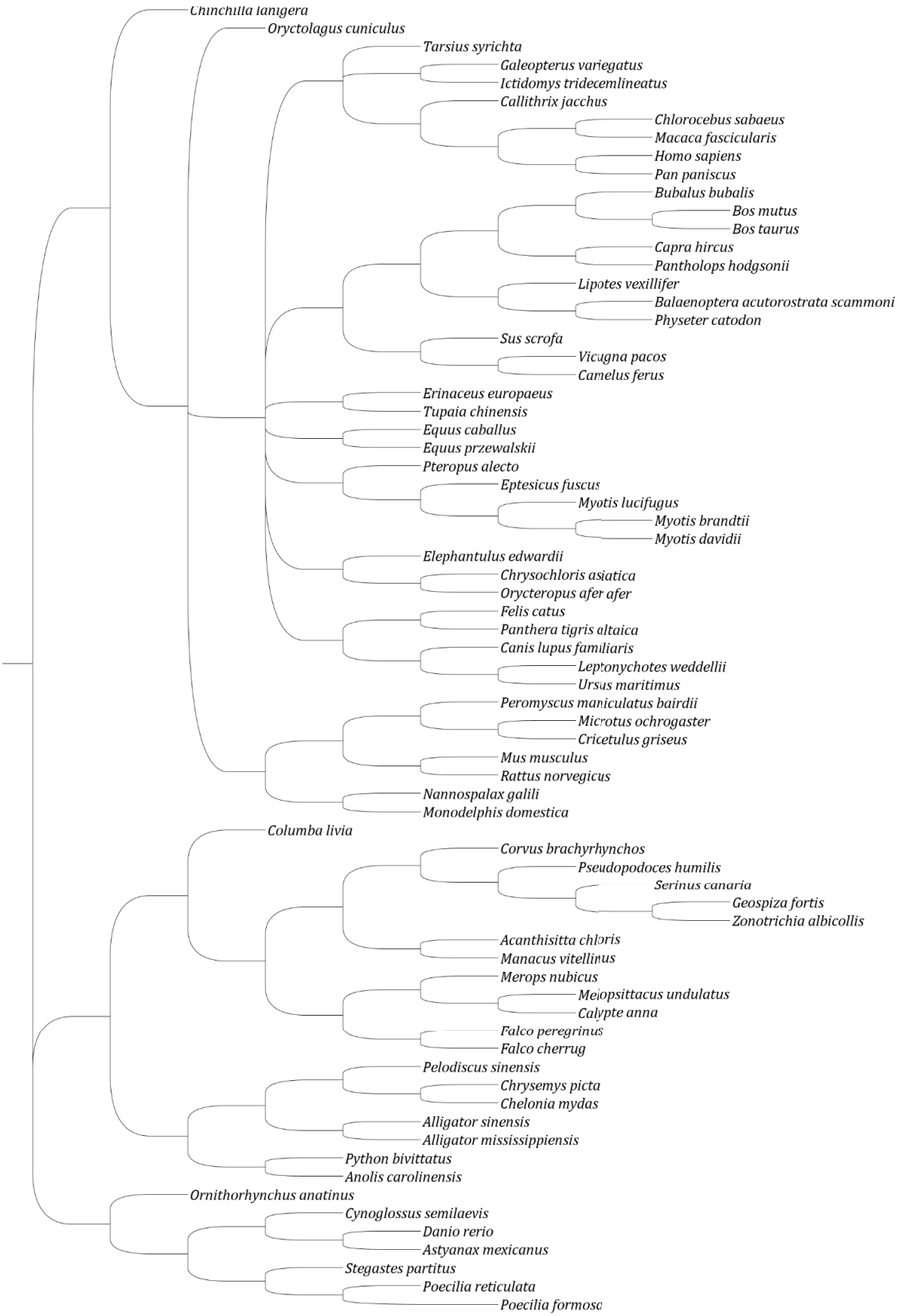

Oryctolagus cuniculus
Galeopterus variegatus
Sus scrofa
Tarsius syrichta
Chinchilla lanigera
Ictidomys tridecemlineatus
Tupaia chinensis
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Vicugna pacos
Camelus ferus
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Erinaceus europaeus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Monodelphis domestica
Python bivittatus
Anolis carolinensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Manacus vitellinus
Pseudopodoces humilis
Corvus brachyrhynchos
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Melopsittacus undulatus
Acanthisitta chloris
Columba livia
Merops nubicus
Calypte anna
Falco peregrinus
Falco cherrug
Ornithorhynchus anatinus
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

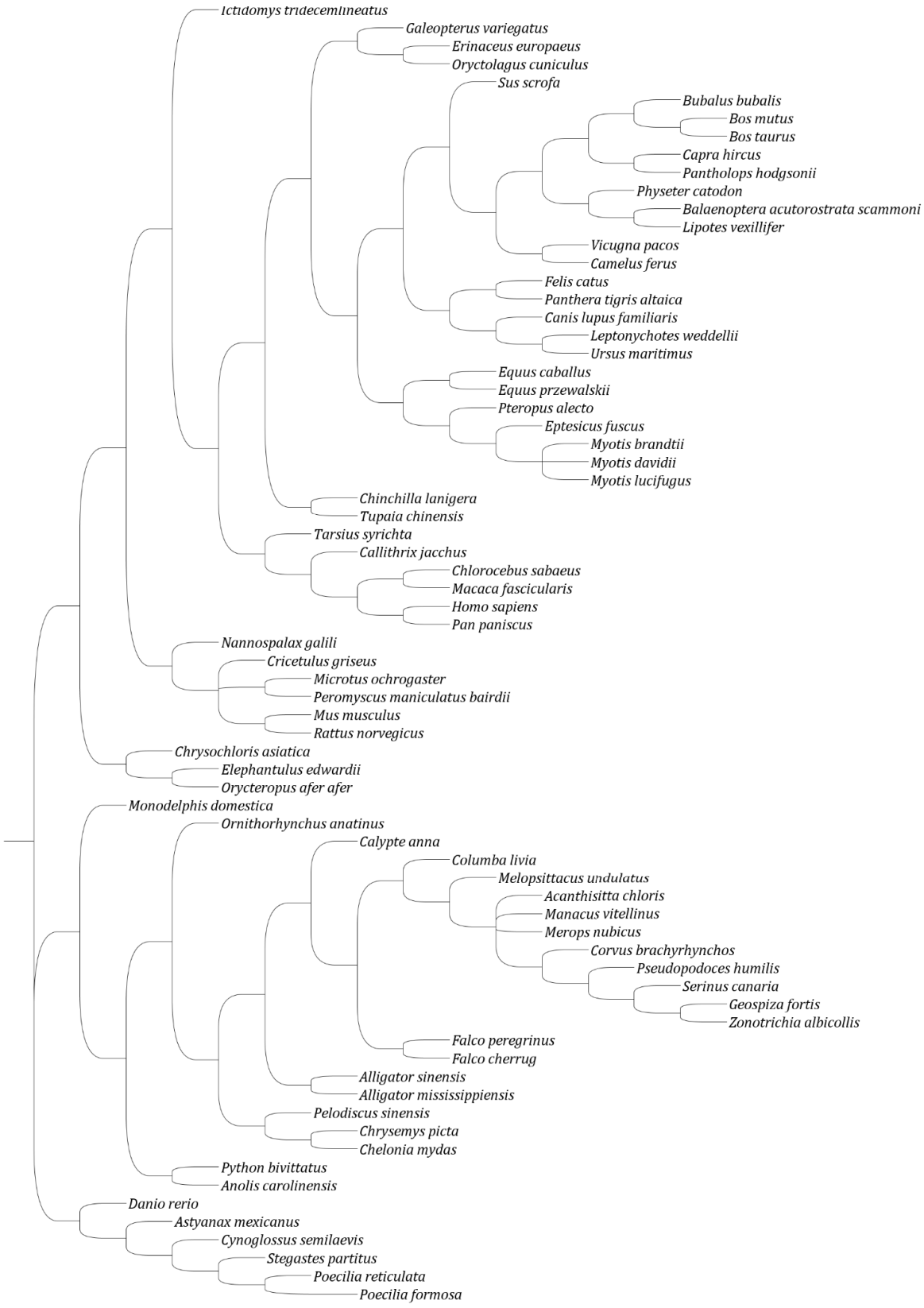

Ornithorhynchus anatinus
Monodelphis domestica
Erinaceus europaeus
Galeopterus variegatus
Tupaia chinensis
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Oryctolagus cuniculus
Nannospalax galili
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Chinchilla lanigera
Ictidomys tridecemlineatus
Bos mutus
Bos taurus
Bubalus bubalis
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Orycteropus afer afer
Elephantulus edwardii
Chrysochloris asiatica
Columba livia
Merops nubicus
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Melopsittacus undulatus
Calypte anna
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

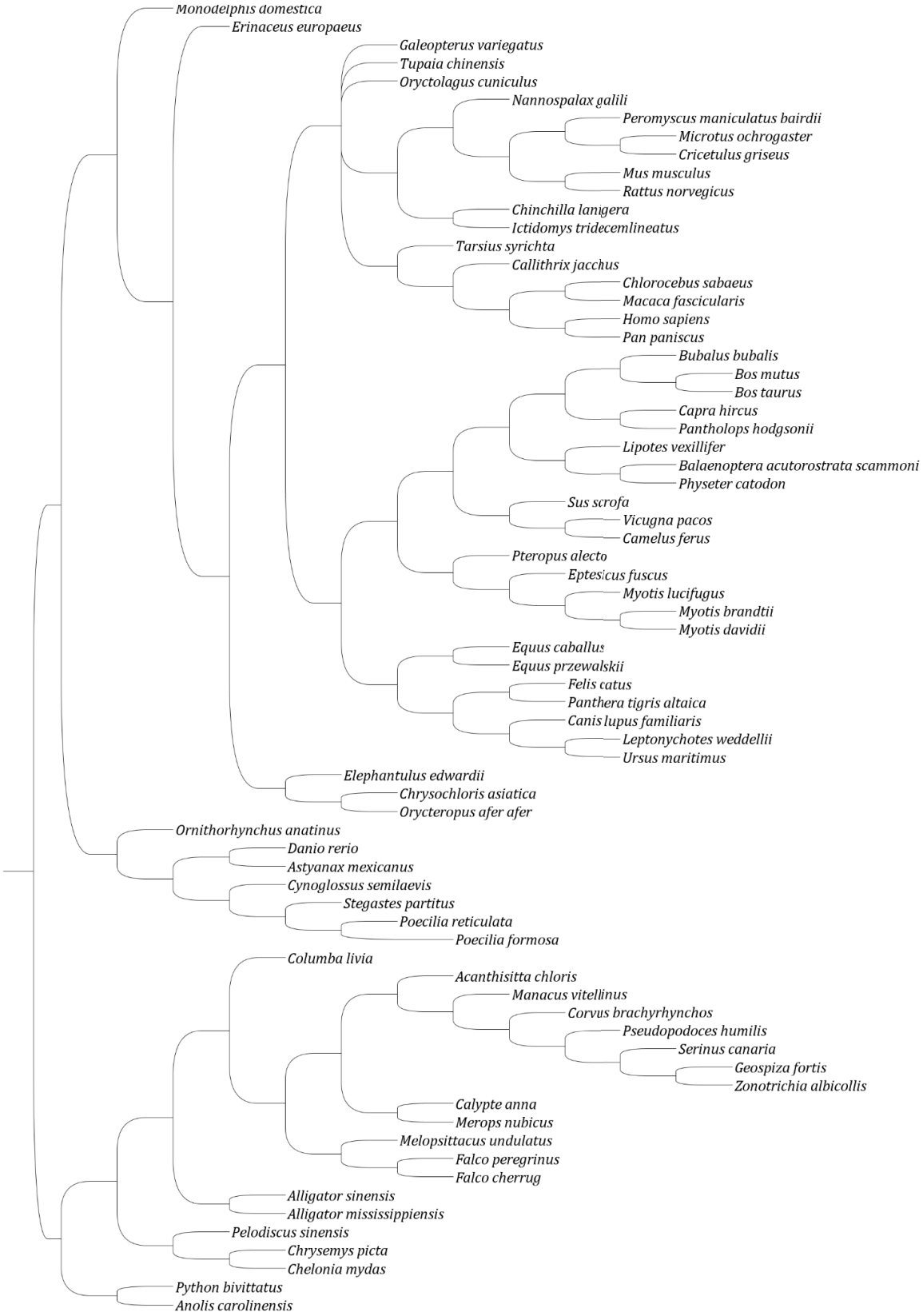

Python bivittatus
Anolis carolinensis
Monodelphis domestica
Galeopterus variegatus
Tupaia chinensis
Tarsius syrichta
Ornithorhynchus anatinus
Oryctolagus cuniculus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Sus scrofa
Vicugna pacos
Camelus ferus
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Ictidomys tridecemlineatus
Erinaceus europaeus
Chinchilla lanigera
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Astyanax mexicanus
Danio rerio
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Melopsittacus undulatus
Calypte anna
Columba livia
Merops nubicus
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas

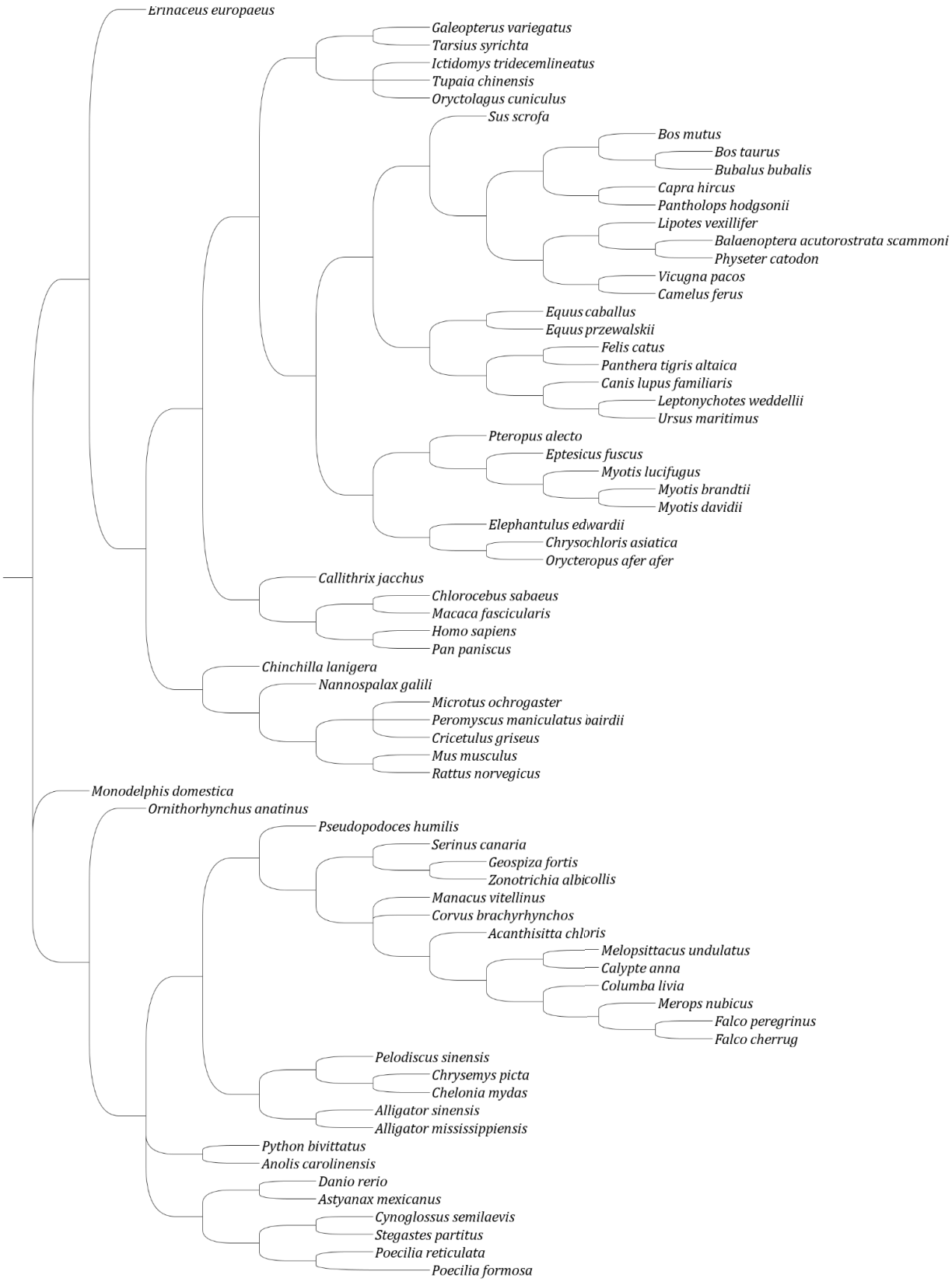

Monodelphis domestica
Galeopterus variegatus
Tupaia chinensis
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Canis lupus familiaris
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Tarsius syrichta
Erinaceus europaeus
Ornithorhynchus anatinus
Oryctolagus cuniculus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Chinchilla lanigera
Ictidomys tridecemlineatus
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Calypte anna
Melopsittacus undulatus
Columba livia
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Merops nubicus
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

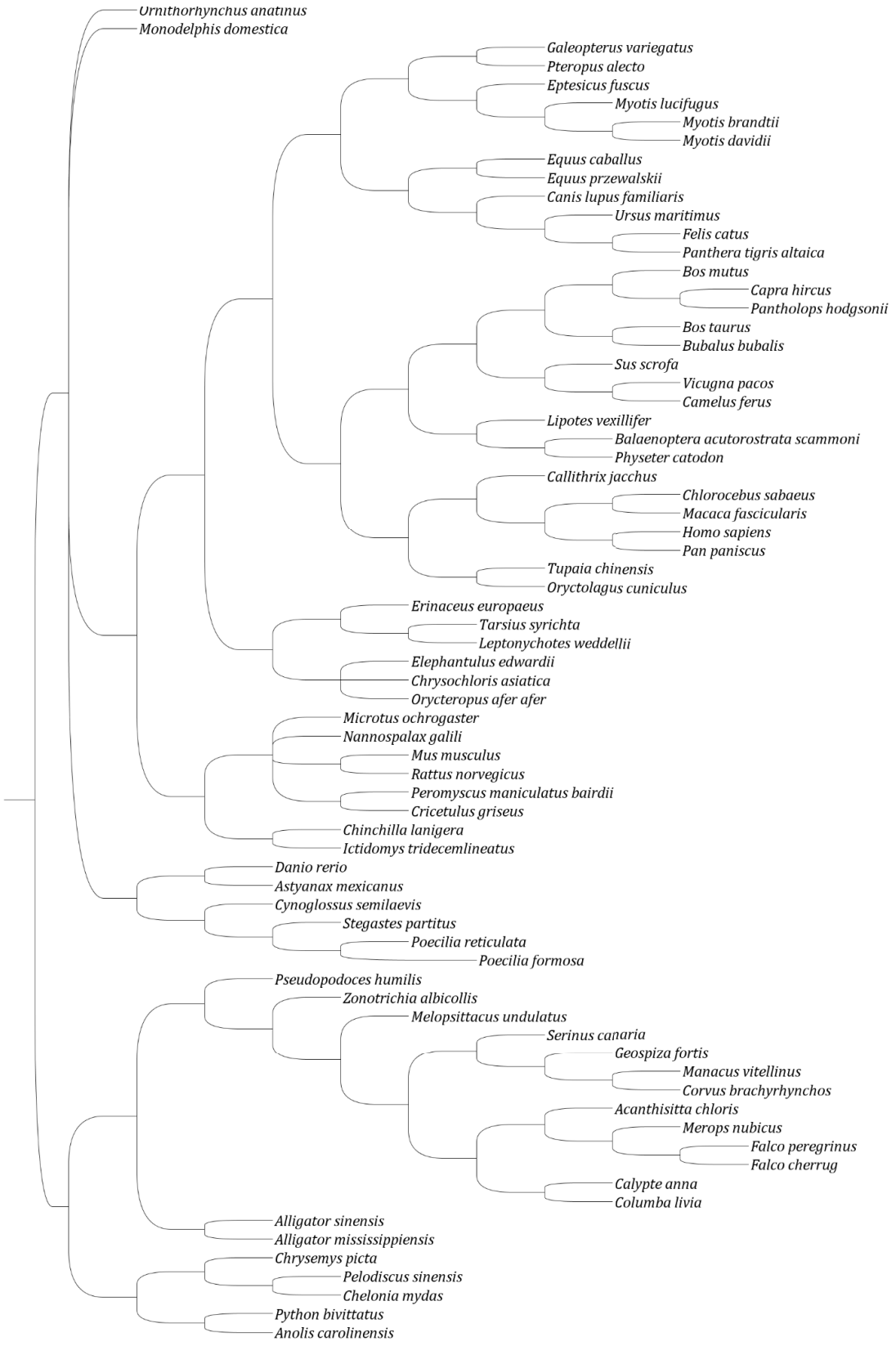

Oryctolagus cuniculus
Galeopterus variegatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Erinaceus europaeus
Tupaia chinensis
Ictidomys tridecemlineatus
Nannospalax galili
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Chinchilla lanigera
Melopsittacus undulatus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Acanthisitta chloris
Manacus vitellinus
Calypte anna
Columba livia
Merops nubicus
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Ornithorhynchus anatinus
Monodelphis domestica
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

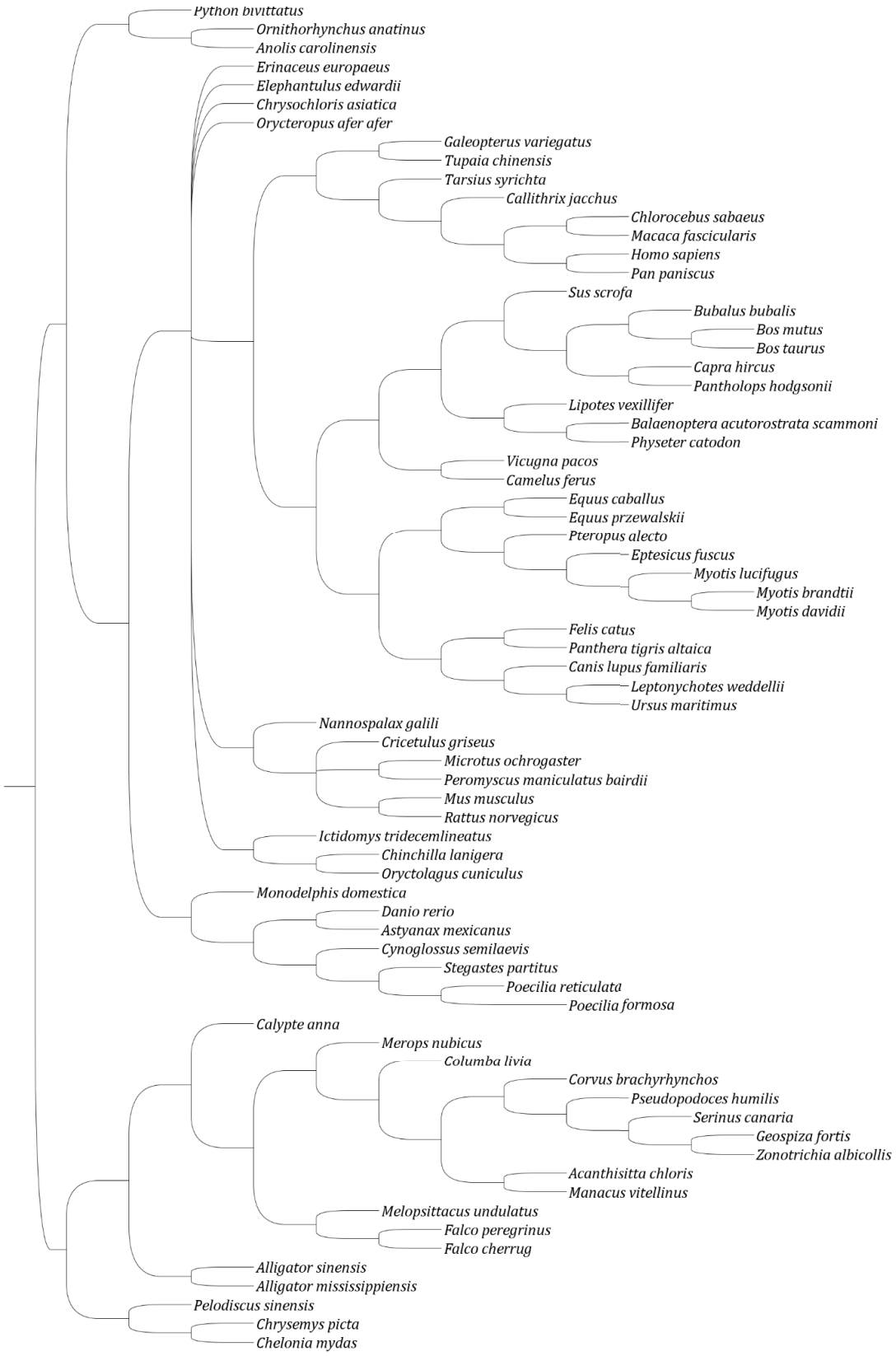

Chinchilla lanigera
Oryctolagus cuniculus
Tarsius syrichta
Galeopterus variegatus
Ictidomys tridecemlineatus
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Erinaceus europaeus
Tupaia chinensis
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Nannospalax galili
Monodelphis domestica
Columba livia
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Acanthisitta chloris
Manacus vitellinus
Merops nubicus
Meiopsittacus undulatus
Calypte anna
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Ornithorhynchus anatinus
Cynoglossus semilaevis
Danio rerio
Astyanax mexicanus
Stegastes partitus
Poecilia reticulata
Poecilia formosa

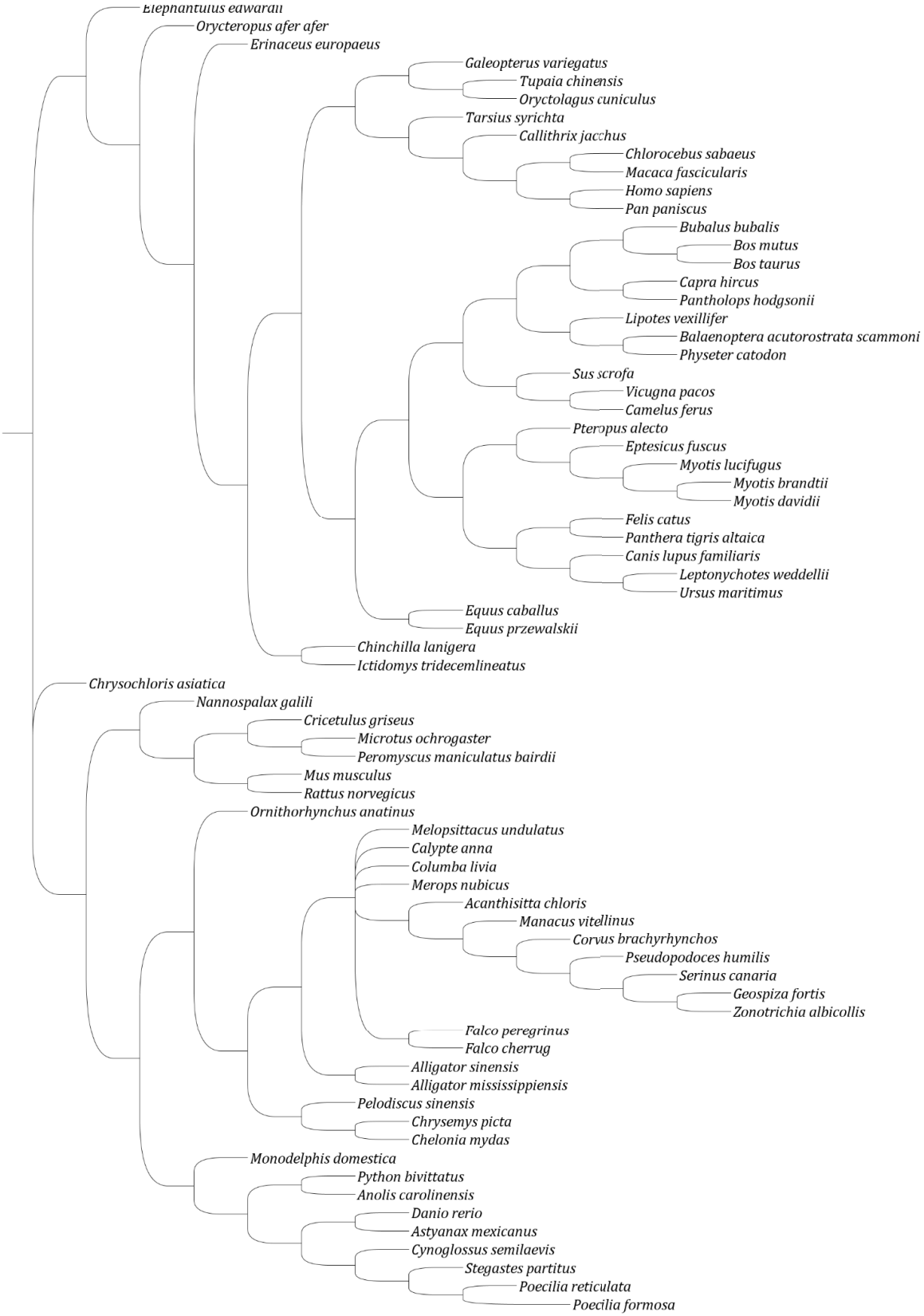

Ictidomys tridecemlineatus
Galeopterus variegatus
Erinaceus europaeus
Oryctolagus cuniculus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Physeter catodon
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Vicugna pacos
Camelus ferus
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis brandtii
Myotis davidii
Myotis lucifugus
Chinchilla lanigera
Tupaia chinensis
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Chrysochloris asiatica
Elephantulus edwardii
Orycteropus afer afer
Monodelphis domestica
Ornithorhynchus anatinus
Calypte anna
Columba livia
Melopsittacus undulatus
Acanthisitta chloris
Manacus vitellinus
Merops nubicus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

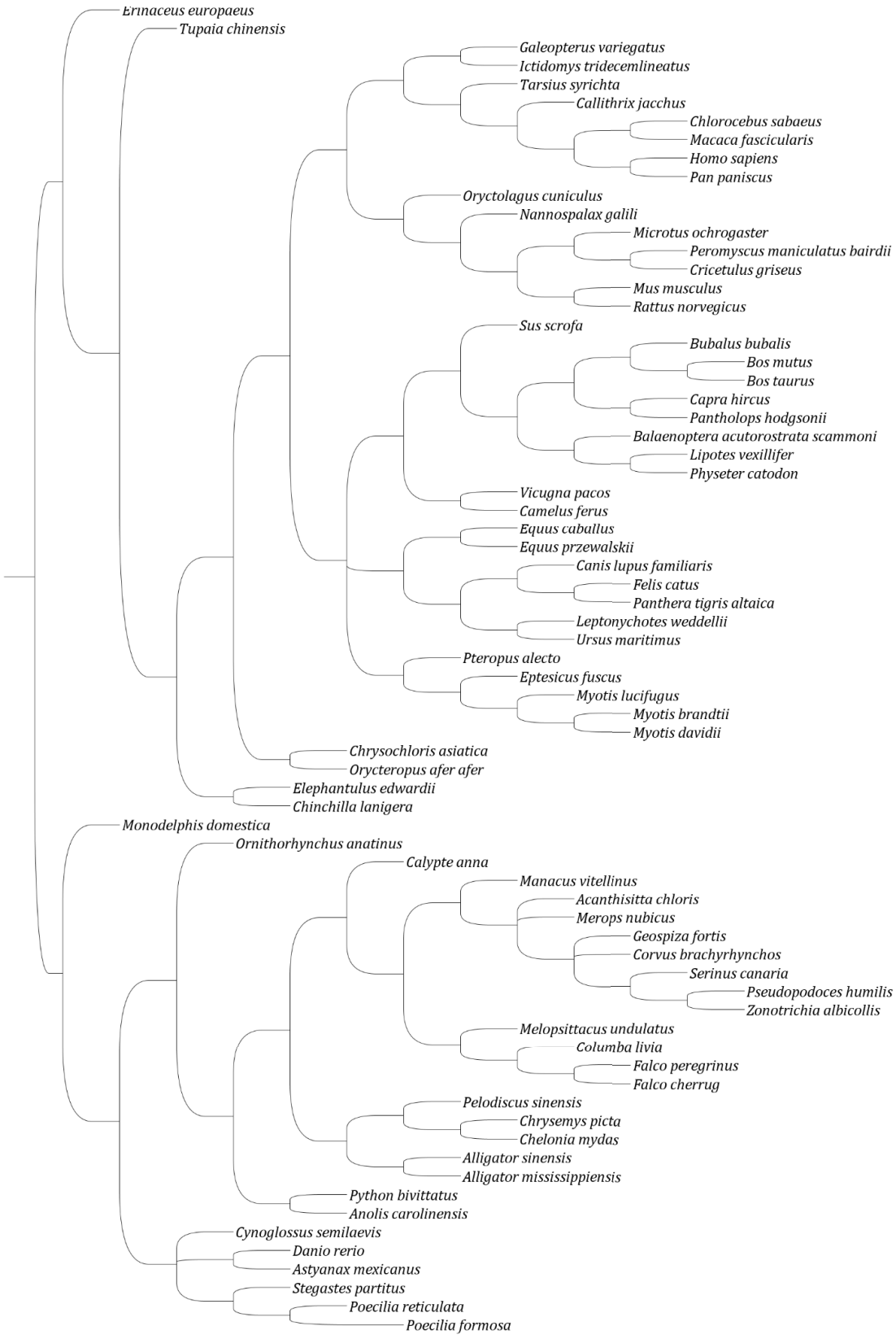

Monodelphis domestica
Erinaceus europaeus
Galeopterus variegatus
Tupaia chinensis
Oryctolagus cuniculus
Nannospalax galili
Peromyscus maniculatus bairdii
Microtus ochrogaster
Cricetulus griseus
Mus musculus
Rattus norvegicus
Chinchilla lanigera
Ictidomys tridecemlineatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Ornithorhynchus anatinus
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Columba livia
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Calypte anna
Merops nubicus
Melopsittacus undulatus
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Python bivittatus
Anolis carolinensis

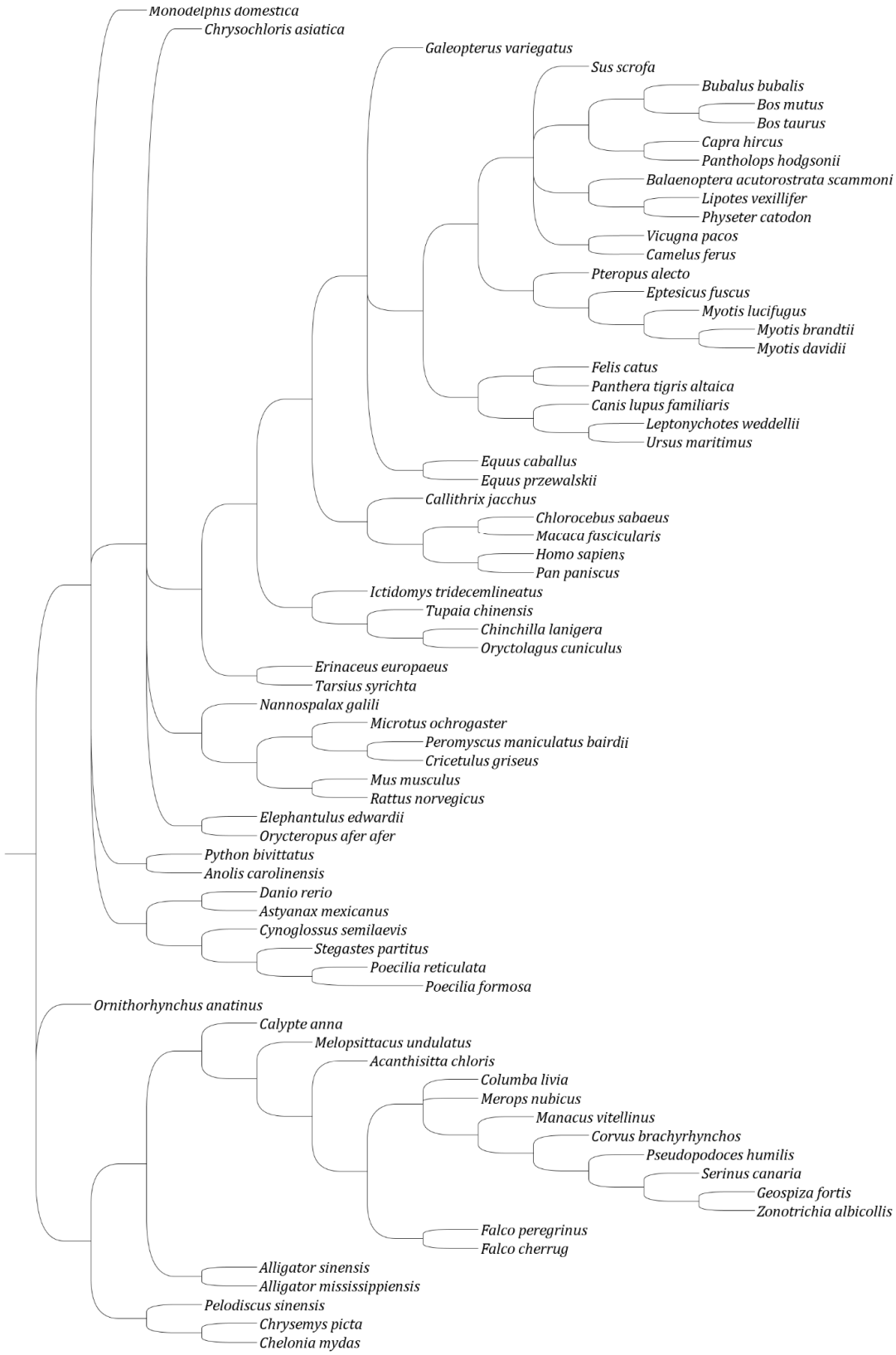

Erinaceus europaeus
Galeopterus variegatus
Tarsius syrichta
Ictidomys tridecemlineatus
Tupaia chinensis
Oryctolagus cuniculus
Sus scrofa
Bos mutus
Bos taurus
Bubalus bubalis
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Chinchilla lanigera
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Monodelphis domestica
Ornithorhynchus anatinus
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Manacus vitellinus
Corvus brachyrhynchos
Acanthisitta chloris
Melopsittacus undulatus
Calypte anna
Columba livia
Merops nubicus
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa

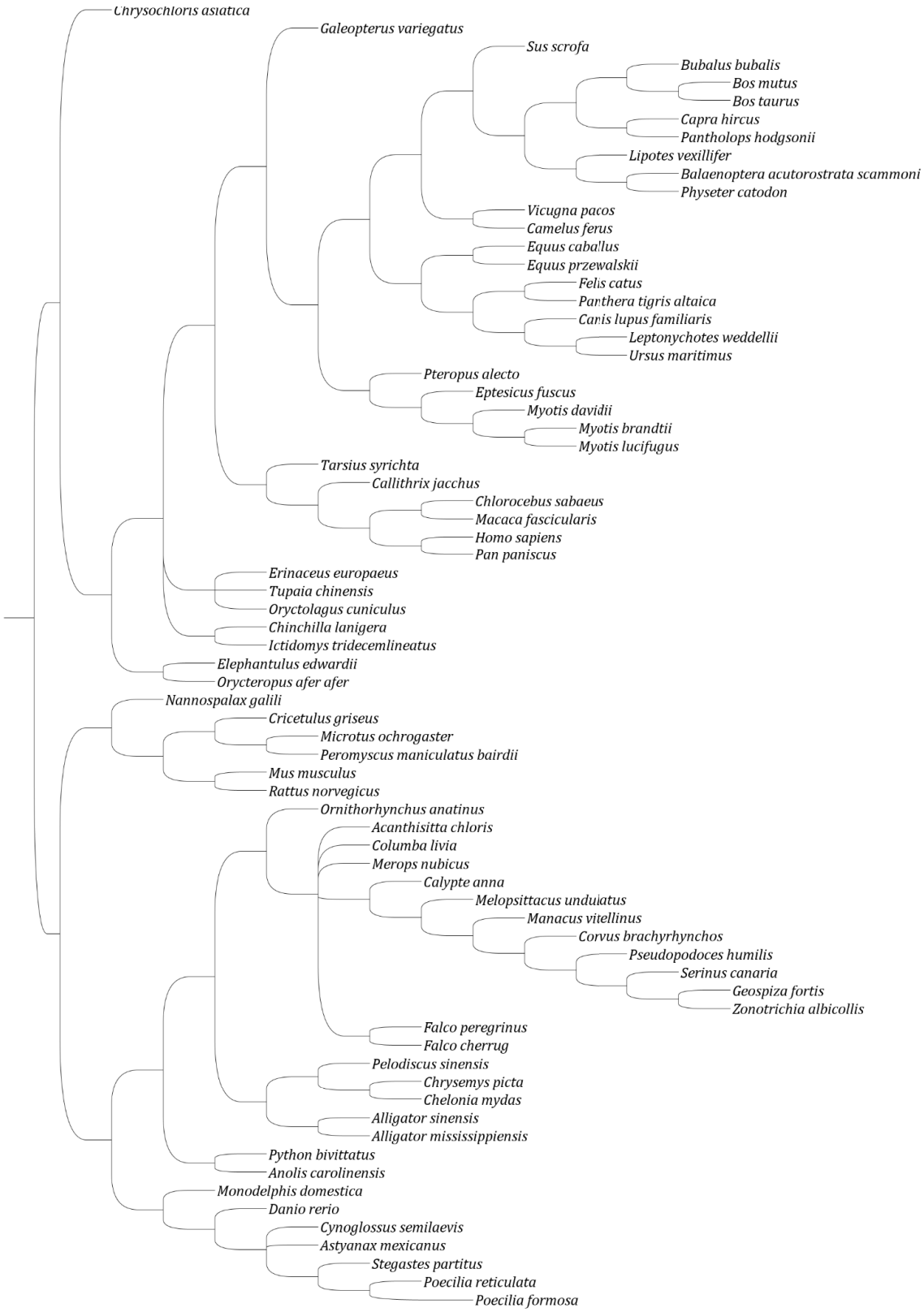

Ornithorhynchus anatinus
Monodelphis domestica
Galeopterus variegatus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Equus caballus
Equus przewalskii
Canis lupus familiaris
Ursus maritimus
Felis catus
Panthera tigris altaica
Bos mutus
Capra hircus
Pantholops hodgsonii
Bos taurus
Bubalus bubalis
Sus scrofa
Vicugna pacos
Camelus ferus
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Tupaia chinensis
Oryctolagus cuniculus
Erinaceus europaeus
Tarsius syrichta
Leptonychotes weddellii
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Microtus ochrogaster
Nannospalax galili
Mus musculus
Rattus norvegicus
Peromyscus maniculatus bairdii
Cricetulus griseus
Chinchilla lanigera
Ictidomys tridecemlineatus
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Pseudopodoces humilis
Zonotrichia albicollis
Melopsittacus undulatus
Serinus canaria
Geospiza fortis
Manacus vitellinus
Corvus brachyrhynchos
Acanthisitta chloris
Merops nubicus
Falco peregrinus
Falco cherrug
Calypte anna
Columba livia
Alligator sinensis
Alligator mississippiensis
Chrysemys picta
Pelodiscus sinensis
Chelonia mydas
Python bivittatus
Anolis carolinensis


Python bivittatus
Ornithorhynchus anatinus
Anolis carolinensis
Erinaceus europaeus
Elephantulus edwardii
Chrysochloris asiatica
Orycteropus afer afer
Galeopterus variegatus
Tupaia chinensis
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Ictidomys tridecemlineatus
Chinchilla lanigera
Oryctolagus cuniculus
Monodelphis domestica
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Calypte anna
Merops nubicus
Columba livia
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Acanthisitta chloris
Manacus vitellinus
Melopsittacus undulatus
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas


Elephantulus edwardii
Orycteropus afer afer
Erinaceus europaeus
Galeopterus variegatus
Tupaia chinensis
Oryctolagus cuniculus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Sus scrofa
Vicugna pacos
Camelus ferus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Equus caballus
Equus przewalskii
Chinchilla lanigera
Ictidomys tridecemlineatus
Chrysochloris asiatica
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Ornithorhynchus anatinus
Melopsittacus undulatus
Calypte anna
Columba livia
Merops nubicus
Acanthisitta chloris
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Monodelphis domestica
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa


Erinaceus europaeus
Tupaia chinensis
Galeopterus variegatus
Ictidomys tridecemlineatus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Oryctolagus cuniculus
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Canis lupus familiaris
Felis catus
Panthera tigris altaica
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Chrysochloris asiatica
Orycteropus afer afer
Elephantulus edwardii
Chinchilla lanigera
Monodelphis domestica
Ornithorhynchus anatinus
Calypte anna
Manacus vitellinus
Acanthisitta chloris
Merops nubicus
Geospiza fortis
Corvus brachyrhynchos
Serinus canaria
Pseudopodoces humilis
Zonotrichia albicollis
Melopsittacus undulatus
Columba livia
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Cynoglossus semilaevis
Danio rerio
Astyanax mexicanus
Stegastes partitus
Poecilia reticulata
Poecilia formosa


Monodelphis domestica
Chrysochloris asiatica
Galeopterus variegatus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Balaenoptera acutorostrata scammoni
Lipotes vexillifer
Physeter catodon
Vicugna pacos
Camelus ferus
Pteropus alecto
Eptesicus fuscus
Myotis lucifugus
Myotis brandtii
Myotis davidii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Equus caballus
Equus przewalskii
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Ictidomys tridecemlineatus
Tupaia chinensis
Chinchilla lanigera
Oryctolagus cuniculus
Erinaceus europaeus
Tarsius syrichta
Nannospalax galili
Microtus ochrogaster
Peromyscus maniculatus bairdii
Cricetulus griseus
Mus musculus
Rattus norvegicus
Elephantulus edwardii
Orycteropus afer afer
Python bivittatus
Anolis carolinensis
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Stegastes partitus
Poecilia reticulata
Poecilia formosa
Ornithorhynchus anatinus
Calypte anna
Melopsittacus undulatus
Acanthisitta chloris
Columba livia
Merops nubicus
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Alligator sinensis
Alligator mississippiensis
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas


Chrysochloris asiatica
Galeopterus variegatus
Sus scrofa
Bubalus bubalis
Bos mutus
Bos taurus
Capra hircus
Pantholops hodgsonii
Lipotes vexillifer
Balaenoptera acutorostrata scammoni
Physeter catodon
Vicugna pacos
Camelus ferus
Equus caballus
Equus przewalskii
Felis catus
Panthera tigris altaica
Canis lupus familiaris
Leptonychotes weddellii
Ursus maritimus
Pteropus alecto
Eptesicus fuscus
Myotis davidii
Myotis brandtii
Myotis lucifugus
Tarsius syrichta
Callithrix jacchus
Chlorocebus sabaeus
Macaca fascicularis
Homo sapiens
Pan paniscus
Erinaceus europaeus
Tupaia chinensis
Oryctolagus cuniculus
Chinchilla lanigera
Ictidomys tridecemlineatus
Elephantulus edwardii
Orycteropus afer afer
Nannospalax galili
Cricetulus griseus
Microtus ochrogaster
Peromyscus maniculatus bairdii
Mus musculus
Rattus norvegicus
Ornithorhynchus anatinus
Acanthisitta chloris
Columba livia
Merops nubicus
Calypte anna
Melopsittacus undulatus
Manacus vitellinus
Corvus brachyrhynchos
Pseudopodoces humilis
Serinus canaria
Geospiza fortis
Zonotrichia albicollis
Falco peregrinus
Falco cherrug
Pelodiscus sinensis
Chrysemys picta
Chelonia mydas
Alligator sinensis
Alligator mississippiensis
Python bivittatus
Anolis carolinensis
Monodelphis domestica
Danio rerio
Cynoglossus semilaevis
Astyanax mexicanus
Stegastes partitus
Poecilia reticulata
Poecilia formosa

Supplemental Table S1, Chapter 2: Species names arranged taxonomically. The percent missing data is located next to each species name. The first number is the percent of missing data from stop codons used in constructing Figure 3. The second number is the percent of missing data from all codons used in constructing Figure 2. Ex. "Microtus ochrogaster, 17.23%, 17.28%" means that the Microtus ochrogaster reference genome was missing 17.23% of the stop codon instances used in constructing the tree in Figure 3, and 17.28% of all codons used in constructing the tree in Figure 2. Each species was required to have annotations for a minimum of 10% of the genes used in this experiment.

Euteleostomi
Sarcopterygii
Amniota
Mammalia
Theria
Metatheria
Monodelphis domestica, 24.61%, 25.38%
Eutheria
Afrotheria
Tubulidentata
Orycteropus afer afer, 17.09%, 17.10%
Chrysochloridae
Chrysochloris asiatica, 18.67%, 19.82%
Macroscelidea
Elephantulus edwardii, 22.01%, 23.62%
Boreoeutheria
Euarchontoglires
Dermoptera
Galeopterus variegatus, 24.89%, 24.06%
Scandentia
Tupaia chinensis, 24.67%, 24.71%
Glires
Rodentia
Hystricognathi
Chinchilla lanigera, 16.89%, 18.08%
Sciurognathi
Sciuridae
Ictidomys tridecemlineatus, 15.96%, 16.76%
Muroidea
Spalacidae
Nannospalax galili, 18.05%, 17.33%
Muridae
Murinae
Rattus
Rattus norvegicus, 22.55%, 20.68%
Mus

Mus musculus, 12.06%, 11.84%
Cricetidae
Neotominae
Peromyscus maniculatus bairdii, 17.83%, 17.72%
Arvicolinae
Microtus ochrogaster, 17.23%, 17.28%
Cricetinae
Cricetulus griseus, 29.43%, 29.60%
Lagomorpha
Oryctolagus cuniculus, 38.23%, 38.50%
Primates
Haplorrhini
Simiiformes
Catarrhini
Hominoidea
Homininae
Homo
Homo sapiens, 5.54%, 5.46%
Pan
Pan paniscus, 18.79%, 17.97%
Cercopithecoidea
Cercopithecinae
Chlorocebus
Chlorocebus sabaeus, 12.61%, 12.61%
Macaca
Macaca fascicularis, 26.08%, 25.86%
Platyrrhini
Callithrix jacchus, 30.52%, 28.65%
Tarsiiformes
Tarsius syrichta, 25.74%, 26.45%
Laurasiatheria
Perissodactyla
Equus
Equus przewalskii, 28.27%, 28.39%
Equus caballus, 28.38%, 28.73%
Cetartiodactyla
Tylopoda
Camelidae
Camelus
Camelus ferus, 23.60%, 25.32%
Vicugna
Vicugna pacos, 22.98%, 23.25%
Ruminantia
Bovidae

Antilopinae
Pantholops hodgsonii, 23.51%, 24.38%
Caprinae
Capra hircus, 24.36%, 25.48%
Bovinae
Bubalus
Bubalus bubalis, 19.78%, 19.85%
Bos
Bos mutus, 24.06%, 24.01%
Bos taurus, 23.76%, 23.07%
Suina
Sus scrofa, 36.94%, 37.74%
Cetacea
Mysticeti
Balaenoptera acutorostrata scammoni, 21.39%, 21.72%
Odontoceti
Lipotidae
Lipotes vexillifer, 22.34%, 22.26%
Physeteridae
Physeter catodon, 25.57%, 26.40%
Carnivora
Feliformia
Felidae
Pantherinae
Panthera tigris altaica, 23.80%, 24.85%
Felinae
Felis catus, 22.64%, 21.92%
Caniformia
Ursidae
Ursus maritimus, 29.19%, 29.21%
Phocidae
Leptonychotes weddellii, 24.28%, 24.87%
Canidae
Canis lupus familiaris, 18.70%, 19.30%
Chiroptera
Microchiroptera
Vespertilionidae
Myotis
Myotis brandtii, 24.34%, 25.04%
Myotis davidii, 24.54%, 25.26%
Myotis lucifugus, 33.92%, 34.72%
Eptesicus
Eptesicus fuscus, 17.67%, 17.83%

Megachiroptera
Pteropus alecto, 19.01%, 18.94%
Insectivora
Erinaceus europaeus, 18.66%, 18.91%
Prototheria
Ornithorhynchus anatinus, 54.95%, 55.91%
Sauropsida
Sauria
Lepidosauria
Toxicofera
Serpentes
Python bivittatus, 40.85%, 42.83%
Iguania
Anolis carolinensis, 41.77%, 42.04%
Archelosauria
Archosauria
Dinosauria
Neognathae
Coraciiformes
Merops nubicus, 80.51%, 80.85%
Psittaciformes
Melopsittacus undulatus, 45.77%, 48.77%
Trochiliformes
Calypte anna, 40.17%, 42.42%
Passeriformes
Pipridae
Manacus vitellinus, 44.69%, 46.96%
Paridae
Pseudopodoces humilis, 33.77%, 35.59%
Corvoidea
Corvus brachyrhynchos, 40.77%, 42.88%
Acanthisittidae
Acanthisitta chloris, 76.21%, 76.47%
Thraupidae
Geospiza fortis, 43.99%, 46.87%
Passerellidae
Zonotrichia albicollis, 43.78%, 48.02%
Passeroidea
Serinus canaria, 39.56%, 41.92%
Falconiformes
Falco
Falco cherrug, 39.46%, 42.29%
Falco peregrinus, 39.50%, 41.68%
Columbiformes

Columba livia, 43.65%, 45.47%
Crocodylia
Alligator
Alligator sinensis, 33.01%, 34.66%
Alligator mississippiensis, 34.04%, 36.08%
Testudines
Cryptodira
Trionychia
Pelodiscus sinensis, 37.79%, 39.60%
Durocryptodira
Testudinoidea
Chrysemys picta, 24.32%, 24.33%
Americhelydia
Chelonia mydas, 35.83%, 37.08%
Actinopterygii
Clupeocephala
Euteleosteomorpha
Percomorphaceae
Carangaria
Cynoglossus semilaevis, 42.52%, 43.02%
Ovalentaria
Ovalentaria incertae sedis
Stegastes partitus, 40.62%, 41.79%
Atherinomorphae
Poecilia
Poecilia formosa, 42.47%, 43.75%
Poecilia reticulata, 38.68%, 40.24%
Otomorpha
Otophysa
Characiphysae
Astyanax mexicanus, 62.43%, 65.87%
Cypriniphysae
Danio rerio, 62.94%, 63.16%

Appendix 2: Supplementary Figures and Tables for Chapter 3

Supplementary Figures:

Supplementary Figure 1, Chapter 3: Fungi: The Open Tree of Life with annotated character state changes. This figure is too long to display on a normal page. The pdf of the tree is available upon request.

Supplementary Figure 2, Chapter 3: Invertebrates: The Open Tree of Life with annotated character state changes. This figure is too long to display on a normal page. The pdf of the tree is available upon request.

Supplementary Figure 3, Chapter 3: Plants: The Open Tree of Life with annotated character state changes. This figure is too long to display on a normal page. The pdf of the tree is available upon request.

Supplementary Figure 4, Chapter 3: Protozoa: The Open Tree of Life with annotated character state changes. This figure is too long to display on a normal page. The pdf of the tree is available upon request.

Supplementary Figure 5, Chapter 3: Mammals: The Open Tree of Life with annotated character state changes. This figure is too long to display on a normal page. The pdf of the tree is available upon request.

Supplementary Figure 6, Chapter 3: Other Vertebrates: The Open Tree of Life with annotated character state changes. This figure is too long to display on a normal page. The pdf of the tree is available upon request.

Supplementary Tables

Supplementary Table 1, Chapter 3: All species: For each parsimony informative codon character state, the name of the ortholog and codon, the number of gains, the number of losses, the number of unknown gains/losses from the root node, the number of species in the smaller group, the number of total species with that ortholog, the percent of species in the smaller group, and the total number of gains/losses divided by number of species in the smaller group. The first 100 lines out of 890,815 total lines are presented in this dissertation. The full table is available upon request.

Gene_And_Codon_Name	Num_Origin	Num_Loss	Root_Loss(0/1)	Total_Species_In_Smaller_Group	Total_Species	Percent_Species_In_Min	Total_Origin_And_Loss/Total_Species_In_Smaller_Group
ULBP1_GAC	0	1	0	2	6	0.333333333	0.5
ULBP1_GAG	0	1	0	2	6	0.333333333	0.5
CUNH1ORF64_ATT	0	2	0	5	13	0.384615385	0.4
ACEB_GAA	3	6	0	12	99	0.121212121	0.75
CUNH21ORF140_GCC	0	1	0	1	19	0.052631579	1
CUNH1ORF64_ATA	1	3	1	6	13	0.461538462	0.833333333
CUNH21ORF140_GCG	6	1	0	7	19	0.368421053	1
ULBP1_GAT	0	1	0	2	6	0.333333333	0.5
RPA1_CCA	0	1	0	1	202	0.004950495	1
FGL2_CCA	0	2	0	2	185	0.010810811	1
CNBP_TTG	7	4	0	18	186	0.096774194	0.611111111
YUNB_GGG	2	16	0	19	247	0.076923077	0.947368421
CNBP_TTA	1	0	0	1	186	0.005376344	1
CNBP_TTC	0	2	0	2	186	0.010752688	1
MTRF1L_GTA	0	9	0	11	177	0.062146893	0.818181818
OXLT_TAG	35	0	0	35	218	0.160550459	1
MTRF1L_GTC	0	4	0	4	177	0.02259887	1
MTRF1L_GTG	0	2	0	2	177	0.011299435	1
OXLT_TAA	29	13	0	79	218	0.362385321	0.53164557
CNBP_TTT	2	9	0	13	186	0.069892473	0.846153846
DR1_GGC	0	1	2	3	165	0.018181818	1
MGAT4A_TGT	0	2	0	2	188	0.010638298	1
TSTA3_GCA	0	2	0	2	174	0.011494253	1
PSAE_CCC	0	0	3	3	9	0.333333333	1
OXLT_TAT	0	5	0	5	218	0.02293578	1

PSAE_CCG	4	0	0	4	9	0.444444444	1
MTRF1L_GTT	0	7	0	7	177	0.039548023	1
SLC35B2_TTT	0	1	0	1	167	0.005988024	1
PRKCDBP_CGA	2	9	0	13	78	0.166666667	0.846153846
ROGDI_CAA	0	3	0	3	177	0.016949153	1
PLEKHB2_GGC	0	2	0	2	162	0.012345679	1
SLC35B2_TTG	0	7	0	15	167	0.089820359	0.466666667
SLC35B2_TTA	12	14	1	49	167	0.293413174	0.551020408
ROGDI_CAT	3	16	0	25	177	0.141242938	0.76
DSBE_GTA	0	1	1	2	5	0.4	1
PRKCDBP_CGT	9	7	0	35	78	0.448717949	0.457142857
GRX3_GCG	1	1	0	2	6	0.333333333	1
ACPS_CGG	21	126	0	185	837	0.221027479	0.794594595
ACPS_CGA	103	148	0	396	837	0.47311828	0.633838384
GRX3_GCC	0	1	0	1	6	0.166666667	1
ACPS_CGC	12	64	0	101	837	0.120669056	0.752475248
GRX3_GCA	0	1	0	1	6	0.166666667	1
FLAR_CTT	0	0	1	1	2	0.5	1
GTF2A2_GGA	0	12	0	17	194	0.087628866	0.705882353
PDCL_CGC	0	3	0	3	188	0.015957447	1
ACPS_CGT	8	96	0	135	837	0.161290323	0.77037037
PTPMT1_TGT	6	18	0	27	184	0.14673913	0.888888889
PDCL_CGG	2	5	0	11	188	0.058510638	0.636363636
ZNF569_CAT	0	1	0	1	49	0.020408163	1
COG5_TAC	0	1	0	1	177	0.005649718	1
MLX_TTG	8	10	1	32	188	0.170212766	0.59375
TRIM10_AGT	0	1	0	1	75	0.013333333	1
FAM173B_GAG	1	6	0	8	145	0.055172414	0.875
FAM173B_GAA	0	2	0	2	145	0.013793103	1
FAM173B_GAC	0	1	0	1	145	0.006896552	1
DSBE_GTG	0	1	0	1	5	0.2	1
FAM173B_GAT	2	3	0	5	145	0.034482759	1
GATAD2B_CGA	1	2	0	5	165	0.03030303	0.6
ZNF470_CGT	0	3	0	3	33	0.090909091	1
GCY1_CTC	0	0	2	2	4	0.5	1
BZPG_TCG	0	1	0	1	4	0.25	1
GCY1_CTA	0	1	0	2	4	0.5	0.5
RBM19_TAT	0	3	0	3	168	0.017857143	1

CYSE_CCC	83	111	0	252	1640	0.153658537	0.76984127
CYSE_CCA	258	162	0	764	1640	0.465853659	0.54973822
CYSE_CCG	3	80	0	87	1640	0.05304878	0.954022989
YEDO_CGG	0	0	1	1	3	0.333333333	1
UQCC3_TCC	0	2	0	3	6	0.5	0.666666667
ZNF101_TGT	0	1	0	1	18	0.055555556	1
UQCC3_TCA	0	1	1	2	6	0.333333333	1
HNRNPL_AGG	0	1	0	1	122	0.008196721	1
UQCC3_TCG	1	1	0	3	6	0.5	0.666666667
FAM84A_ACA	19	10	0	69	163	0.423312883	0.420289855
CYSE_CCT	125	238	0	514	1640	0.313414634	0.706225681
GP63-1_GTA	0	0	1	2	4	0.5	0.5
FAM84A_ACG	0	2	0	2	163	0.012269939	1
AMRS_GAG	0	1	0	1	339	0.002949853	1
ZNF101_TGA	4	2	1	8	18	0.444444444	0.875
USP8_CGC	0	3	0	3	183	0.016393443	1
ZNF101_TGC	1	3	0	4	18	0.222222222	1
UQCC3_TCT	2	0	0	3	6	0.5	0.666666667
LRRC70_CCC	1	4	0	7	121	0.05785124	0.714285714
SKY1_TAA	1	0	1	3	7	0.428571429	0.666666667
SKY1_TAG	1	0	0	1	7	0.142857143	1
LRRC70_CCG	1	18	0	21	121	0.173553719	0.904761905
KCNF1_AAT	9	10	0	33	183	0.180327869	0.575757576
CSM2_GAT	0	2	0	2	67	0.029850746	1
TSTA3_GCT	0	2	0	2	174	0.011494253	1
MAN2A2_ACT	0	2	0	2	152	0.013157895	1
YABP_AGC	6	21	0	41	275	0.149090909	0.658536585
YABP_AGG	56	13	0	93	275	0.338181818	0.741935484
CSM2_GAG	1	2	0	3	67	0.044776119	1
CSM2_GAC	1	6	0	8	67	0.119402985	0.875
KCNF1_AAA	5	10	0	34	183	0.18579235	0.441176471
DFNB31_CGT	0	6	0	6	95	0.063157895	1
FGL2_CCG	3	15	0	20	185	0.108108108	0.9
ZNF101_GTT	0	2	1	3	18	0.166666667	1
SRSF2_GCT	11	13	2	50	127	0.393700787	0.52
MLX_TTT	2	5	0	9	188	0.04787234	0.777777778
MRAS_CGT	12	17	0	73	191	0.382198953	0.397260274

Supplementary Table 2, Chapter 3: Archaea: The first 100 lines out of 18,552 total lines are presented in this dissertation. The full table is available upon request.

Gene_And_Codon_Name	Num_Origin	Num_Loss	Root_Loss(0/1)	Total_Species_In_Smaller_Group	Total_Species	Percent_Species_In_Min	Total_Origin_And_Loss/Total_Species_In_Smaller_Group
PYRC_CTG	0	1	0	1	11	0.090909091	1
PYK_GCT	3	7	0	13	88	0.147727273	0.769230769
MENC_TCA	4	0	0	4	10	0.4	1
PYRC_CTC	0	2	0	2	11	0.181818182	1
PFDA_TTC	5	7	0	30	73	0.410958904	0.4
FEN_ACG	0	0	1	1	26	0.038461538	1
GVPN_TAA	5	1	1	8	28	0.285714286	0.875
ACS_CGT	0	8	0	9	78	0.115384615	0.888888889
BTUD_GTA	0	3	0	3	15	0.2	1
LEUS_GAT	0	2	0	2	50	0.04	1
PYK_GCA	0	6	0	6	88	0.068181818	1
COBN_TCT	0	11	0	13	41	0.317073171	0.846153846
ARCC_CGC	0	4	0	4	16	0.25	1
ACS_CGC	0	1	0	2	78	0.025641026	0.5
BTUD_GTT	0	3	0	3	15	0.2	1
ACS_CGA	0	9	0	10	78	0.128205128	0.9
CAS3_CCT	0	2	0	2	17	0.117647059	1
ECTC_CCC	0	1	0	1	3	0.333333333	1
NTHB_CAA	1	0	0	1	5	0.2	1
HEML_CGT	0	3	0	4	15	0.266666667	0.75
HMGA_CAC	0	2	0	2	88	0.022727273	1
ARTF_TGC	0	0	2	2	4	0.5	1
MSRB_ATT	13	4	1	19	75	0.253333333	0.947368421
SURE_TGG	0	1	0	1	62	0.016129032	1
PHNG_GTT	0	1	0	1	7	0.142857143	1
SFSA_TAA	3	0	0	3	9	0.333333333	1
RPOA2_ATA	12	6	0	43	95	0.452631579	0.418604651
ARTF_TGT	0	0	2	2	4	0.5	1
HEML_CGG	2	2	0	6	15	0.4	0.666666667
HEML_CGA	2	1	0	7	15	0.466666667	0.428571429
HEML_CGC	2	0	0	6	15	0.4	0.333333333
CSM4_TGA	0	1	0	1	4	0.25	1
ARGF_TGA	16	10	2	42	93	0.451612903	0.666666667
RPIA_AAT	0	1	0	2	10	0.2	0.5
CAS2_CAC	5	4	0	12	45	0.266666667	0.75
GVPN_AGA	0	5	0	7	28	0.25	0.714285714
CAS2_CAA	7	8	2	20	45	0.444444444	0.85

GYRB_CTT	4	2	1	9	71	0.126760563	0.777777778
CAS2_CAG	0	2	0	4	45	0.088888889	0.5
TFX_GCA	0	0	1	1	13	0.076923077	1
CADA_TTG	0	3	0	3	51	0.058823529	1
LEUS_ATT	0	2	0	2	50	0.04	1
TRPD_AAT	16	8	0	41	85	0.482352941	0.585365854
GRXC_ACA	0	2	0	2	7	0.285714286	1
SUFB_GCG	0	1	0	1	69	0.014492754	1
TFX_GCT	0	3	0	3	13	0.230769231	1
CAS2_CAT	12	1	0	17	45	0.377777778	0.764705882
GRXC_ACT	0	0	1	1	7	0.142857143	1
TRPD_AAA	2	5	0	13	85	0.152941176	0.538461538
LEUS_ATA	5	1	0	13	50	0.26	0.461538462
CAS4A_TAA	1	0	0	1	5	0.2	1
RPL7AE_ATC	0	2	0	2	70	0.028571429	1
AHAC_CGG	0	1	0	1	38	0.026315789	1
DPH2_GCC	0	1	0	1	94	0.010638298	1
AHAC_CGA	1	3	0	5	38	0.131578947	0.8
GVPN_AGG	1	4	0	6	28	0.214285714	0.833333333
AHAC_CGC	1	1	0	3	38	0.078947368	0.666666667
GCVH_ATA	6	1	1	23	82	0.280487805	0.347826087
THIE_CGC	2	2	1	7	65	0.107692308	0.714285714
GCVH_ATC	0	2	0	2	82	0.024390244	1
KYNU_CCA	11	1	0	14	28	0.5	0.857142857
CMR5_AGG	0	1	0	1	5	0.2	1
AHAC_CGT	1	3	0	5	38	0.131578947	0.8
CMR5_AGC	0	1	0	1	5	0.2	1
MENC_TCT	3	0	0	3	10	0.3	1
CMR5_AGA	0	0	1	1	5	0.2	1
KYNU_CCT	6	1	0	9	28	0.321428571	0.777777778
DPH2_GCG	0	3	0	6	94	0.063829787	0.5
GCVH_ATT	13	3	1	30	82	0.365853659	0.566666667
LEUD_AGT	14	1	0	18	73	0.246575342	0.833333333
PCM_CCT	0	1	0	2	7	0.285714286	0.5
GLTA_CGC	0	1	0	1	20	0.05	1
TRPD_TTG	18	7	0	42	85	0.494117647	0.595238095
GLTA_CGA	3	0	0	3	20	0.15	1
SPED_AAC	0	1	0	1	4	0.25	1
GLTA_CGT	0	4	0	4	20	0.2	1
MOAC_GCA	0	7	0	8	29	0.275862069	0.875
PCM_CCA	0	2	0	3	7	0.428571429	0.666666667
SAT_GAA	0	1	0	1	14	0.071428571	1

TMK_CAA	0	2	0	2	8	0.25	1
LEUC_GCA	3	2	0	9	67	0.134328358	0.555555556
NRDD_TAG	1	0	0	1	21	0.047619048	1
MOAC_GCG	0	0	1	1	29	0.034482759	1
CYSE_CCC	1	0	1	2	51	0.039215686	1
CYSE_CCA	8	4	0	22	51	0.431372549	0.545454545
CYSE_CCG	0	1	0	1	51	0.019607843	1
LEUC_GTT	1	12	0	23	67	0.343283582	0.565217391
CCA_TTA	0	1	0	1	5	0.2	1
CDHB_TTG	0	1	0	2	9	0.222222222	0.5
RDGB_TGA	9	2	0	15	58	0.25862069	0.733333333
RDGB_TGC	7	2	0	10	58	0.172413793	0.9
CDHB_TTC	0	1	0	1	9	0.111111111	1
RPS19P_CAA	0	1	0	3	10	0.3	0.333333333
CDHB_TTA	0	2	1	3	9	0.333333333	1
RDGB_TGG	0	1	0	1	58	0.017241379	1
PURK_CCA	0	1	1	2	4	0.5	1
CYSE_CCT	12	3	0	22	51	0.431372549	0.681818182
FRHA_CTG	0	0	1	3	8	0.375	0.333333333
RPIA_AAA	0	1	0	1	10	0.1	1
RDGB_TGT	5	1	0	11	58	0.189655172	0.545454545

Supplementary Table 3, Chapter 3: Bacteria: The first 100 lines out of 183,541 total lines are presented in this dissertation. The full table is available upon request.

Gene_And_Codon_Name	Num_Origin	Num_Loss	Root_Loss(0/1)	Total_Species_In_Smaller_Group	Total_Species	Percent_Species_In_Min	Total_Origin_And_Loss/Total_Species_In_Smaller_Group
FIMD_GGA	0	1	0	2	5	0.4	0.5
BPHC_TGT	12	1	0	17	83	0.204819277	0.764705882
FUCA_ACT	13	2	0	30	73	0.410958904	0.5
ICAD_AGC	0	2	0	2	9	0.222222222	1
OXLT_TAG	34	0	0	35	218	0.160550459	0.971428571
PSAE_CCT	0	0	1	1	2	0.5	1
CLOSI_CGA	1	0	0	1	5	0.2	1
BPHC_TGA	10	5	2	25	83	0.301204819	0.68
BPHC_TGC	0	15	0	15	83	0.180722892	1
PSAE_CCC	0	0	1	1	2	0.5	1
GNTT_CGA	1	0	0	1	4	0.25	1
PSAE_CCG	0	0	1	1	2	0.5	1
HPRK_ACG	0	9	0	18	94	0.191489362	0.5
RPSN_AAT	225	56	0	392	3236	0.121137206	0.716836735
HPRK_ACA	6	10	0	31	94	0.329787234	0.516129032
HPRK_ACC	0	7	0	7	94	0.074468085	1
PTCC_TGT	0	0	1	1	3	0.333333333	1
RPSN_AAG	12	67	0	91	3236	0.028121137	0.868131868
HPRK_ACT	8	7	0	32	94	0.340425532	0.46875
RPSN_AAC	196	148	2	1286	3236	0.397404203	0.269051322
RPSN_AAA	134	76	0	1026	3236	0.317058096	0.204678363
PURR_GGC	0	1	0	1	28	0.035714286	1
ACPS_CGG	21	125	0	184	836	0.220095694	0.793478261
ACPS_CGA	104	146	0	397	836	0.474880383	0.629722922
ACPS_CGC	12	62	1	100	836	0.119617225	0.75
FLAR_CTT	0	0	1	1	2	0.5	1
GPT_GCG	0	1	0	1	11	0.090909091	1
ACPS_CGT	10	94	0	135	836	0.161483254	0.77037037
FRCK_GTC	0	1	0	1	7	0.142857143	1
DHAK_CTC	2	6	0	8	191	0.041884817	1
GLTA_CGG	115	104	0	498	1227	0.405867971	0.439759036
DHAK_CTA	12	7	0	54	191	0.282722513	0.351851852
DHAK_CTG	0	7	0	7	191	0.036649215	1
GLTA_CGC	14	19	0	47	1227	0.038304808	0.70212766
GLTA_CGA	137	78	0	353	1227	0.287693562	0.609065156
GLTA_CGT	3	61	0	69	1227	0.056234719	0.927536232
FDXH_GGG	7	23	0	41	432	0.094907407	0.731707317

RPH_ACT	22	18	0	50	196	0.255102041	0.8
FDXH_GGA	56	67	1	216	432	0.5	0.574074074
DHAK_CTT	19	5	0	85	191	0.445026178	0.282352941
PSEI_TAA	73	49	0	177	450	0.393333333	0.689265537
CYSE_CCC	82	110	1	250	1589	0.157331655	0.772
YBGF_CTT	0	1	0	1	8	0.125	1
CYSE_CCA	250	156	0	735	1589	0.462555066	0.552380952
CYSE_CCG	3	79	0	86	1589	0.054122089	0.953488372
CYSE_CCT	112	235	0	485	1589	0.305223411	0.715463918
BASR_TCT	0	1	1	2	5	0.4	1
YABP_AGT	1	52	0	57	275	0.207272727	0.929824561
CSM2_GAT	0	2	0	2	62	0.032258065	1
BPHC_AAA	11	12	1	39	83	0.469879518	0.615384615
YABP_AGA	14	43	1	124	275	0.450909091	0.467741935
YABP_AGC	6	21	0	41	275	0.149090909	0.658536585
BASR_TCA	0	0	1	2	5	0.4	0.5
BASR_TCC	0	1	0	1	5	0.2	1
YABP_AGG	56	13	0	93	275	0.338181818	0.741935484
CSM2_GAG	1	1	0	2	62	0.032258065	1
BPHC_AAT	15	0	1	32	83	0.385542169	0.5
CSM2_GAC	1	6	0	8	62	0.129032258	0.875
CSM2_GAA	0	1	0	1	62	0.016129032	1
YQIA_GAC	0	0	1	1	5	0.2	1
HEMG_CAT	34	71	0	158	803	0.196762142	0.664556962
CSM2_TCC	10	7	0	27	62	0.435483871	0.62962963
CSM2_TCA	4	14	0	23	62	0.370967742	0.782608696
CSM2_TCG	14	1	0	19	62	0.306451613	0.789473684
PAAD_CAA	52	86	0	322	723	0.445366528	0.428571429
PAAD_CAT	49	59	0	263	723	0.363762102	0.410646388
GALR_TTA	0	1	0	1	9	0.111111111	1
CSM2_TCT	3	13	0	24	62	0.387096774	0.666666667
DRRA_CTA	0	1	0	2	4	0.5	0.5
CSPA2_TCC	0	0	1	1	2	0.5	1
CSPA2_TCA	0	0	1	1	2	0.5	1
NTRC_GTT	27	121	0	168	814	0.206388206	0.880952381
PURF_AAT	0	7	0	8	98	0.081632653	0.875
LENA_CTA	0	1	0	1	7	0.142857143	1
LENA_CTC	0	3	0	3	7	0.428571429	1
LENA_CTG	0	3	0	3	7	0.428571429	1
PURF_AAC	0	1	0	1	98	0.010204082	1
NTRC_GTA	78	98	1	238	814	0.292383292	0.743697479
PURF_AAA	1	7	0	8	98	0.081632653	1
NTRC_GTC	0	4	0	4	814	0.004914005	1

PURF_AAG	0	1	0	1	98	0.010204082	1
NTRC_GTG	0	1	0	1	814	0.001228501	1
COBM_TAG	118	29	0	157	1380	0.113768116	0.936305732
CRL_CGC	1	0	0	1	4	0.25	1
PHRA_GGA	0	0	1	1	4	0.25	1
COBM_TAC	2	34	0	39	1380	0.02826087	0.923076923
PHRA_GGC	0	0	1	1	4	0.25	1
CRL_CGG	0	0	1	1	4	0.25	1
SOXD_TCT	1	0	0	1	6	0.166666667	1
NTRC_TTC	0	5	0	5	814	0.006142506	1
NTRC_TTA	87	21	0	172	814	0.211302211	0.627906977
NTRC_TTG	0	34	0	35	814	0.042997543	0.971428571
SOXD_TCA	0	0	1	1	6	0.166666667	1
SOXD_TCG	0	1	0	2	6	0.333333333	0.5
OMPF_CTA	1	0	0	1	4	0.25	1
SOXD_AAA	0	0	1	1	6	0.166666667	1
OMPF_CTG	0	1	0	1	4	0.25	1
NTRC_TTT	51	104	0	162	814	0.199017199	0.956790123
NOSZ_GGG	2	4	0	9	58	0.155172414	0.666666667
NOSZ_GGA	2	9	0	12	58	0.206896552	0.916666667

Supplementary Table 4, Chapter 3: Fungi: The first 100 lines out of 27,054 total lines are presented in this dissertation. The full table is available upon request.

Gene_And_Codon_Name	Num_Origin	Num_Loss	Root_Loss(0/1)	Total_Species_In_Smaller_Group	Total_Species	Percent_Species_In_Min	Total_Origin_And_Loss/Total_Species_In_Smaller_Group
IES6_ATA	1	0	1	2	4	0.5	1
EAP1_CCG	0	1	0	2	4	0.5	0.5
EAP1_CCC	0	1	0	1	4	0.25	1
TBF1_TGC	0	1	1	2	4	0.5	1
SNF3_CGG	0	2	0	2	5	0.4	1
TWF1_ATC	0	0	1	1	4	0.25	1
SNF3_CGC	1	0	0	1	5	0.2	1
RIP1_GCA	0	1	0	1	7	0.142857143	1
RIP1_GCG	0	1	1	2	7	0.285714286	1
VPS27_CCG	0	1	0	1	4	0.25	1
ALG11_TGA	0	1	2	3	7	0.428571429	1
ALG11_TGC	0	2	0	2	7	0.285714286	1
TIM23_GAC	0	0	1	1	6	0.166666667	1
TIM23_GAG	0	1	0	1	6	0.166666667	1
HIS7_AGG	0	2	0	2	6	0.333333333	1
HIS7_AGC	0	1	0	2	6	0.333333333	0.5
RGA1_TGA	1	0	0	1	4	0.25	1
CDC4_TGA	2	0	0	2	5	0.4	1
CDC4_TGC	0	0	1	1	5	0.2	1
TOP2_GGC	0	1	0	1	7	0.142857143	1
CDC4_TGG	0	0	1	1	5	0.2	1
LEU2_CAG	0	1	0	1	6	0.166666667	1
GCY1_CTT	0	1	0	1	4	0.25	1
MCM1_GTT	0	0	1	1	5	0.2	1
STE11_CGA	0	1	0	1	5	0.2	1
HMX1_AGC	1	0	1	2	4	0.5	1
ORC4_TAA	0	1	0	1	4	0.25	1
QCR7_ACG	2	0	0	2	7	0.285714286	1
HMX1_AGG	0	0	1	1	4	0.25	1
GCY1_CTG	1	0	0	1	4	0.25	1
MCM1_GTC	0	1	0	1	5	0.2	1
GCY1_CTC	0	0	2	2	4	0.5	1
RHB1_CAC	1	0	1	2	5	0.4	1
PSF3_GTG	0	1	1	2	4	0.5	1
PSF3_GTA	0	1	0	1	4	0.25	1
PSF3_GTC	0	0	1	1	4	0.25	1
SKY1_TAA	1	0	0	1	4	0.25	1

SKY1_TAG	0	0	1	1	4	0.25	1
HYR1_AGT	0	0	1	1	4	0.25	1
GCV1_TCG	0	1	0	1	5	0.2	1
OST1_TAG	1	0	0	1	7	0.142857143	1
HYR1_AGG	1	0	0	1	4	0.25	1
HEM15_CGC	1	0	1	2	5	0.4	1
HEM15_CGA	0	2	0	2	5	0.4	1
HEM15_CGG	0	1	1	2	5	0.4	1
OST1_GGG	0	2	0	2	7	0.285714286	1
MMS2_TCA	0	1	0	1	4	0.25	1
IAH1_TGT	0	1	0	1	4	0.25	1
RPL10_AAT	0	0	2	2	4	0.5	1
MMS2_TCG	1	0	0	1	4	0.25	1
SUT1_GGG	0	2	0	2	6	0.333333333	1
IAH1_TGC	0	1	0	1	4	0.25	1
IAH1_TGA	1	0	0	1	4	0.25	1
SDA1_GCG	0	1	0	1	5	0.2	1
TUB2_GCG	0	1	0	2	4	0.5	0.5
TRR1_TTT	0	1	0	1	5	0.2	1
PRE10_AAT	0	1	0	1	4	0.25	1
ULP1_TGC	0	1	1	2	4	0.5	1
RMT2_CGG	0	1	0	2	4	0.5	0.5
HXT2_TAG	0	0	1	1	3	0.333333333	1
TYR1_GGG	0	1	0	1	5	0.2	1
PRE10_AAC	0	1	0	1	4	0.25	1
MRPL25_CGC	1	0	1	2	4	0.5	1
PRE10_TGC	1	0	1	2	4	0.5	1
MRPL25_CGG	0	0	1	1	4	0.25	1
DIT2_CCG	0	1	0	1	4	0.25	1
SPE2_GAC	0	1	0	1	5	0.2	1
PRE10_TGT	0	1	0	1	4	0.25	1
DPM1_GCA	0	1	0	1	6	0.166666667	1
PET9_CTG	1	0	0	1	4	0.25	1
PET9_CTC	0	0	2	2	4	0.5	1
PET9_CTA	1	0	0	1	4	0.25	1
ATS1_TCC	0	0	1	1	4	0.25	1
PET9_CTT	0	1	0	1	4	0.25	1
ATS1_TCG	0	1	1	2	4	0.5	1
RAD14_CAG	0	1	0	1	4	0.25	1
RAD14_CAC	0	1	0	1	4	0.25	1
NCB2_TCC	0	1	0	1	4	0.25	1

PIM1_CGG	0	0	1	1	4	0.25	1
LYS9_ATA	1	0	0	1	5	0.2	1
RRP46_TGC	0	0	1	1	4	0.25	1
CHS7_CAC	0	1	0	1	6	0.166666667	1
GPD1_AGT	0	1	0	1	7	0.142857143	1
VAS1_ACA	0	1	0	1	5	0.2	1
UBC1_TGG	0	1	0	1	4	0.25	1
UBC1_TGA	1	0	1	2	4	0.5	1
UBC1_TGC	0	1	0	1	4	0.25	1
TRM1_CTC	0	1	0	1	5	0.2	1
NIK1_TGA	0	0	1	1	4	0.25	1
TRM1_CTG	0	1	0	1	5	0.2	1
RPL3_TTT	0	1	0	1	4	0.25	1
UBC1_TGT	0	1	1	2	4	0.5	1
GPD1_AGG	1	0	0	1	7	0.142857143	1
GPD1_AGA	0	1	0	1	7	0.142857143	1
GPD1_AGC	0	1	0	3	7	0.428571429	0.333333333
RPL3_TTA	0	0	2	2	4	0.5	1
XYL2_AGT	0	1	0	1	5	0.2	1
RIM101_CGT	0	1	0	1	5	0.2	1
XYL2_AGG	1	0	0	1	5	0.2	1
IDH1_ACA	0	1	0	1	6	0.166666667	1

Supplementary Table 5, Chapter 3: Invertebrates: The first 100 lines out of 1,692 total lines are presented in this dissertation. The full table is available upon request.

Gene_And_Codon_Name	Num_Origin	Num_Loss	Root_Loss(0/1)	Total_Species_In_Smaller_Group	Total_Species	Percent_Species_In_Min	Total_Origin_And_Loss/Total_Species_In_Smaller_Group
GPB5_TAA	0	1	0	3	6	0.5	0.333333333
GPB5_TAG	1	0	0	1	6	0.166666667	1
PDF_TCA	0	1	0	1	4	0.25	1
GPB5_TAT	0	1	0	1	6	0.166666667	1
PDF_TCT	0	1	0	1	4	0.25	1
CPR5_ATA	0	0	1	2	4	0.5	0.5
CAPA_TAT	0	1	0	1	4	0.25	1
UQCR11_GAC	0	1	1	2	5	0.4	1
UQCR11_GAG	0	2	0	2	5	0.4	1
CAPA_TAA	0	1	0	1	4	0.25	1
CAPA_TAG	1	0	0	1	4	0.25	1
UQCR11_GAT	1	0	1	2	5	0.4	1
RPS8_AGG	0	1	1	2	4	0.5	1
IMD_CTC	0	0	1	2	4	0.5	0.5
IMD_CTA	0	1	0	1	4	0.25	1
ND5_AGC	4	2	0	16	33	0.484848485	0.375
CPR16_CAT	0	1	0	1	4	0.25	1
ND5_AGG	3	4	0	16	33	0.484848485	0.4375
RPL39_CGG	1	0	0	1	5	0.2	1
RPL39_CGA	0	1	1	2	5	0.4	1
RPL39_CGC	0	1	1	2	5	0.4	1
CPR16_CAA	0	1	0	1	4	0.25	1
CPR16_CAC	0	1	0	1	4	0.25	1
SXL_GTT	0	1	0	1	4	0.25	1
ND5_AGT	0	1	0	1	33	0.03030303	1
COX3_GAG	4	3	1	17	42	0.404761905	0.470588235
CDK5_TCA	0	1	0	1	5	0.2	1
CDK5_TCC	0	1	0	1	5	0.2	1
ECR_CGC	0	0	1	1	4	0.25	1
SYT4_TGA	0	1	0	1	4	0.25	1
ECR_TAT	0	1	0	1	4	0.25	1
COX2_TTG	2	1	1	12	29	0.413793103	0.333333333
COX2_TTC	0	5	0	5	29	0.172413793	1
OBP6_CAT	0	1	0	1	4	0.25	1
IMD_TAT	0	1	0	1	4	0.25	1
CPR14_TGT	1	0	0	1	4	0.25	1
EVE_TAT	0	1	0	1	4	0.25	1

IMD_TAA	0	0	1	2	4	0.5	0.5
IMD_TAG	0	0	1	2	4	0.5	0.5
OBP6_CAC	0	1	1	2	4	0.5	1
CPR14_TGG	0	0	2	2	4	0.5	1
SXL_GTA	0	1	0	1	4	0.25	1
GSTD3_TCA	1	0	0	1	4	0.25	1
GSTD3_TCC	0	1	0	1	4	0.25	1
COX3_GGT	0	1	0	1	42	0.023809524	1
GSTD3_TCT	1	0	0	1	4	0.25	1
COX3_GGC	4	6	1	17	42	0.404761905	0.647058824
ATP6_AAA	0	1	0	1	45	0.022222222	1
COX3_GGG	3	7	1	16	42	0.380952381	0.6875
RPL32_CGC	0	1	0	1	4	0.25	1
GSTZ1_GGG	1	0	0	1	4	0.25	1
ND2_GAT	1	5	0	10	37	0.27027027	0.6
ND2_GAC	3	2	0	8	37	0.216216216	0.625
ND2_GAA	0	2	0	2	37	0.054054054	1
ND2_GAG	0	2	0	12	37	0.324324324	0.166666667
RAB7_GTG	0	1	0	1	4	0.25	1
COX3_GAC	4	6	0	17	42	0.404761905	0.588235294
GSTD1_TCT	1	0	0	1	4	0.25	1
GSTZ1_CGG	1	0	0	1	4	0.25	1
GSTD1_TCC	0	1	0	1	4	0.25	1
RPS12_CAG	1	0	0	2	4	0.5	0.5
RPS12_CAA	0	1	0	1	4	0.25	1
RPS12_CAC	0	0	1	1	4	0.25	1
PTEN_TGA	0	1	0	2	4	0.5	0.5
GSTT1_CTG	0	0	1	1	4	0.25	1
CPR16_ACT	0	1	0	1	4	0.25	1
GPB5_GCG	0	1	2	3	6	0.5	1
GPB5_GCA	0	1	0	1	6	0.166666667	1
CPR16_ACA	0	0	2	2	4	0.5	1
CPR16_ACC	0	1	0	1	4	0.25	1
GPB5_GCT	0	1	0	1	6	0.166666667	1
ETH_TTA	0	2	1	3	6	0.5	1
COX1_GAC	0	3	0	3	27	0.111111111	1
ND4_TTC	1	4	0	7	32	0.21875	0.714285714
COX1_GAG	1	2	1	12	27	0.444444444	0.333333333
ND4_CTG	3	3	0	12	32	0.375	0.5
ND4_CTC	5	1	0	9	32	0.28125	0.666666667
ND4_CTA	5	3	0	9	32	0.28125	0.888888889
RPS12_ACC	0	0	1	1	4	0.25	1
AKHR_GAA	0	1	0	1	4	0.25	1

COX2_CGT	1	7	0	9	29	0.310344828	0.888888889
ETH_CAT	0	2	0	3	6	0.5	0.666666667
COX2_CGG	7	1	0	10	29	0.344827586	0.8
ETH_CAA	0	1	0	1	6	0.166666667	1
COX2_CGC	8	0	0	8	29	0.275862069	1
COX2_CGA	2	2	0	7	29	0.24137931	0.571428571
CPR16_TTT	0	1	0	1	4	0.25	1
OBP6_GTA	0	1	0	1	4	0.25	1
OBP6_GTG	0	0	1	1	4	0.25	1
CPR16_TTA	0	0	1	2	4	0.5	0.5
CPR4_CGC	0	1	0	1	4	0.25	1
CPR16_TTG	0	0	2	2	4	0.5	1
OBP6_TTA	0	1	0	1	4	0.25	1
ND3_CTT	0	11	0	15	41	0.365853659	0.733333333
OBP6_TTG	0	1	0	1	4	0.25	1
EYG_GAT	0	1	0	1	4	0.25	1
EVE_CTA	0	1	0	2	4	0.5	0.5
ND3_CTA	2	4	0	8	41	0.195121951	0.75
ND3_CTC	7	1	0	9	41	0.219512195	0.888888889
ND3_CTG	2	4	0	8	41	0.195121951	0.75

Supplementary Table 6, Chapter 3: Plants: The first 100 lines out of 12,503 total lines are presented in this dissertation. The full table is available upon request.

Gene_And_Codon_Name	Num_Origin	Num_Loss	Root_Loss(0/1)	Total_Species_In_Smaller_Group	Total_Species	Percent_Species_In_Min	Total_Origin_And_Loss/Total_Species_In_Smaller_Group
ISU1_AAT	1	0	0	2	4	0.5	0.5
RER1_CAT	0	0	1	1	4	0.25	1
NAD5_TTA	0	1	0	1	4	0.25	1
PSAL_ACT	0	1	0	2	7	0.285714286	0.5
PSAC_AAG	0	2	0	2	52	0.038461538	1
PSAC_AAA	0	2	0	3	52	0.057692308	0.666666667
ARF2_AGT	0	0	1	1	5	0.2	1
ABP1_TAG	0	1	0	1	4	0.25	1
ISU1_AAA	0	0	1	1	4	0.25	1
ABP1_TAA	1	0	0	1	4	0.25	1
ARF2_AGA	0	0	1	1	5	0.2	1
PSAC_AAT	1	0	1	2	52	0.038461538	1
PSAL_ACG	0	0	1	1	7	0.142857143	1
MSD1_TAA	1	0	0	1	4	0.25	1
PSAL_ACA	0	1	0	2	7	0.285714286	0.5
PSAL_ACC	0	0	1	1	7	0.142857143	1
RPL10_AAT	0	0	1	2	9	0.222222222	0.5
EIF2A_AGA	0	1	0	1	4	0.25	1
PETG_CGC	2	2	0	5	51	0.098039216	0.8
EIF2A_AGG	0	1	0	1	4	0.25	1
PSAE_CCT	0	2	0	2	6	0.333333333	1
RPL10_AAC	0	1	0	1	9	0.111111111	1
ASN2_ATA	0	0	1	2	4	0.5	0.5
RPL10_AAA	0	1	0	1	9	0.111111111	1
EXPB1_CTA	0	0	1	1	4	0.25	1
PSAE_CCC	0	0	1	1	6	0.166666667	1
SPL2_TAG	0	1	0	1	4	0.25	1
PSAE_CCG	0	2	1	3	6	0.5	1
EIF2A_AGT	0	1	0	1	4	0.25	1
PSBM_CCC	0	0	2	3	51	0.058823529	0.666666667
COG4_CCC	0	1	0	1	4	0.25	1
PSBM_CCA	1	0	0	1	51	0.019607843	1
PSBM_CCG	2	0	0	4	51	0.078431373	0.5
COG4_CCG	0	1	0	1	4	0.25	1
UBC3_TCG	0	1	0	2	4	0.5	0.5
RPL13_TCC	0	1	1	2	5	0.4	1

RPL13_TCA	1	0	0	1	5	0.2	1
RPL13_TCG	1	0	0	2	5	0.4	0.5
ORF103C_GGC	1	0	1	2	4	0.5	1
SPL2_CAT	0	1	0	1	4	0.25	1
PSBM_CCT	1	1	0	6	51	0.117647059	0.333333333
RPS9_GTC	0	1	1	2	5	0.4	1
ORF103C_GGA	0	0	1	1	4	0.25	1
RPL13_TCT	0	1	0	1	5	0.2	1
CEMA_TCC	0	2	0	2	51	0.039215686	1
GPX4_CCT	0	1	0	1	4	0.25	1
CEMA_TCG	0	2	0	4	51	0.078431373	0.5
RPL35_GCT	0	1	0	1	6	0.166666667	1
PSAH_CGT	1	0	0	1	5	0.2	1
ORF103C_GGG	0	1	0	1	4	0.25	1
RPL35_GCA	2	0	0	2	6	0.333333333	1
MYB1_TAC	0	1	0	1	4	0.25	1
RPL35_GCC	0	0	1	1	6	0.166666667	1
PSAH_CGA	1	0	0	1	5	0.2	1
PSAH_CGG	2	0	0	2	5	0.4	1
RPL35_GCG	0	0	1	1	6	0.166666667	1
RPS4_AAG	0	7	0	10	57	0.175438596	0.7
ORF101B_GAC	0	0	2	2	5	0.4	1
ORF101B_GAA	0	1	0	1	5	0.2	1
ORF101B_GAG	0	0	2	2	5	0.4	1
WRKY36_TAA	1	0	0	2	4	0.5	0.5
BZIP1_TAT	0	1	1	2	4	0.5	1
EXPA5_CAT	0	0	1	1	4	0.25	1
YCF15_ACG	0	1	0	1	7	0.142857143	1
TIM_TGT	0	1	0	1	5	0.2	1
ORF101B_GAT	0	1	0	2	5	0.4	0.5
XTH1_GAT	0	0	1	1	4	0.25	1
PSAD_TAA	0	1	0	1	5	0.2	1
PSAD_TAG	1	0	0	1	5	0.2	1
RPL30_ACA	0	0	1	1	4	0.25	1
ORF101C_GGT	0	1	0	1	4	0.25	1
RPL30_ACT	0	1	0	1	4	0.25	1
XTH1_GAA	0	0	1	1	4	0.25	1
PSAD_TAT	0	0	1	1	5	0.2	1
BZIP1_GCC	0	1	0	1	4	0.25	1
CLPP_AGC	0	8	0	12	47	0.255319149	0.666666667
CLPP_AGG	0	1	1	2	47	0.042553191	1
PSBY_TGC	1	0	0	1	5	0.2	1
BBS9_TTA	0	0	1	2	4	0.5	0.5

SIP1_ATT	0	1	0	1	4	0.25	1
MRPL11_CCC	0	0	1	1	4	0.25	1
ORF101A_TTT	0	1	0	1	5	0.2	1
MRPL11_CCA	0	1	0	1	4	0.25	1
SDH4_CTC	1	0	0	2	10	0.2	0.5
BBS9_TTT	0	1	0	1	4	0.25	1
RER1_ACA	0	1	0	2	4	0.5	0.5
ORF101A_TTA	1	0	0	1	5	0.2	1
PSBA_GGG	10	2	1	26	52	0.5	0.5
MRPL11_CCT	1	0	0	2	4	0.5	0.5
PSBA_GGC	0	1	0	1	52	0.019230769	1
ORF101A_TTG	0	1	0	1	5	0.2	1
PSBA_GGA	0	1	0	1	52	0.019230769	1
RPL12_GGA	0	0	1	1	4	0.25	1
YCF15_CAC	0	1	0	2	7	0.285714286	0.5
ATPI_CGT	1	1	0	9	52	0.173076923	0.222222222
RPL12_GGG	0	0	1	1	4	0.25	1
YCF1_CAC	0	1	0	1	41	0.024390244	1
YCF1_CAA	0	1	0	1	41	0.024390244	1
GSK3_CAT	0	0	1	2	4	0.5	0.5
PSAI_ATT	0	1	0	1	52	0.019230769	1

Supplementary Table 7, Chapter 3: Protozoa: The first 100 lines out of 14,800 total lines are presented in this dissertation. The full table is available upon request.

Gene_And_Codon_Name	Num_Origin	Num_Loss	Root_Loss(0/1)	Total_Species_In_Smaller_Group	Total_Species	Percent_Species_In_Min	Total_Origin_And_Loss/Total_Species_In_Smaller_Group
PGAM_AGT	0	0	1	1	4	0.25	1
LSM7_CGA	0	0	1	1	4	0.25	1
PRMT2_CCC	1	0	1	2	5	0.4	1
LSM7_CGC	0	0	1	1	4	0.25	1
PRMT2_CCA	0	1	0	1	5	0.2	1
PRMT2_CCG	0	1	0	2	5	0.4	0.5
LSM7_CGG	0	0	1	1	4	0.25	1
GEMA_TCG	0	1	0	1	4	0.25	1
CSNK2B_GTC	0	1	0	1	4	0.25	1
PGAM_AGA	1	0	0	1	4	0.25	1
CSNK2B_GTG	0	1	0	2	4	0.5	0.5
RPIA_AAC	0	1	0	1	4	0.25	1
PRMT2_CCT	0	1	0	1	5	0.2	1
APM4_GTC	0	1	0	1	4	0.25	1
APM4_GTG	0	1	0	1	4	0.25	1
PURD_ACT	0	0	1	1	4	0.25	1
ABCA4_CAT	0	1	0	1	5	0.2	1
BUD31_CAC	0	1	0	2	4	0.5	0.5
HUS1_GTG	0	1	0	1	4	0.25	1
GPI14_TAG	0	0	1	2	4	0.5	0.5
COG3_GGC	0	1	0	2	5	0.4	0.5
GPI14_TAA	0	0	1	2	4	0.5	0.5
RFC1_AGA	0	1	0	1	5	0.2	1
PDE4_AGG	1	0	1	2	4	0.5	1
PURD_GGC	0	1	0	1	4	0.25	1
SMC1_GTA	0	1	0	1	6	0.166666667	1
PDE4_AGA	0	0	1	1	4	0.25	1
ACGA_GCG	0	1	0	1	4	0.25	1
HEMF_CAA	0	0	1	1	4	0.25	1
RPL13_TCA	0	2	0	2	5	0.4	1
RPL13_TCG	0	0	1	2	5	0.4	0.5
HEMF_CAG	0	1	0	1	4	0.25	1
RPE_GCG	0	0	1	3	6	0.5	0.333333333
H1_CGG	0	0	1	1	4	0.25	1
PDE4_AGT	0	0	1	1	4	0.25	1
H1_CGA	1	0	1	2	4	0.5	1

H1_CGC	0	0	1	1	4	0.25	1
ALG11_TGC	0	1	0	2	5	0.4	0.5
SOD2_TAG	2	0	0	2	5	0.4	1
QSOX_TTA	0	1	0	1	4	0.25	1
SOD2_TAC	0	1	0	1	5	0.2	1
LEUS_ATA	0	1	0	1	4	0.25	1
SOD2_TAA	0	2	0	2	5	0.4	1
VAMP7A_GTC	0	1	0	1	4	0.25	1
RPA43_TGG	0	0	1	1	4	0.25	1
VAMP7A_GTG	0	1	0	2	4	0.5	0.5
SOD2_GCG	0	0	1	2	5	0.4	0.5
RAB21_GTC	0	1	0	1	4	0.25	1
RPL35_GCA	0	0	1	1	4	0.25	1
AAT19_TAA	0	0	1	2	4	0.5	0.5
RPL35_GCG	0	0	1	1	4	0.25	1
GTF2A2_GGT	0	1	0	1	4	0.25	1
TRYR_TGT	0	0	1	2	4	0.5	0.5
RPA43_TGC	1	0	1	2	4	0.5	1
ARL5_GGC	0	0	1	1	5	0.2	1
TOP2_GGG	1	0	1	2	7	0.285714286	1
PSMB7_GCG	0	0	1	1	4	0.25	1
ABCD3_GGG	0	0	1	2	5	0.4	0.5
PFP1_TCG	0	1	0	1	4	0.25	1
SF3B5_TCC	0	1	1	2	4	0.5	1
RFC4_CGT	0	1	0	1	5	0.2	1
PFP1_TCC	1	0	0	1	4	0.25	1
SF3B5_TCG	1	0	0	1	4	0.25	1
PRP19_GTG	0	1	1	2	4	0.5	1
PTER_CCT	0	1	0	1	4	0.25	1
SMC1_GCT	0	1	0	1	6	0.166666667	1
MMSDH_TGC	0	1	0	1	4	0.25	1
ABCD3_GGC	1	0	1	2	5	0.4	1
BZPG_TCG	0	1	0	1	4	0.25	1
USP7_TAA	1	0	0	2	4	0.5	0.5
RPL30_ACG	0	1	1	2	6	0.333333333	1
CYSB_CCC	1	0	1	2	7	0.285714286	1
RPL30_ACC	0	1	0	1	6	0.166666667	1
CYSB_CCG	0	1	0	1	7	0.142857143	1
RPL30_ACA	2	0	0	2	6	0.333333333	1
ARF1_TCT	0	1	0	1	4	0.25	1
AAT1.3_ATA	0	1	0	1	4	0.25	1
CLPP_AGT	0	1	0	1	4	0.25	1
LSM7_CGT	0	1	0	1	4	0.25	1

RAB6_AGC	0	2	0	3	6	0.5	0.666666667
RPL30_ACT	1	1	1	3	6	0.5	1
RAB6_AGG	1	1	1	3	6	0.5	1
SPT6_GGG	1	0	1	2	4	0.5	1
CLPP_AGA	0	1	0	1	4	0.25	1
TAT_GGG	1	0	1	2	7	0.285714286	1
CLPP_AGC	0	0	1	1	4	0.25	1
MYBC_CGA	1	0	1	2	4	0.5	1
TAT_GGC	0	1	0	2	7	0.285714286	0.5
MYBC_CGC	1	0	0	1	4	0.25	1
RAGC_AGC	0	1	0	1	4	0.25	1
RAB1A_TTG	0	1	0	1	5	0.2	1
CSN1_GCG	0	1	0	1	5	0.2	1
RAGC_AGG	1	0	0	1	4	0.25	1
PPA1_CAG	0	1	0	2	4	0.5	0.5
DDX6_GTG	0	1	0	1	4	0.25	1
LAP_GCG	0	1	1	4	8	0.5	0.5
NT3_ACA	0	1	0	1	4	0.25	1
PDSA_TCG	0	1	0	1	4	0.25	1
ABCG4_TGA	1	0	0	1	5	0.2	1
MED4_GGC	0	1	0	1	4	0.25	1

Supplementary Table 8, Chapter 3: Mammals: The first 100 lines out of 491,758 total lines are presented in this dissertation. The full table is available upon request.

Gene_And_Codon_Name	Num_Origin	Num_Loss	Root_Loss (0/1)	Total_Species_In_Smaller_Group	Total_Species	Percent_Species_In_Min	Total_Origin_And_Loss/Total_Species_In_Smaller_Group
ULBP1_GAC	0	1	0	2	6	0.333333333	0.5
CUNH8ORF59_TCG	1	0	0	1	15	0.066666667	1
RFC1_AGG	0	1	0	1	84	0.011904762	1
ULBP1_GAG	0	1	0	2	6	0.333333333	0.5
CUNH21ORF140_GCT	0	3	0	4	15	0.266666667	0.75
CUNH21ORF140_GCA	0	2	0	2	15	0.133333333	1
CUNH21ORF140_GCG	5	1	0	6	15	0.4	1
ULBP1_GAT	0	1	0	2	6	0.333333333	0.5
RPA1_CCA	0	1	0	1	93	0.010752688	1
FGL2_CCA	0	1	0	1	90	0.011111111	1
CNBP_TTG	3	1	0	6	95	0.063157895	0.666666667
CNBP_TTA	1	0	0	1	95	0.010526316	1
CNBP_TTC	0	1	0	1	95	0.010526316	1
MTRF1L_GTA	0	6	0	6	86	0.069767442	1
PRSS50_TAG	1	0	0	1	78	0.012820513	1
MTRF1L_GTC	0	2	0	2	86	0.023255814	1
CUNH1ORF64_ATT	0	2	0	5	13	0.384615385	0.4
MTRF1L_GTG	0	2	0	2	86	0.023255814	1
PRSS50_TAA	1	1	0	2	78	0.025641026	1
CNBP_TTT	0	6	0	6	95	0.063157895	1
PYY_GCA	6	14	0	35	77	0.454545455	0.571428571
PYY_GCG	2	6	0	10	77	0.12987013	0.8
MGAT4A_TGT	0	1	0	1	91	0.010989011	1
PRSS50_TAT	0	2	0	2	78	0.025641026	1
MTRF1L_GTT	0	7	0	7	86	0.081395349	1
RGS2_CTA	1	7	1	17	89	0.191011236	0.529411765
NDEL1_TAA	4	20	0	30	92	0.326086957	0.8
NDEL1_TAC	2	6	0	12	92	0.130434783	0.666666667
NDEL1_TAG	1	0	0	1	92	0.010869565	1
ARX_AAA	0	3	0	3	68	0.044117647	1
PRKCDBP_CGA	2	9	0	13	78	0.166666667	0.846153846
SMIM5_TAA	1	2	0	3	91	0.032967033	1
ROGDI_CAA	0	3	0	3	84	0.035714286	1
SMIM5_TAC	10	3	0	29	91	0.318681319	0.448275862
SLC35B2_TTG	0	4	1	13	87	0.149425287	0.384615385
SLC35B2_TTA	4	7	0	18	87	0.206896552	0.611111111
ROGDI_CAT	3	14	1	24	84	0.285714286	0.75

SMIM5_TAT	2	0	0	2	91	0.021978022	1
DOK1_GTC	1	7	0	8	89	0.08988764	1
PRKCDBP_CGT	9	7	0	35	78	0.448717949	0.457142857
FATE1_CGA	4	10	0	20	70	0.285714286	0.7
FATE1_CGC	1	3	0	4	70	0.057142857	1
PTPMT1_TGC	1	8	0	14	84	0.166666667	0.642857143
PTPMT1_TGA	2	6	1	16	84	0.19047619	0.5625
SNRNP40_GGG	0	1	0	1	89	0.011235955	1
SSNA1_GTA	4	1	0	8	82	0.097560976	0.625
CUNH1ORF64_ATA	1	3	1	6	13	0.461538462	0.833333333
SSNA1_GTG	0	1	0	1	82	0.012195122	1
DMXL2_TAG	2	8	1	18	73	0.246575342	0.611111111
CCDC63_AGT	0	1	0	1	87	0.011494253	1
PTPMT1_TGT	1	6	0	7	84	0.083333333	1
FATE1_CGT	0	5	0	6	70	0.085714286	0.833333333
DMXL2_TAA	5	2	0	13	73	0.178082192	0.538461538
GNG5_AGC	0	13	1	14	72	0.194444444	1
SSNA1_GTT	9	1	0	12	82	0.146341463	0.833333333
FAM228B_ATC	0	4	0	9	59	0.152542373	0.444444444
MLX_TTG	2	7	1	12	92	0.130434783	0.833333333
ODF2L_CCA	0	9	0	11	85	0.129411765	0.818181818
FAM173B_GAG	1	5	0	7	87	0.08045977	0.857142857
FAM173B_GAA	0	1	0	1	87	0.011494253	1
FGF22_GCT	2	7	0	18	76	0.236842105	0.5
FAM173B_GAC	0	1	0	1	87	0.011494253	1
ORC4_TAG	2	0	0	3	88	0.034090909	0.666666667
ORC4_TAA	5	0	0	8	88	0.090909091	0.625
FAM173B_GAT	2	3	0	5	87	0.057471264	1
FGF22_GCC	0	1	0	1	76	0.013157895	1
FGF22_GCA	3	7	0	17	76	0.223684211	0.588235294
ODF2L_CCT	4	8	0	20	85	0.235294118	0.6
FGF22_GCG	1	4	1	6	76	0.078947368	1
RBM19_TAT	0	3	0	3	80	0.0375	1
TMEM259_ACA	0	2	0	2	72	0.027777778	1
UQCC3_TCC	0	1	0	2	4	0.5	0.5
ZNF101_TGT	0	1	0	1	17	0.058823529	1
UQCC3_TCA	0	0	1	1	4	0.25	1
UQCC3_TCG	0	1	0	1	4	0.25	1
FAM84A_ACA	4	7	0	31	80	0.3875	0.35483871
GNG5_AGA	3	5	0	16	72	0.222222222	0.5
MSANTD4_GCA	0	1	0	1	88	0.011363636	1
SLC35D2_AGA	0	10	0	12	90	0.133333333	0.833333333
ZNF101_TGA	4	2	0	7	17	0.411764706	0.857142857

USP8_CGC	0	2	0	2	84	0.023809524	1
ZNF101_TGC	1	3	0	4	17	0.235294118	1
UQCC3_TCT	1	0	0	2	4	0.5	0.5
ALAS1_GCG	1	1	0	7	81	0.086419753	0.285714286
LRRC70_CCC	1	4	0	7	53	0.132075472	0.714285714
MAN2A2_ACG	0	1	0	1	87	0.011494253	1
AP5Z1_AGA	0	1	0	1	83	0.012048193	1
EEF1AKMT3_TCA	0	4	0	5	13	0.384615385	0.8
CTSH_TCT	0	1	0	3	88	0.034090909	0.333333333
KCNF1_AAT	9	7	0	30	92	0.326086957	0.533333333
UIMC1_CGT	0	7	0	8	91	0.087912088	0.875
CD83_CGT	7	2	0	20	86	0.23255814	0.45
TSTA3_GCT	0	2	0	2	89	0.02247191	1
CTSH_TCA	0	1	0	1	88	0.011363636	1
MAN2A2_ACT	0	2	0	2	87	0.022988506	1
CTSH_TCG	4	14	0	35	88	0.397727273	0.514285714
CD83_CGA	1	6	1	12	86	0.139534884	0.666666667
MTMR11_TGA	4	6	0	18	72	0.25	0.555555556
CD83_CGC	3	9	0	13	86	0.151162791	0.923076923
KCNF1_AAA	5	10	0	34	92	0.369565217	0.441176471

Supplementary Table 9, Chapter 3: The first 100 lines out of 456,905 total lines are presented in this dissertation. The full table is available upon request.

Gene_And_Codon_Name	Num_Origin	Num_Loss	Root_Loss(0/1)	Total_Species_In_Smaller_Group	Total_Species	Percent_Species_In_Min	Total_Origin_And_Loss/Total_Species_In_Smaller_Group
GNG5_AGG	16	2	0	23	97	0.237113402	0.782608696
TCTEX1D1_AGG	1	2	1	5	78	0.064102564	0.8
TCTEX1D1_AGA	10	5	0	34	78	0.435897436	0.441176471
RAC3_ACT	0	12	0	19	75	0.253333333	0.631578947
CUNH21ORF140_GCC	0	1	0	1	4	0.25	1
CUNH21ORF140_GCA	0	0	1	1	4	0.25	1
PLEKHB1_CGG	1	4	0	5	56	0.089285714	1
CUNH21ORF140_GCG	0	0	1	1	4	0.25	1
FGL2_CCA	0	1	0	1	95	0.010526316	1
CNBP_TTG	4	3	0	12	91	0.131868132	0.583333333
FGL2_CCG	2	14	0	18	95	0.189473684	0.888888889
PYY_GCT	0	1	0	1	6	0.166666667	1
CAAP1_GTA	6	11	0	20	91	0.21978022	0.85
MTRF1L_GTC	0	2	0	2	91	0.021978022	1
PYY_GCC	0	0	1	1	6	0.166666667	1
TSTA3_GCG	2	8	0	11	85	0.129411765	0.909090909
PYY_GCA	0	0	1	1	6	0.166666667	1
PYY_GCG	0	2	0	3	6	0.5	0.666666667
TSTA3_GCA	0	2	0	2	85	0.023529412	1
TMEM9_CGA	12	5	0	41	97	0.422680412	0.414634146
NDEL1_TAA	8	4	0	13	82	0.158536585	0.923076923
NDEL1_TAG	3	0	0	3	82	0.036585366	1
SMIM5_TAA	4	0	0	11	71	0.154929577	0.363636364
SMIM5_TAC	9	7	0	32	71	0.450704225	0.5
SLC35B2_TTG	0	2	0	2	80	0.025	1
NDEL1_TAT	0	1	0	1	82	0.012195122	1
SLC35B2_TTA	7	7	1	31	80	0.3875	0.483870968
ROGDI_CAT	0	1	0	1	92	0.010869565	1
SMIM5_TAT	4	10	0	31	71	0.436619718	0.451612903
FATE1_CGA	0	3	0	3	45	0.066666667	1
FATE1_CGC	0	1	0	1	45	0.022222222	1
PTPMT1_TGC	0	1	0	1	99	0.01010101	1
PTPMT1_TGA	7	4	0	31	99	0.313131313	0.35483871
FATE1_CGG	0	1	0	1	45	0.022222222	1
FAM228B_ATT	1	5	1	7	76	0.092105263	1
SSNA1_GTA	5	6	0	24	60	0.4	0.458333333

SSNA1_GTC	0	4	0	4	60	0.066666667	1
TCTEX1D1_AGT	6	9	0	21	78	0.269230769	0.714285714
SSNA1_GTG	3	8	1	25	60	0.416666667	0.48
DMXL2_TAG	1	8	1	35	95	0.368421053	0.285714286
CCDC63_AGT	0	3	0	5	54	0.092592593	0.6
PTPMT1_TGT	5	11	0	19	99	0.191919192	0.842105263
SNRNP40_GGT	7	7	0	36	94	0.382978723	0.388888889
DMXL2_TAA	8	3	1	24	95	0.252631579	0.5
SSNA1_GTT	10	5	0	28	60	0.466666667	0.535714286
FAM228B_ATA	2	3	0	10	76	0.131578947	0.5
FAM228B_ATC	0	4	0	6	76	0.078947368	0.666666667
ORC4_TAT	1	7	0	9	91	0.098901099	0.888888889
GRB7_GAT	0	2	0	2	36	0.055555556	1
HNRNPL_AGA	0	3	0	3	40	0.075	1
ODF2L_CCA	5	11	0	24	63	0.380952381	0.666666667
FAM173B_GAG	0	1	0	1	58	0.017241379	1
FAM173B_GAA	0	1	0	1	58	0.017241379	1
FGF22_GCT	2	8	0	11	56	0.196428571	0.909090909
ODF2L_CCG	10	0	0	12	63	0.19047619	0.833333333
ORC4_TAG	2	1	1	19	91	0.208791209	0.210526316
ORC4_TAA	6	2	0	9	91	0.098901099	0.888888889
ORC4_TAC	0	2	0	2	91	0.021978022	1
FGF22_GCA	0	9	0	14	56	0.25	0.642857143
ODF2L_CCT	3	10	0	14	63	0.222222222	0.928571429
FGF22_GCG	1	9	1	16	56	0.285714286	0.6875
USP8_CGT	0	1	0	1	98	0.010204082	1
HNRNPL_AGG	0	1	0	1	40	0.025	1
COMMD4_TAG	3	0	0	3	72	0.041666667	1
ARFGAP2_GCG	0	1	0	1	93	0.010752688	1
ITGAV_CGG	6	6	0	17	81	0.209876543	0.705882353
USP8_CGC	0	1	0	1	98	0.010204082	1
ALAS1_GCG	1	6	0	8	100	0.08	0.875
EEF1AKMT3_TCG	0	2	1	3	8	0.375	1
EEF1AKMT3_TCA	0	2	0	2	8	0.25	1
TCEANC2_CAT	0	4	0	4	102	0.039215686	1
KCNF1_AAT	0	3	0	3	91	0.032967033	1
CD83_CGT	8	10	0	33	79	0.417721519	0.545454545
MTMR11_TGT	0	6	0	6	48	0.125	1
TCEANC2_CAA	1	5	0	10	102	0.098039216	0.6
SLC46A2_GCG	3	10	0	15	89	0.168539326	0.866666667
CTSH_TCC	0	2	0	2	89	0.02247191	1
TCEANC2_CAG	1	3	0	4	102	0.039215686	1
CD83_CGA	8	10	0	32	79	0.405063291	0.5625

MTMR11_TGA	1	6	0	10	48	0.208333333	0.7
CD83_CGC	7	10	1	37	79	0.46835443	0.486486486
DFNB31_CGT	0	3	0	3	45	0.066666667	1
CD83_CGG	11	6	1	29	79	0.367088608	0.620689655
CUNH5ORF49_AGC	1	1	1	3	18	0.166666667	1
FAM104A_TGG	1	3	1	7	72	0.097222222	0.714285714
DONSON_TAT	0	1	0	1	96	0.010416667	1
FAM104A_TGA	0	1	0	1	72	0.013888889	1
CUNH5ORF49_AGG	0	0	1	1	18	0.055555556	1
FAM104A_TGC	1	11	0	15	72	0.208333333	0.8
SRSF2_GCT	1	4	2	7	50	0.14	1
MRAS_CGT	3	5	0	28	97	0.288659794	0.285714286
FAM104A_TGT	1	5	0	9	72	0.125	0.666666667
DONSON_TAA	6	1	0	23	96	0.239583333	0.304347826
CUNH5ORF49_AGT	1	3	0	5	18	0.277777778	0.8
DONSON_TAG	5	5	0	40	96	0.416666667	0.25
MRAS_CGC	11	4	1	19	97	0.195876289	0.842105263
MRAS_CGA	1	4	0	5	97	0.051546392	1
MRAS_CGG	0	3	0	3	97	0.030927835	1
CD99L2_ATA	5	1	0	12	53	0.226415094	0.5
TRIL_TTT	0	1	0	3	71	0.042253521	0.333333333

Supplementary Table 10, Chapter 3: All species: The number of species in the smaller clade, the total number of species, the number of clades with this cladal distribution, the number of groups with this cladal distribution which are expected to be consistent with the OTL based on random chance, the number of observed groups that were consistent with the OTL. The first 100 lines out of 62,417 total lines are presented in this dissertation. The full table is available upon request.

Number_of_species_in_smaller_clade	Total_number_of_species	Number_of_clades_with_this_separation	Number_of_expected_monophyletic_groups	Number_of_observed_monophyletic_groups
134	2035	1	1.00E-33	0
7	2051	2	1.95E-17	0
12	687	1	2.73E-24	0
17	979	1	3.38E-35	0
476	1168	1	1.00E-33	0
950	2161	1	1.00E-33	0
21	727	1	1.91E-39	0
5	178	342	8.65E-06	1
279	2814	1	1.00E-33	0
229	1378	1	1.00E-33	0
337	866	1	1.00E-33	0
173	1532	1	1.00E-33	0
67	137	21	3.74E-39	0
51	524	1	4.06E-71	0
39	171	53	4.27E-37	0
53	166	26	8.63E-43	0
559	1153	1	1.00E-33	0
76	201	1	5.92E-57	0
110	2429	1	1.00E-33	0
661	2163	1	1.00E-33	0
99	312	1	1.41E-83	0
25	701	1	4.83E-45	0
168	912	1	1.00E-33	0
122	635	1	1.00E-33	0
85	373	1	1.02E-85	0
32	367	1	1.03E-45	0
449	2797	1	1.00E-33	0
960	4115	1	1.00E-33	0
328	748	1	1.00E-33	0
2	1867	1	0.000535906	0
493	1008	1	1.00E-33	0
4	781	1	1.27E-08	0
1177	2509	1	1.00E-33	0
82	841	1	4.25E-115	0

74	543	2	3.77E-92	0
29	334	1	2.31E-41	0
264	2694	1	1.00E-33	0
142	919	1	1.00E-33	0
15	143	56	6.98E-18	0
255	522	1	1.00E-33	0
38	228	2	4.12E-43	0
2034	4909	1	1.00E-33	0
42	294	1	4.56E-51	0
77	578	1	4.69E-97	0
8	87	46	8.56E-09	0
123	396	1	2.06E-105	0
284	2376	1	1.00E-33	0
56	323	1	1.99E-63	0
19	313	1	1.34E-29	0
81	468	1	2.70E-92	0
153	751	1	1.00E-33	0
358	999	2	1.00E-33	0
28	185	74	4.24E-31	0
51	200	3	8.81E-48	0
119	308	1	3.25E-88	0
74	299	1	1.65E-71	0
58	1204	1	4.15E-99	0
277	662	1	1.00E-33	0
60	266	1	1.61E-60	0
62	456	1	2.49E-77	0
16	1760	1	2.91E-37	0
50	1564	1	4.08E-94	0
2	399	1	0.002512563	0
110	1492	2	1.00E-33	0
57	712	1	1.30E-84	0
41	179	46	4.16E-39	0
276	1454	1	1.00E-33	0
64	252	3	1.98E-60	0
27	992	1	7.10E-52	0
11	1447	1	9.37E-26	0
1127	2861	1	1.00E-33	0
52	243	1	1.21E-53	0
125	267	2	6.33E-79	0
444	980	1	1.00E-33	0
239	853	1	1.00E-33	0
13	467	2	1.05E-23	0
170	1228	1	1.00E-33	0

98	407	2	4.86E-96	0
61	257	1	4.93E-60	0
45	764	1	1.39E-72	0
227	745	1	1.00E-33	0
174	992	1	1.00E-33	0
47	338	2	1.48E-57	0
15	1060	1	4.26E-32	0
197	449	1	1.00E-33	0
54	586	1	1.06E-76	0
426	2761	1	1.00E-33	0
3	152	317	0.02799117	10
593	1192	1	1.00E-33	0
26	219	4	8.97E-33	0
10	740	2	1.16E-20	0
18	1166	1	2.98E-38	0
28	637	1	3.86E-48	0
12	90	58	1.59E-12	0
108	512	1	3.04E-113	0
14	568	2	2.29E-26	0
19	1166	1	4.68E-40	0
114	404	1	3.24E-103	0
166	727	1	1.00E-33	0
242	1065	1	1.00E-33	0

Supplementary Table 11, Chapter 3: Archaea: The first 100 lines out of 2,321 total lines are presented in this dissertation. The full table is available upon request.

Number_of_species_in_smaller_clade	Total_number_of_species	Number_of_clades_with_this_separation	Number_of_expected_monophyletic_groups	Number_of_observed_monophyletic_groups
14	74	4	4.64E-14	0
25	78	7	1.28E-19	0
4	36	22	0.003361345	0
5	178	2	5.06E-08	0
8	63	6	1.22E-08	0
7	25	12	8.92E-05	4
11	90	3	5.90E-13	0
16	47	1	1.95E-12	0
12	42	5	1.58E-09	0
2	78	15	0.194805195	1
5	195	1	1.75E-08	0
15	30	4	5.16E-08	0
19	91	5	1.32E-18	0
20	75	8	3.74E-17	0
42	88	4	3.49E-25	0
5	84	2	1.09E-06	0
20	58	8	1.26E-14	0
23	95	2	1.27E-21	0
8	87	2	3.72E-10	0
32	77	1	5.22E-22	0
14	77	1	6.55E-15	0
2	214	3	0.014084507	0
4	35	5	0.000835561	0
8	38	8	7.77E-07	0
7	22	9	0.000165856	0
11	83	3	1.40E-12	0
16	38	13	1.39E-09	0
2	73	17	0.236111111	5
4	178	1	1.10E-06	0
29	214	1	1.25E-35	0
12	50	2	6.86E-11	0
16	181	2	7.06E-22	0
17	57	1	2.40E-14	0
20	82	7	4.62E-18	0
5	93	4	1.43E-06	0
61	144	1	8.52E-42	0
10	104	2	7.97E-13	0

3	11	62	1.377777778	5
25	58	10	1.33E-15	0
8	94	5	5.28E-10	0
32	68	1	8.39E-20	0
51	160	2	2.98E-42	0
14	84	1	1.89E-15	0
21	46	3	9.46E-13	0
4	42	7	0.00065666	0
8	45	4	1.04E-07	0
7	15	33	0.010989011	0
47	148	1	3.00E-39	0
36	80	3	9.21E-23	0
2	64	13	0.206349206	0
5	53	2	7.39E-06	0
27	72	2	1.13E-19	0
12	41	8	3.46E-09	0
13	181	1	6.02E-19	0
33	148	1	4.46E-33	0
17	50	2	5.97E-13	0
20	89	2	2.24E-19	0
5	70	2	2.31E-06	0
17	195	1	9.81E-24	0
4	144	1	2.10E-06	0
8	69	3	3.09E-09	0
32	91	8	6.14E-24	0
33	67	2	2.85E-19	0
14	95	3	9.96E-16	0
3	129	6	0.000738189	0
4	17	24	0.042857143	0
27	81	7	9.11E-21	0
12	81	7	6.68E-13	0
36	95	6	7.91E-26	0
2	91	18	0.2	0
17	64	4	1.09E-14	0
4	160	3	4.56E-06	0
5	62	2	3.83E-06	0
26	85	3	1.95E-21	0
27	65	3	4.99E-18	0
17	43	5	3.00E-11	0
58	159	1	2.06E-44	0
5	79	9	6.31E-06	0
10	26	12	5.87E-06	0
30	89	1	6.61E-24	0

31	85	4	7.39E-23	0
8	76	6	3.02E-09	0
32	82	5	2.16E-22	0
33	68	2	1.49E-19	0
14	38	11	3.09E-09	0
18	93	2	1.42E-18	0
21	48	3	3.07E-13	0
4	24	15	0.008469791	0
11	45	4	1.61E-09	0
11	70	1	2.94E-12	0
37	80	4	1.00E-22	0
2	82	19	0.234567901	0
5	39	16	0.000216758	0
9	60	1	4.51E-10	0
18	178	1	4.79E-24	0
5	72	1	1.03E-06	0
6	78	5	2.53E-07	0
30	64	11	1.45E-17	0
31	82	4	2.84E-22	0
37	178	1	2.02E-38	0

Supplementary Table 12, Chapter 3: Bacteria: The first 100 lines out of 58,369 total lines are presented in this dissertation. The full table is available upon request.

Number_of_species_in_smaller_clade	Total_number_of_species	Number_of_clades_with_this_separation	Number_of_expected_monophyletic_groups	Number_of_observed_monophyletic_groups
33	440	1	2.31E-49	0
43	201	1	3.30E-44	0
125	786	1	1.00E-33	0
109	365	1	1.65E-95	0
16	1030	1	9.44E-34	0
6	479	1	4.91E-12	0
21	727	1	1.91E-39	0
5	178	7	1.77E-07	0
81	468	1	2.70E-92	0
337	866	1	1.00E-33	0
173	1532	1	1.00E-33	0
67	137	3	5.35E-40	0
510	1171	1	1.00E-33	0
60	710	1	1.08E-87	0
661	2163	1	1.00E-33	0
99	312	1	1.41E-83	0
212	1017	1	1.00E-33	0
85	373	1	1.02E-85	0
87	459	1	1.80E-95	0
80	711	1	4.58E-107	0
351	2083	1	1.00E-33	0
493	1008	1	1.00E-33	0
207	774	1	1.00E-33	0
172	715	1	1.00E-33	0
74	543	2	3.77E-92	0
142	919	1	1.00E-33	0
15	143	7	8.73E-19	0
38	228	2	4.12E-43	0
578	1821	1	1.00E-33	0
54	2106	1	6.10E-107	0
8	87	6	1.12E-09	0
85	473	1	2.11E-95	0
1062	2283	1	1.00E-33	0
561	1341	1	1.00E-33	0
464	2493	1	1.00E-33	0
86	240	1	5.14E-67	0
17	287	2	3.20E-26	0
19	313	1	1.34E-29	0

153	751	1	1.00E-33	0
5	374	2	2.52E-09	0
325	1482	1	1.00E-33	0
49	910	1	4.28E-81	0
404	2880	1	1.00E-33	0
51	200	2	5.88E-48	0
14	508	2	9.94E-26	0
74	299	1	1.65E-71	0
34	809	1	1.91E-59	0
113	263	1	4.31E-77	0
60	266	1	1.61E-60	0
2909	6899	1	1.00E-33	0
83	731	2	1.73E-110	0
429	955	1	1.00E-33	0
463	2699	1	1.00E-33	0
834	2061	1	1.00E-33	0
585	1425	1	1.00E-33	0
2	399	1	0.002512563	0
44	2538	1	3.55E-94	0
57	712	1	1.30E-84	0
64	252	1	6.59E-61	0
27	992	1	7.10E-52	0
209	901	1	1.00E-33	0
304	683	1	1.00E-33	0
239	853	1	1.00E-33	0
59	435	1	1.35E-73	0
41	217	1	1.61E-44	0
32	193	1	1.76E-36	0
53	166	1	3.32E-44	0
45	764	2	2.78E-72	0
227	745	1	1.00E-33	0
47	338	6	4.43E-57	0
15	1060	1	4.26E-32	0
5	1425	1	5.86E-12	0
3	152	13	0.001147903	0
222	812	1	1.00E-33	0
12	987	1	4.93E-26	0
374	763	1	1.00E-33	0
12	90	7	1.92E-13	0
130	2017	1	1.00E-33	0
253	1074	1	1.00E-33	0
1202	2493	1	1.00E-33	0
5	1455	1	5.39E-12	0

166	727	1	1.00E-33	0
171	955	1	1.00E-33	0
242	1065	1	1.00E-33	0
445	1924	2	1.00E-33	0
487	2653	1	1.00E-33	0
32	727	1	3.22E-55	0
113	1256	1	1.00E-33	0
129	323	1	2.34E-93	0
41	527	1	5.41E-61	0
25	402	3	1.26E-38	0
465	3039	1	1.00E-33	0
50	967	1	1.14E-83	0
73	446	1	5.56E-85	0
645	3296	1	1.00E-33	0
15	51	5	5.33E-12	0
149	1009	1	1.00E-33	0
272	2780	1	1.00E-33	0
98	987	1	4.99E-137	0
652	1390	1	1.00E-33	0

Supplementary Table 13, Chapter 3: Fungi:

Number_of_species_in_smaller_clade	Total_number_of_species	Number_of_clades_with_this_separation	Number_of_expected_monophyletic_groups	Number_of_observed_monophyletic_groups
2	7	516	86	103
4	10	8	0.095238095	0
2	6	998	199.6	186
4	9	96	1.714285714	4
5	10	3	0.023809524	0
2	9	125	15.625	31
4	8	63	1.8	5
2	8	168	24	26
3	10	14	0.388888889	2
2	10	10	1.111111111	3
3	8	148	7.047619048	21
3	9	117	4.178571429	14
3	6	434	43.4	74
3	7	441	29.4	58
2	5	3074	768.5	635
2	4	3574	1191.333333	857

Supplementary Table 14, Chapter 3: Invertebrates:

Number_of_species_in_smaller_clade	Total_number_of_species	Number_of_clades_with_this_separation	Number_of_expected_monophyletic_groups	Number_of_observed_monophyletic_groups
4	41	4	0.000404858	0
2	41	2	0.05	0
9	41	3	3.90E-08	0
10	46	2	2.26E-09	0
14	46	4	5.48E-11	0
2	5	67	16.75	16
17	46	1	1.55E-12	0
18	45	4	5.83E-12	0
13	45	3	1.42E-10	0
3	45	3	0.003171247	0
4	46	4	0.000281889	0
2	42	4	0.097560976	0
8	41	9	4.83E-07	0
7	37	1	5.13E-07	0
2	29	4	0.142857143	0
2	27	5	0.192307692	1
7	41	3	7.82E-07	0
14	43	2	7.84E-11	0
20	45	4	2.84E-12	0
11	42	2	1.78E-09	0
2	6	26	5.2	2
11	46	3	9.40E-10	0
12	45	3	3.91E-10	0
18	46	2	1.81E-12	0
2	32	3	0.096774194	1
19	45	4	3.89E-12	0
8	38	1	9.71E-08	0
10	41	5	1.83E-08	0
16	38	3	3.20E-10	0
6	41	2	3.04E-06	0
16	41	4	9.94E-11	0
11	41	1	1.18E-09	0
3	27	2	0.006153846	0
13	38	2	1.08E-09	0
20	42	4	1.63E-11	0
5	27	1	6.69E-05	0
2	37	4	0.111111111	0
4	32	4	0.000889878	0
9	45	6	3.39E-08	0

5	37	3	5.09E-05	0
12	27	2	2.59E-07	0
10	42	5	1.43E-08	0
7	38	2	8.60E-07	0
6	42	1	1.33E-06	0
15	43	3	5.68E-11	0
14	33	3	8.64E-09	0
7	32	2	2.72E-06	0
15	33	1	2.12E-09	0
21	46	2	6.31E-13	0
4	42	5	0.000469043	0
2	38	2	0.054054054	0
8	45	7	1.83E-07	0
6	29	2	2.04E-05	0
5	43	3	2.68E-05	0
4	37	5	0.00070028	0
6	32	1	5.89E-06	0
16	32	4	1.33E-08	0
13	29	4	1.31E-07	0
3	6	10	1	2
17	45	3	7.20E-12	0
15	46	2	1.20E-11	0
12	41	6	2.60E-09	0
19	43	5	1.41E-11	0
18	42	2	1.32E-11	0
3	33	2	0.004032258	1
2	43	4	0.095238095	0
8	42	2	8.90E-08	0
22	46	1	2.65E-13	0
5	46	1	6.71E-06	0
9	43	5	4.24E-08	0
6	45	10	9.21E-06	0
16	46	1	2.90E-12	0
12	29	5	2.33E-07	0
10	32	3	1.49E-07	0
16	45	4	1.74E-11	0
10	29	3	4.34E-07	0
11	45	4	1.61E-09	0
15	45	3	2.61E-11	0
12	46	3	2.96E-10	0
3	43	12	0.013937282	0
2	33	1	0.03125	0
11	29	1	7.62E-08	0

6	27	2	3.04E-05	0
5	33	1	2.78E-05	0
9	38	2	5.18E-08	0
6	46	4	3.27E-06	0
10	37	3	3.19E-08	0
17	38	1	7.77E-11	0
16	42	5	7.88E-11	0
14	37	3	1.30E-09	0
12	32	4	4.72E-08	0
18	37	3	3.49E-10	0
15	32	3	1.13E-08	0
13	37	1	7.99E-10	0
4	27	4	0.001538462	0
9	29	3	9.65E-07	0
21	42	4	1.49E-11	0
3	38	2	0.003003003	0
4	33	2	0.000403226	0
10	43	1	2.24E-09	0
9	33	1	9.51E-08	0
7	33	2	2.21E-06	0
6	43	1	1.18E-06	0
10	38	4	3.22E-08	0
15	42	6	1.70E-10	0
14	38	2	5.61E-10	0
12	37	1	1.66E-09	0
18	38	3	1.89E-10	0
8	29	4	3.38E-06	0
5	29	1	4.88E-05	0
4	43	3	0.000261324	0
3	37	2	0.003174603	0
8	46	3	6.61E-08	0
5	42	5	4.94E-05	0
8	33	4	1.19E-06	0
7	27	2	8.69E-06	0
16	33	8	1.41E-08	0
11	33	3	4.65E-08	0
15	41	5	2.15E-10	0
12	42	2	6.33E-10	0
19	42	1	4.95E-12	0
18	43	1	3.93E-12	0
14	29	4	1.07E-07	0
13	43	1	9.04E-11	0
4	29	4	0.001221001	0

9	27	5	3.20E-06	0
3	32	3	0.006451613	0
8	43	2	7.41E-08	0
5	45	3	2.21E-05	0
21	45	5	2.84E-12	0
9	42	7	7.33E-08	0
23	46	1	2.43E-13	0
10	33	1	3.57E-08	0
7	43	3	5.72E-07	0
2	4	239	79.66666667	99
13	33	2	8.86E-09	0
19	41	4	3.53E-11	0
13	46	6	2.09E-10	0
3	42	4	0.004878049	0
4	45	7	0.000528541	0
2	45	11	0.25	2
5	32	2	6.36E-05	0
13	41	1	1.79E-10	0
9	37	3	9.91E-08	0
7	29	5	1.33E-05	0
11	38	4	1.15E-08	0
17	41	1	1.59E-11	0
17	37	3	4.11E-10	0
16	43	3	3.04E-11	0
14	42	5	2.84E-10	0
12	33	5	3.88E-08	0
11	43	3	2.04E-09	0
17	42	10	9.70E-11	0
3	41	2	0.002564103	0
10	27	4	1.28E-06	0
2	46	1	0.022222222	0
8	37	6	7.19E-07	0
6	37	1	2.65E-06	0
9	32	3	3.80E-07	0
16	37	4	7.18E-10	0
11	37	2	7.87E-09	0
7	45	7	9.92E-07	0
12	38	3	3.51E-09	0
19	46	3	1.75E-12	0
15	38	1	1.64E-10	0
20	41	1	7.62E-12	0
5	41	3	3.28E-05	0
9	46	6	2.78E-08	0

22	45	5	2.48E-12	0
6	38	2	4.59E-06	0
10	45	4	5.64E-09	0
14	45	6	1.16E-10	0
13	27	4	4.14E-07	0
11	32	4	9.02E-08	0
14	32	2	9.70E-09	0
12	43	3	7.01E-10	0
3	29	3	0.007936508	0
8	27	2	3.04E-06	0
15	37	1	2.63E-10	0
13	42	7	8.86E-10	0
20	46	1	4.10E-13	0
3	46	6	0.006060606	0

Supplementary Table 15, Chapter 3: Plants: The first 100 lines out of 478 total lines are presented in this dissertation. The full table is available upon request.

Number_of_species_in_smaller_clade	Total_number_of_species	Number_of_clades_with_this_separation	Number_of_expected_monophyletic_groups	Number_of_observed_monophyletic_groups
5	57	5	1.36E-05	0
16	47	10	1.95E-11	0
6	54	5	1.74E-06	0
17	46	2	3.09E-12	0
7	55	11	4.26E-07	0
11	48	7	1.35E-09	0
13	58	3	4.24E-12	0
2	27	4	0.153846154	0
24	52	9	4.57E-14	0
14	43	2	7.84E-11	0
20	58	5	7.85E-15	0
2	53	12	0.230769231	1
26	54	2	2.21E-15	0
22	48	1	7.97E-14	0
3	40	7	0.009446694	1
9	27	2	1.28E-06	0
6	23	1	3.80E-05	0
2	47	20	0.434782609	7
8	38	2	1.94E-07	0
7	22	3	5.53E-05	0
16	38	1	1.07E-10	0
6	41	1	1.52E-06	0
7	44	2	3.28E-07	0
12	50	3	1.03E-10	0
8	18	10	0.000514192	1
13	51	11	9.06E-11	0
2	18	33	1.941176471	9
14	50	4	1.52E-11	0
25	58	1	1.33E-16	0
15	43	1	1.89E-11	0
20	47	5	1.20E-12	0
3	17	14	0.116666667	3
23	50	4	8.05E-14	0
6	14	2	0.001554002	0
4	42	3	0.000281426	0
2	38	6	0.162162162	0
8	45	1	2.61E-08	0
5	43	7	6.25E-05	0

3	55	37	0.025856045	0
4	48	10	0.000616713	0
10	47	9	8.17E-09	0
8	51	13	1.30E-07	2
5	53	7	2.59E-05	0
11	46	2	6.27E-10	0
9	50	14	3.10E-08	1
16	51	11	4.89E-12	0
6	58	2	4.78E-07	0
12	41	1	4.33E-10	0
17	50	9	2.69E-12	0
13	44	1	6.52E-11	0
15	48	1	2.93E-12	0
19	53	4	9.37E-14	0
2	7	110	18.33333333	53
26	56	2	6.48E-16	0
5	10	3	0.023809524	0
20	54	3	2.52E-14	0
3	22	11	0.052380952	0
21	55	2	6.22E-15	0
4	17	9	0.016071429	0
18	56	3	4.40E-14	0
6	27	6	9.12E-05	0
8	58	1	3.78E-09	0
16	42	4	6.31E-11	0
10	48	8	5.87E-09	0
17	43	1	6.01E-12	0
7	56	8	2.76E-07	1
18	48	4	1.46E-12	0
15	57	1	1.72E-13	0
19	58	2	6.44E-15	0
21	48	2	2.05E-13	0
2	56	36	0.654545455	2
6	18	12	0.001939237	3
4	38	3	0.0003861	3
7	27	4	1.74E-05	0
6	52	27	1.15E-05	1
7	49	2	1.63E-07	0
11	50	9	1.10E-09	0
18	43	2	7.85E-12	0
12	53	4	6.62E-11	0
13	56	8	1.82E-11	0
24	54	1	1.60E-15	0

25	55	2	1.43E-15	0
15	44	1	1.28E-11	0
21	57	2	2.55E-15	0
2	51	66	1.32	23
26	52	2	8.07E-15	0
22	54	1	3.14E-15	0
3	42	10	0.012195122	1
27	57	1	1.50E-16	0
23	55	5	6.41E-15	0
6	21	3	0.000193498	0
4	45	2	0.000151012	0
2	45	6	0.136363636	3
8	40	1	6.50E-08	0
3	48	10	0.009250694	2
6	47	13	9.48E-06	4
7	46	4	4.91E-07	0
11	43	2	1.36E-09	0
13	49	6	8.61E-11	0
14	48	1	7.11E-12	0

Supplementary Table 16, Chapter 3: Protozoa:

Number_of_species_in_smaller_clade	Total_number_of_species	Number_of_clades_with_this_separation	Number_of_expected_monophyletic_groups	Number_of_observed_monophyletic_groups
2	7	193	32.16666667	91
4	10	13	0.154761905	8
2	6	550	110	200
4	9	58	1.035714286	0
5	10	5	0.03968254	0
2	9	40	5	4
4	8	94	2.685714286	4
2	8	159	22.71428571	33
3	10	14	0.388888889	3
2	10	27	3	12
3	8	171	8.142857143	23
3	12	3	0.054545455	0
3	9	41	1.464285714	4
2	12	6	0.545454545	4
3	6	245	24.5	114
6	12	1	0.002164502	0
3	7	194	12.93333333	98
2	5	1381	345.25	705
4	12	3	0.018181818	1
2	4	2401	800.3333333	1145
5	12	3	0.009090909	0

Supplementary Table 17, Chapter 3: Mammals: The first 100 lines out of 2,163 total lines are presented in this dissertation. The full table is available upon request.

Number_of_species_in_smaller_clade	Total_number_of_species	Number_of_clades_with_this_separation	Number_of_expected_monophyletic_groups	Number_of_observed_monophyletic_groups
14	74	76	8.81E-13	0
4	36	80	0.012223071	0
8	63	64	1.30E-07	0
7	25	35	0.000260037	0
11	90	668	1.31E-10	1
16	47	28	5.47E-11	0
2	78	957	12.42857143	91
12	59	35	1.54E-10	0
17	64	56	1.53E-13	0
19	91	375	9.90E-17	0
20	75	65	3.03E-16	0
42	88	232	2.03E-23	0
5	84	960	0.000522415	5
11	35	28	2.14E-07	0
20	58	31	4.87E-14	0
23	95	10	6.33E-21	0
8	87	799	1.49E-07	5
11	54	41	2.10E-09	0
32	77	77	4.02E-20	0
14	77	140	9.17E-13	0
4	35	73	0.012199198	2
8	38	22	2.14E-06	0
7	22	30	0.000552853	1
11	83	376	1.76E-10	0
16	38	28	2.99E-09	0
2	73	478	6.638888889	47
26	67	49	4.67E-17	0
12	50	49	1.68E-09	0
17	57	28	6.72E-13	0
20	82	164	1.08E-16	0
5	93	689	0.000246586	16
3	11	353	7.844444444	33
23	84	200	2.89E-18	0
8	94	187	1.97E-08	1
32	68	40	3.35E-18	0
33	74	53	1.04E-19	0
14	84	323	6.11E-13	0
21	46	30	9.46E-12	0

4	42	84	0.007879925	2
8	45	27	7.05E-07	0
7	15	287	0.095571096	5
11	72	115	2.49E-10	0
12	90	638	1.75E-11	4
23	49	21	7.67E-13	0
36	80	107	3.29E-21	0
2	64	280	4.444444444	32
5	53	68	0.000251177	1
26	74	65	2.80E-18	0
27	72	43	2.44E-18	0
12	41	27	1.17E-08	0
17	50	39	1.16E-11	0
20	89	427	4.79E-17	0
5	70	155	0.000179294	0
6	92	841	1.81E-05	0
30	86	223	4.98E-21	0
8	69	109	1.12E-07	0
32	91	211	1.62E-22	0
33	67	45	6.42E-18	0
14	95	12	3.98E-15	0
21	55	28	8.71E-14	0
4	17	101	0.180357143	4
9	43	29	2.46E-07	0
12	81	298	2.84E-11	0
36	95	4	5.27E-26	0
37	87	182	8.50E-23	0
2	91	2992	33.24444444	261
15	30	17	2.19E-07	0
5	62	90	0.000172462	1
26	85	204	1.32E-19	0
9	59	45	2.35E-08	0
12	32	21	2.48E-07	0
17	43	47	2.82E-10	0
5	79	403	0.000282524	1
6	87	1089	3.13E-05	3
10	26	42	2.06E-05	0
30	89	320	2.12E-21	0
31	85	206	3.81E-21	0
8	76	191	9.62E-08	0
32	82	122	5.26E-21	0
33	68	43	3.21E-18	0
14	38	25	7.02E-09	0

18	93	225	1.60E-16	0
21	48	25	2.56E-12	0
4	24	30	0.016939582	0
29	91	257	1.66E-21	0
11	70	67	1.97E-10	1
12	72	99	3.87E-11	0
37	80	105	2.64E-21	0
2	82	1587	19.59259259	154
5	39	47	0.000636727	0
26	92	226	1.41E-20	0
9	60	65	2.93E-08	0
17	44	12	4.53E-11	0
5	72	194	0.000199663	1
6	78	333	1.69E-05	0
30	64	42	5.53E-17	0
31	82	134	9.51E-21	0
33	93	116	2.04E-23	0
14	41	27	2.24E-09	0
18	84	269	1.32E-15	0

Supplementary Table 18, Chapter 3: Other Vertebrates: The first 100 lines out of 2,771 total lines are presented in this dissertation. The full table is available upon request.

Number_of_species_in_smaller_clade	Total_number_of_species	Number_of_clades_with_this_separation	Number_of_expected_monophyletic_groups	Number_of_observed_monophyletic_groups
14	74	88	1.02E-12	0
4	36	226	0.034530176	8
8	63	139	2.83E-07	1
7	25	90	0.000668668	2
11	90	193	3.80E-11	0
16	47	69	1.35E-10	0
2	78	718	9.324675325	79
12	59	110	4.83E-10	3
17	64	88	2.40E-13	0
19	91	102	2.69E-17	0
20	75	71	3.31E-16	0
42	88	54	4.72E-24	0
5	84	340	0.000185022	7
6	98	453	7.03E-06	0
10	97	243	1.87E-10	1
30	100	74	8.40E-24	2
11	35	96	7.32E-07	0
20	58	103	1.62E-13	0
23	95	110	6.96E-20	0
8	87	217	4.04E-08	0
32	77	72	3.76E-20	0
13	101	161	1.53E-13	0
14	77	96	6.29E-13	0
4	35	200	0.03342246	6
8	38	160	1.55E-05	0
7	22	132	0.002432552	2
11	83	122	5.70E-11	0
16	38	81	8.65E-09	0
2	73	548	7.611111111	64
26	67	40	3.81E-17	0
12	50	85	2.92E-09	0
52	105	5	3.22E-30	0
17	57	74	1.78E-12	0
19	96	125	1.12E-17	0
20	82	70	4.62E-17	0
5	93	477	0.000170714	9
6	101	296	3.93E-06	1
10	104	75	2.99E-11	0

31	103	35	5.91E-25	0
35	96	95	1.38E-24	0
3	11	366	8.133333333	47
23	84	69	9.98E-19	0
8	94	312	3.29E-08	0
32	68	73	6.12E-18	0
33	74	58	1.14E-19	0
14	84	122	2.31E-13	0
18	107	5	2.56E-19	0
21	46	68	2.15E-11	0
4	42	244	0.022889306	8
8	45	160	4.18E-06	1
7	15	161	0.053613054	8
29	62	47	2.45E-16	0
12	90	184	5.04E-12	0
23	49	65	2.37E-12	0
36	80	49	1.50E-21	0
2	64	454	7.206349206	62
5	53	184	0.000679656	2
26	74	54	2.33E-18	0
9	50	117	2.59E-07	1
12	41	69	2.98E-08	0
17	50	80	2.39E-11	0
19	105	21	3.16E-19	0
20	89	77	8.64E-18	0
5	70	198	0.000229034	3
6	92	342	7.35E-06	0
30	86	69	1.54E-21	0
35	105	9	3.09E-27	0
8	69	158	1.63E-07	0
32	91	74	5.68E-23	0
33	67	34	4.85E-18	0
14	95	165	5.48E-14	1
18	98	127	3.36E-17	0
37	98	80	1.57E-25	0
38	97	92	1.77E-25	0
21	55	69	2.15E-13	0
4	17	181	0.323214286	7
11	65	124	8.19E-10	1
12	81	132	1.26E-11	0
36	95	75	9.88E-25	0
37	87	68	3.18E-23	0
2	91	915	10.16666667	78

15	30	40	5.16E-07	0
5	62	229	0.000438819	4
26	85	72	4.67E-20	0
9	59	137	7.15E-08	0
12	32	69	8.15E-07	1
17	43	78	4.68E-10	0
20	96	139	3.09E-18	0
24	99	109	7.42E-21	0
5	79	281	0.000196996	2
6	87	289	8.30E-06	2
11	22	73	0.000206965	1
10	26	58	2.84E-05	0
30	89	67	4.43E-22	0
31	85	76	1.40E-21	0
40	99	88	2.64E-26	0
23	106	2	8.20E-23	0
8	76	167	8.41E-08	0
32	82	63	2.72E-21	0
33	68	57	4.25E-18	0

Supplementary Files

Supplementary File 1, Chapter 3: All Species: Species relationships from the OTL annotated in Newick format with homologous character state changes labeled. A range is given, which points to the order in the accompanying character state change file. This file is available upon request. However, it is approximately 150 pages long, so it is excluded from this dissertation.

Supplementary File 2, Chapter 3: Archaea: See description for Supplementary File 1

((((((((((PYRODICTIUM_OCCULTUM,PYRODICTIUM_DELANEYI),HYPERTHERMUS_BUTYLICUS,PYROLOBUS_FUMARII),(AEROPYRUM_CAMINI,(CALDISPHAERA_LAGUNENSIS,ACIDILOBUS_SACCHAROVORANS))"228"),((THERMOSPHAERA_AGGREGANS,(DESULFUROCOCCUS_AMYLOLYTICUS,DESULFUROCOCCUS_MUCOSUS)),(STAPHYLOTHERMUS_MARINUS,STAPHYLOTHERMUS_HELLENICUS)"1-35")"326-327")"365-375",((((SULFOLOBUS_SOLFATARICUS,SULFOLOBUS_ISLANDICUS)"92-127",(SULFOLOBUS_TOKODAII,SULFOLOBUS_ACIDOCALDARIUS))"264-325",((CANDIDATUS_ACIDIANUS_COPAHUENSIS,ACIDIANUS_HOSPITALIS),(METALLOSPHAERA_SEDULA,METALLOSPHAERA_CUPRINA,METALLOSPHAERA_YELLOWSTONENSIS))),SULFOLOBUS_SP._JCM_16833,SULFOLOBUS_SP._A20,SULFOLOBUS_METALLICUS)"380-384")"700-720",(IGNICOCCUS_ISLANDICUS,IGNICOCCUS_HOSPITALIS)"161-163"),(((CALDIVIRGA_MAQUILINGENSIS,CALDIVIRGA_SP._MU80),(VULCANISAETA_DISTRIBUTA,VULCANISAETA_MOUTNOVSKIA,VULCANISAETA_SP._EB80)"227"),VULCANISAETA_THERMOPHILA,((THERMOPROTEUS_TENAX,THERMOPROTEUS_UZONIENSIS)"134-142",PYROBACULUM_ISLANDICUM)"168-170"))"746-750",(THERMOFILUM_PENDENS,THERMOFILUM_CARBOXYDITROPHUS))"751-754",(((((PALAEOCOCCUS_PACIFICUS,PALAEOCOCCUS_FERROPHILUS),(((THERMOCOCCUS_SP._AM4,THERMOCOCCUS_GAMMATOLERANS),THERMOCOCCUS_KODAKARENSIS),THERMOCOCCUS_ONNURINEUS)),(THERMOCOCCUS_CHITONOPHAGUS,PYROCOCCUS_SP._NA2,PYROCOCCUS_SP._ST04,((PYROCOCCUS_ABYSSI,PYROCOCCUS_HORIKOSHII),PYROCOCCUS_FURIOSUS),PYROCOCCUS_YAYANOSII))"385-386",THERMOCOCCUS_BAROPHILUS),THERMOCOCCUS_PEPTONOPHILUS,THERMOCOCCUS_PACIFICUS,THERMOCOCCUS_CELERICRESCENS,THERMOCOCCUS_SP._4557,THERMOCOCCUS_PROFUNDUS,THERMOCOCCUS_SICULI,THERMOCOCCUS_RADIOTOLERANS,THERMOCOCCUS_LITORALIS,THERMOCOCCUS_GUAYMASENSIS,THERMOCOCCUS_CLEFTENSIS,THERMOCOCCUS_SP._P6,THERMOCOCCUS_BAROSSII,THERMOCOCCUS_THIOREDUCENS,THERMOCOCCUS_CELER,THERMOCOCCUS_GORGONARIUS,THERMOCOCCUS_SP._2319X1,THERMOCOCCUS_PIEZOPHILUS,THERMOCOCCUS_NAUTILI,THERMOCOCCUS_PARALVINELLAE,THERMOCOCCUS_EURYTHERMALIS,THERMOCOCCUS_ZILLIGII,THERMOCOCCUS_SP._EP1,THERMOCOCCUS_SIBIRICUS,THERMOCOCCUS_SP._PK)"755-800")"839-853",((((((PICROPHILUS_OSHIMAE,PICROPHILUS_TORRIDUS)"41-

44",((ACIDIPLASMA_CUPRICUMULANS,ACIDIPLASMA_AEOLICUM)"143-
147",(FERROPLASMA_ACIDIPHILUM,FERROPLASMA_ACIDARMANUS)))"349",
(THERMOPLASMA_VOLCANIUM,THERMOPLASMA_ACIDOPHILUM)"90-
91")"357",CUNICULIPLASMA_DIVULGATUM),CANDIDATUS_METHANOPLAS
MA_TERMITUM),(NITROSOPUMILUS_SP._SJ,CANDIDATUS_NITROSOPUMILU
S_ADRIATICUS,CANDIDATUS_NITROSOPUMILUS_SALARIA,NITROSOPUMIL
US_SP._NSUB,NITROSOPUMILUS_SP._AR)"329"))"863",((((METHANOCALDOC
OCCUS_VILLOSUS,METHANOCALDOCOCCUS_INFERNUS,((METHANOCALD
OCOCCUS_JANNASCHII,METHANOCALDOCOCCUS_FERVENS)"130",METHAN
OCALDOCOCCUS_VULCANIUS),METHANOCALDOCOCCUS_BATHOARDESCE
NS),((((METHANOTHERMOCOCCUS_OKINAWENSIS,METHANOTHERMOCOC
CUS_THERMOLITHOTROPHICUS)"148",METHANOCOCCUS_AEOLICUS),((MET
HANOCOCCUS_MARIPALUDIS,METHANOCOCCUS_VANNIELII)"149",METHA
NOCOCCUS_VOLTAE)),(METHANOTORRIS_FORMICICUS,METHANOTORRIS_
IGNEUS)))"387",(((((METHANOBREVIBACTER_SP._87.7,METHANOBREVIBACT
ER_GOTTSCHALKII,METHANOBREVIBACTER_SMITHII,METHANOBREVIBAC
TER_CURVATUS,METHANOBREVIBACTER_FILIFORMIS,METHANOBREVIBA
CTER_CUTICULARIS,METHANOBREVIBACTER_ARBORIPHILUS,METHANOB
REVIBACTER_MILLERAE,METHANOBREVIBACTER_WOLINII,METHANOBRE
VIBACTER_ORALIS,METHANOBREVIBACTER_OLLEYAE,METHANOBREVIB
ACTER_SP._A54,METHANOBREVIBACTER_SP._YE315,METHANOBREVIBACT
ER_BOVISKOREANI,METHANOBREVIBACTER_RUMINANTIUM,METHANOBR
EVIBACTER_SP._ABM4,METHANOBREVIBACTER_SP._A27)"388-
666",(METHANOSPHAERA_STADTMANAE,METHANOSPHAERA_SP._A6,METH
ANOSPHAERA_SP._WGK6)"229-
230"),(METHANOBACTERIUM_PALUDIS,METHANOBACTERIUM_FORMICICU
M,METHANOBACTERIUM_VETERUM,METHANOBACTERIUM_ARCTICUM,M
ETHANOBACTERIUM_LACUS,METHANOBACTERIUM_CONGOLENSE,METHA
NOBACTERIUM_SP._A39,METHANOBACTERIUM_SP._MB1)"358"),(METHANO
THERMOBACTER_WOLFEII,METHANOTHERMOBACTER_MARBURGENSIS,M
ETHANOTHERMOBACTER_THERMAUTOTROPHICUS,METHANOTHERMOBA
CTER_SP._CAT2)),METHANOTHERMUS_FERVIDUS)),(((((((((NATRIALBA_CHA
HANNAOENSIS,NATRIALBA_ASIATICA,NATRIALBA_AEGYPTIA,NATRIALBA
_HULUNBEIRENSIS,NATRIALBA_TAIWANENSIS,NATRIALBA_MAGADII,NAT
RIALBA_SP._SSL1)"350-
354",(NATRINEMA_SALACIAE,HALOTERRIGENA_SALINA,NATRINEMA_ALT
UNENSE,HALOTERRIGENA_JEOTGALI,HALOTERRIGENA_DAQINGENSIS,HA
LOTERRIGENA_SACCHAREVITANS,HALOTERRIGENA_THERMOTOLERANS,
HALOTERRIGENA_HISPANICA,NATRONORUBRUM_SEDIMINIS,HALOTERRIG
ENA_TURKMENICA,HALOTERRIGENA_LIMICOLA)"359-364")"667-
669",(HALOBIFORMA_HALOTERRESTRIS,HALOBIFORMA_LACISALSI)"151-
160",(NATRONOBACTERIUM_TEXCOCONENSE,NATRONOBACTERIUM_GREG
ORYI),(HALOPIGER_SALIFODINAE,HALOBIFORMA_NITRATIREDUCENS,HAL
OPIGER_XANADUENSIS),(NATRONOLIMNOBIUS_INNERMONGOLICUS,NATR
ONOLIMNOBIUS_BAERHUENSIS),(NATRINEMA_PALLIDUM,NATRINEMA_VE
RSIFORME,NATRINEMA_GARI),(HALOSTAGNICOLA_KAMEKURAE,HALOST

AGNICOLA_SP._A56,HALOSTAGNICOLA_LARSENII),HALOVIVAX_ASIATICUS,(NATRONORUBRUM_TEXCOCONENSE,NATRONORUBRUM_BANGENSE,NATRONORUBRUM_THIOOXIDANS,NATRONORUBRUM_TIBETENSE,NATRONORUBRUM_SULFIDIFACIENS)"330-335",(NATRONOCOCCUS_OCCULTUS,NATRONOCOCCUS_AMYLOLYTICUS,NATRONOCOCCUS_JEOTGALI)"223-226")"801-816",(((((HALOFERAX_ELONGANS,HALOFERAX_DENITRIFICANS,HALOFERAX_SP._SB3,HALOFERAX_SP._ATB1,HALOFERAX_SP._SB29,HALOFERAX_SP._Q22,HALOFERAX_MUCOSUM,HALOFERAX_MEDITERRANEI,HALOFERAX_SULFURIFONTIS,HALOFERAX_SP._BAB2207,HALOFERAX_ALEXANDRINUS,HALOFERAX_VOLCANII,HALOFERAX_PRAHOVENSE,HALOFERAX_GIBBONSII,HALOFERAX_LARSENII,HALOFERAX_LUCENTENSE),(HALOGEOMETRICUM_RUFUM,HALOGEOMETRICUM_BORINQUENSE)),HALOQUADRATUM_WALSBYI),(HALOGRANUM_SALARIUM,HALOGRANUM_AMYLOLYTICUM,HALOGRANUM_GELATINILYTICUM,HALOGRANUM_RUBRUM)"260-263",HALOBACULUM_GOMORRENSE,(HALORUBRUM_EZZEMOULENSE,HALORUBRUM_SODOMENSE,HALORUBRUM_CHAOVIATOR,HALORUBRUM_CALIFORNIENSE,HALORUBRUM_SP._T3,HALORUBRUM_CORIENSE,HALORUBRUM_LIPOLYTICUM,HALORUBRUM_LITOREUM,HALORUBRUM_HALOPHILUM,HALORUBRUM_SP._SD626R,HALORUBRUM_DISTRIBUTUM,HALORUBRUM_SP._EA8,HALORUBRUM_SP._SD683,HALORUBRUM_TERRESTRE,HALORUBRUM_TEBENQUICHENSE,HALORUBRUM_AIDINGENSE,HALORUBRUM_VACUOLATUM,HALORUBRUM_SACCHAROVORUM,HALORUBRUM_SP._HD13,HALORUBRUM_SP._SD612,HALORUBRUM_SP._EB13,HALORUBRUM_SP._BV1,HALORUBRUM_SP._EA1,HALORUBRUM_HALODURANS,HALORUBRUM_SP._AJ67,HALORUBRUM_SP._IB24,HALORUBRUM_HOCHSTENIUM)"721-732",HALOPARVUM_SEDIMENTI,(HALOPENITUS_PERSICUS,HALOPENITUS_MALEKZADEHII)"36-39",(HALOPELAGIUS_INORDINATUS,HALOPELAGIUS_LONGUS)"164-167",HALOHASTA_LITCHFIELDIAE,(HALOPLANUS_NATANS,HALOPLANUS_VESCUS),(HALOBELLUS_CLAVATUS,HALOBELLUS_RUFUS),(HALOPHILIC_ARCHAEON_DL31,HALOLAMINA_PELAGICA,HALOLAMINA_SP._CBA1230,HALOLAMINA_RUBRA,HALOLAMINA_SEDIMINIS))"817-838",(((HALOARCULA_SP._CBA1127,HALOARCULA_SP._K1,HALOARCULA_SP._CBA1128,HALOARCULA_SP._CBA1115,HALOARCULA_SINAIIENSIS,HALOARCULA_ARGENTINENSIS,HALOARCULA_CALIFORNIAE,HALOARCULA_HISPANICA,HALOARCULA_VALLISMORTIS,HALOARCULA_MARISMORTUI,HALOARCULA_AMYLOLYTICA,HALOARCULA_JAPONICA)"376-379",(HALOMICROBIUM_MUKOHATAEI,HALOMICROBIUM_KATESII)"131"),(HALORHABDUS_UTAHENSIS,HALORHABDUS_TIAMATEA)))"854-862")"864-877",(NATRONOMONAS_MOOLAPENSIS,NATRONOMONAS_PHARAONIS,NATRONOMONAS_SP._CBA1134))"878",(HALOBACTERIUM_HUBEIENSE,HALOBACTERIUM_SP._CBA1132,HALOBACTERIUM_SP._DL1,HALOBACTERIUM_SALINARUM,HALOBACTERIUM_JILANTAIENSE)"328",HALOVENUS_ARANENSIS,HALOGEOMETRICUM_LIMI,((HALALKALICOCCUS_JEOTGALI,HALALKALICOCCUS_PAUCIHALOPHILUS),(HALADAPTATUS_PAUCIHALOPHILUS,HALAD

APTATUS_CIBARIUS,HALADAPTATUS_LITOREUS,HALADAPTATUS_SP._W1,
HALADAPTATUS_SP._R4)),HALARCHAEUM_ACIDIPHILUM,CANDIDATUS_H
ALOBONUM_TYRRELLENSIS,HALANAEROARCHAEUM_SULFURIREDUCENS,
HALOPROFUNDUS_MARISRUBRI,HALAPRICUM_SALINUM,(HALOCOCCUS_S
ACCHAROLYTICUS,HALOCOCCUS_AGARILYTICUS,HALOCOCCUS_SALIFOD
INAE,HALOCOCCUS_THAILANDENSIS,HALOCOCCUS_SEDIMINICOLA,HALO
COCCUS_MORRHUAE,HALOCOCCUS_HAMELINENSIS),(HALOMICROBIUM_Z
HOUII,HALOSIMPLEX_CARLSBADENSE)"132-
133",HALOARCHAEOBIUS_IRANENSIS,(HALORIENTALIS_REGULARIS,HALO
RIENTALIS_PERSICUS))"879-
916",(((METHANOCORPUSCULUM_LABREANUM,METHANOCORPUSCULUM_
BAVARICUM)"128",METHANOSPIRILLUM_HUNGATEI),((((METHANOCULLEU
S_THERMOPHILUS,METHANOCULLEUS_MARISNIGRI,METHANOCULLEUS_B
OURGENSIS,METHANOCULLEUS_HORONOBENSIS,METHANOCULLEUS_CHI
KUGOENSIS,METHANOCULLEUS_SP._MAB1,METHANOCULLEUS_SEDIMINIS
,METHANOCULLEUS_SP._MH98A)"355-
356",METHANOPLANUS_LIMICOLA),METHANOGENIUM_CARIACI,METHANO
MICROBIUM_MOBILE,METHANOLACINIA_PAYNTERI,(METHANOFOLLIS_ET
HANOLICUS,METHANOFOLLIS_LIMINATANS)"150"),(METHANOSPHAERULA
_PALUSTRIS,(METHANOREGULA_BOONEI,METHANOREGULA_FORMICICA)"
45-89",METHANOLINEA_TARDA)"231-259")))"917-
919",((CANDIDATUS_METHANOPEREDENS_NITROREDUCENS,(METHANOSA
ETA_CONCILII,METHANOSAETA_HARUNDINACEA)"129",METHERMICOCCU
S_SHENGLIENSIS,(((METHANOSARCINA_MAZEI,METHANOSARCINA_ACETI
VORANS)"40",METHANOSARCINA_BARKERI,METHANOSARCINA_HORONOB
ENSIS,METHANOSARCINA_VACUOLATA,METHANOSARCINA_SICILIAE,MET
HANOSARCINA_LACUSTRIS,METHANOSARCINA_SOLIGELIDI,METHANOSA
RCINA_SP._2.H.T.1A.6,METHANOSARCINA_SP._2.H.T.1A.8,METHANOSARCIN
A_SP._1.H.A.2.2,METHANOSARCINA_SP._2.H.T.1A.15,METHANOSARCINA_SP.
_2.H.T.1A.3,METHANOSARCINA_SP._WWM596,METHANOSARCINA_SP._2.H.A
.1B.4,METHANOSARCINA_FLAVESCENS,METHANOSARCINA_SP._WH1,METH
ANOSARCINA_SP._MTP4,METHANOSARCINA_SP._KOLKSEE,METHANOSARC
INA_SP._A14,METHANOSARCINA_SP._1.H.T.1A.1,METHANOSARCINA_THER
MOPHILA)"670-
699",((((METHANOCOCCOIDES_METHYLUTENS,METHANOCOCCOIDES_BUR
TONII,METHANOCOCCOIDES_VULCANI),(METHANOHALOPHILUS_HALOPHI
LUS,METHANOHALOPHILUS_PORTUCALENSIS,METHANOHALOPHILUS_MA
HII))"336-
348",METHANOMETHYLOVORANS_HOLLANDICA),(METHANOHALOBIUM_E
VESTIGATUM,METHANOSALSUM_ZHILINAE)),(METHANOLOBUS_VULCANI,
METHANOLOBUS_PROFUNDI,METHANOLOBUS_TINDARIUS)"171-222")"733-
745"),METHANOCELLA_CONRADII))"920-
923",((GEOGLOBUS_ACETIVORANS,GEOGLOBUS_AHANGARI),FERROGLOBU
S_PLACIDUS,(ARCHAEOGLOBUS_VENEFICUS,ARCHAEOGLOBUS_FULGIDUS
,ARCHAEOGLOBUS_PROFUNDUS,ARCHAEOGLOBUS_SULFATICALLIDUS)))"
924-925"));

Supplementary File 3, Chapter 3: Bacteria: See description for Supplementary File 1. This file is available upon request. However, it is approximately 130 pages long, so it is excluded from this dissertation.

Supplementary File 4, Chapter 3: Fungi: See description for Supplementary File 1
((((((((((((((((((NANNIZZIA_GYPSEA,ARTHRODERMA_BENHAMIAE),MICROSPORUM_CANIS,(TRICHOPHYTON_VERRUCOSUM,TRICHOPHYTON_RUBRUM)),((PARACOCCIDIOIDES_LUTZII,PARACOCCIDIOIDES_BRASILIENSIS),HISTOPLASMA_CAPSULATUM,BLASTOMYCES_GILCHRISTII),((COCCIDIOIDES_POSADASII,COCCIDIOIDES_IMMITIS),UNCINOCARPUS_REESII)),(((CAPRONIA_EPIMYCES,CAPRONIA_CORONATA),(FONSECAEA_PEDROSOI,FONSECAEA_MULTIMORPHOSA,FONSECAEA_ERECTA),(EXOPHIALA_SPINIFERA,EXOPHIALA_XENOBIOTICA,EXOPHIALA_MESOPHILA,EXOPHIALA_OLIGOSPERMA,EXOPHIALA_AQUAMARINA,EXOPHIALA_DERMATITIDIS),PHIALOPHORA_ATTAE,RHINOCLADIELLA_MACKENZIEI,CONIOSPORIUM_APOLLINIS,(CLADOPHIALOPHORA_CARRIONII,CLADOPHIALOPHORA_BANTIANA,CLADOPHIALOPHORA_YEGRESII,CLADOPHIALOPHORA_IMMUNDA,CLADOPHIALOPHORA_PSAMMOPHILA)),CYPHELLOPHORA_EUROPAEA),((ASPERGILLUS_TERREUS,ASPERGILLUS_FUMIGATUS,ASPERGILLUS_FLAVUS,ASPERGILLUS_NOMIUS,ASPERGILLUS_NIGER,ASPERGILLUS_FISCHERI,ASPERGILLUS_ACULEATUS,ASPERGILLUS_CLAVATUS),(TALAROMYCES_MARNEFFEI,TALAROMYCES_STIPITATUS,TALAROMYCES_ATROROSEUS),(PENICILLIUM_EXPANSUM,PENICILLIUM_RUBENS,PENICILLIUM_DIGITATUM),RASAMSONIA_EMERSONII),ENDOCARPON_PUSILLUM),((METARHIZIUM_ROBERTSII,METARHIZIUM_ANISOPLIAE),METARHIZIUM_MAJUS)),METACORDYCEPS_CHLAMYDOSPORIA),PURPUREOCILLIUM_LILACINUM),((((((PODOSPORA_PAUCISETA,(THIELAVIA_TERRESTRIS,(CHAETOMIUM_GLOBOSUM,CHAETOMIUM_THERMOPHILUM)),(NEUROSPORA_TETRASPERMA,NEUROSPORA_CRASSA)),(GROSMANNIA_CLAVIGERA,SPOROTHRIX_SCHENCKII)),PHAEOACREMONIUM_ALEOPHILUM),(TRICHODERMA_GAMSII,(PESTALOTIOPSIS_FICI,EUTYPA_LATA))),(TRICHODERMA_VIRENS,TRICHODERMA_REESEI)"9-28")"1625",(ISARIA_FUMOSOROSEA,(CORDYCEPS_MILITARIS,BEAUVERIA_BASSIANA))))"1928",GIBBERELLA_ZEAE),(SCLEROTINIA_SCLEROTIORUM,(DREPANOPEZIZA_PUNCTIFORMIS,PHIALOCEPHALA_SCOPIFORMIS),GLAREA_LOZOYENSIS),(((ZYMOSEPTORIA_TRITICI,SPHAERULINA_MUSIVA,MYCOSPHAERELLA_FIJIENSIS),BAUDOINIA_PANAMERICANA),(DIPLODIA_CORTICOLA,NEOFUSICOCCUM_PARVUM),(PARACONIOTHYRIUM_SPORULOSUM,(((BIPOLARIS_MAYDIS,BIPOLARIS_SOROKINIANA,BIPOLARIS_VICTORIAE,BIPOLARIS_ORYZAE,BIPOLARIS_ZEICOLA),ALTERNARIA_ALTERNATA,PYRENOPHORA_TERES,SETOSPHAERIA_TURCICA),LEPTOSPHAERIA_MACULANS),PARASTAGONOSPORA_NODORUM),AUREOBASIDIUM_PULLULANS)"1657",TRICHODERMA_ATROVIRIDE,(COLLETOTRICHUM_FIORINIAE,COLLETOTRICHUM_GRAMINICOLA,COLLETOTRICHUM_GLOEOSPORIOIDES,COLLETOTRICHUM_HIGGINSIANUM),PSEUDALLESCHERIA_BOYDII,MAGNAPORTHE_ORYZAE,XYLONA_HEVEAE,VERTICILLIUM_ALFALFAE,VERTICILLIUM_DAHLIAE,GIBBERELLA_CORONICOLA,GIBBERELLA_FUJIKUROI,FUSARIUM_OXYSPORUM)"1929-

1982",TUBER_MELANOSPORUM,(ARTHROBOTRYS_OLIGOSPORA,DACTYLELLINA_
HAPTOTYLA)),(SUGIYAMAELLA_LIGNOHABITANS,(KAZACHSTANIA_AFRICANA,(
SACCHAROMYCES_CEREVISIAE,SACCHAROMYCES_EUBAYANUS),ZYGOSACCHA
ROMYCES_ROUXII,(NAUMOVOZYMA_CASTELLII,NAUMOVOZYMA_DAIRENENSIS)
,(TETRAPISISPORA_BLATTAE,TETRAPISISPORA_PHAFFII),KLUYVEROMYCES_LAC
TIS,VANDERWALTOZYMA_POLYSPORA,TORULASPORA_DELBRUECKII,LACHANC
EA_THERMOTOLERANS,(EREMOTHECIUM_GOSSYPII,EREMOTHECIUM_SINECAUD
UM,EREMOTHECIUM_CYMBALARIAE))"1626-
1630",(MEYEROZYMA_GUILLIERMONDII,SPATHASPORA_PASSALIDARUM,SCHEFF
ERSOMYCES_STIPITIS,(LODDEROMYCES_ELONGISPORUS,(CANDIDA_ALBICANS,C
ANDIDA_DUBLINIENSIS,CANDIDA_ORTHOPSILOSIS,CANDIDA_TROPICALIS)"29-
313"),HYPHOPICHIA_BURTONII,(DEBARYOMYCES_HANSENII,DEBARYOMYCES_FA
BRYI),BABJEVIELLA_INOSITOVORA)"321-
1624",((PICHIA_KUDRIAVZEVII,PICHIA_MEMBRANIFACIENS),(OGATAEA_POLYMO
RPHA,OGATAEA_PARAPOLYMORPHA)),(METSCHNIKOWIA_BICUSPIDATA,CLAVIS
PORA_LUSITANIAE),(KOMAGATAELLA_PHAFFII,(WICKERHAMOMYCES_ANOMAL
US,WICKERHAMOMYCES_CIFERRII),CYBERLINDNERA_JADINII),YARROWIA_LIPO
LYTICA,ASCOIDEA_RUBESCENS)"1658-1927")"1983-
1994",(SAITOELLA_COMPLICATA,((SCHIZOSACCHAROMYCES_POMBE,SCHIZOSAC
CHAROMYCES_OCTOSPORUS,SCHIZOSACCHAROMYCES_CRYOPHILUS),HASEGA
WAEA_JAPONICA),(PNEUMOCYSTIS_CARINII,PNEUMOCYSTIS_MURINA,PNEUMOC
YSTIS_JIROVECII)))"1995-
2017",(((((((((TRAMETES_VERSICOLOR,DICHOMITUS_SQUALENS)"1-
2",(FIBROPORIA_RADICULOSA,OLIGOPORUS_PLACENTUS),PHANEROCHAETE_CA
RNOSA),SERPULA_LACRYMANS),((GLOEOPHYLLUM_TRABEUM,PUNCTULARIA_S
TRIGOSOZONATA)"3-
8",((((COPRINOPSIS_CINEREA,LACCARIA_BICOLOR),SCHIZOPHYLLUM_COMMUNE,
AGARICUS_BISPORUS,MONILIOPHTHORA_RORERI)"314-
316",CONIOPHORA_PUTEANA),(STEREUM_HIRSUTUM,HETEROBASIDION_IRREGUL
ARE)))"317-320")"1631-
1656",FOMITIPORIA_MEDITERRANEA),AURICULARIA_SUBGLABRA),(KOCKOVAEL
LA_IMPERATAE,((CRYPTOCOCCUS_PINUS,CRYPTOCOCCUS_AMYLOLENTUS,CRYP
TOCOCCUS_BESTIOLAE,CRYPTOCOCCUS_DEJECTICOLA,CRYPTOCOCCUS_NEOFO
RMANS,CRYPTOCOCCUS_GATTII_VGI),TREMELLA_MESENTERICA,TSUCHIYAEA_
WINGFIELDII,KWONIELLA_MANGROVENSIS),TRICHOSPORON_ASAHII),(WALLEMI
A_MELLICOLA,WALLEMIA_ICHTHYOPHAGA)),((MALASSEZIA_GLOBOSA,MALASS
EZIA_SYMPODIALIS,MALASSEZIA_PACHYDERMATIS),((PSEUDOZYMA_HUBEIENSI
S,PSEUDOZYMA_ANTARCTICA,PSEUDOZYMA_FLOCCULOSA,KALMANOZYMA_BR
ASILIENSIS),USTILAGO_MAYDIS),TILLETIARIA_ANOMALA)),(MIXIA_OSMUNDAE,(
RHODOTORULA_GRAMINIS,RHODOTORULA_TORULOIDES),PUCCINIA_GRAMINIS)
))"2018-
2019",(LOBOSPORANGIUM_TRANSVERSALE,PHYCOMYCES_BLAKESLEEANUS)),SPI
ZELLOMYCES_PUNCTATUS),(((ENCEPHALITOZOON_ROMALEAE,ENCEPHALITOZO

ON_INTESTINALIS,ENCEPHALITOZOON_HELLEM,ENCEPHALITOZOON_CUNICULI),(NOSEMA_CERANAE,VITTAFORMA_CORNEAE),VAVRAIA_CULICIS),ENTEROCYTOZOON_BIENEUSI)),(MARINOMONAS_FUNGIAE,COLLIMONAS_FUNGIVORANS));

Supplementary File 5, Chapter 3: Invertebrates: See description for Supplementary File 1

(((((((((((((((((((((((((((CYPHOMYRMEX_COSTATUS,((ACROMYRMEX_ECHINATIOR,(TRACHYMYRMEX_SEPTENTRIONALIS,TRACHYMYRMEX_CORNETZI,TRACHYMYRMEX_ZETEKI)),(ATTA_CEPHALOTES,ATTA_COLOMBICA))),WASMANNIA_AUROPUNCTATA),(MONOMORIUM_PHARAONIS,SOLENOPSIS_INVICTA)),VOLLENHOVIA_EMERYI),POGONOMYRMEX_BARBATUS),CAMPONOTUS_FLORIDANUS),(PSEUDOMYRMEX_GRACILIS,LINEPITHEMA_HUMILE)),(HARPEGNATHOS_SALTATOR,DINOPONERA_QUADRICEPS)),(((((((APIS_MELLIFERA,((APIS_DORSATA,APIS_FLOREA),APIS_CERANA))"4",(BOMBUS_TERRESTRIS,BOMBUS_IMPATIENS)),EUFRIESEA_MEXICANA),CERATINA_CALCARATA),HABROPODA_LABORIOSA),MEGACHILE_ROTUNDATA),DUFOUREA_NOVAEANGLIAE),(POLISTES_CANADENSIS,POLISTES_DOMINULA)),((CERATOSOLEN_SOLMSI,COPIDOSOMA_FLORIDANUM,TRICHOGRAMMA_PRETIOSUM,NASONIA_VITRIPENNIS),((DIACHASMA_ALLOEUM,FOPIUS_ARISANUS),MICROPLITIS_DEMOLITOR)))"20-56",ORUSSUS_ABIETINUS),CEPHUS_CINCTUS),(ATHALIA_ROSAE,NEODIPRION_LECONTEI)),((((((HELICOVERPA_ARMIGERA,BOMBYX_MORI),AMYELOIS_TRANSITELLA),((PAPILIO_POLYTES,PAPILIO_MACHAON,PAPILIO_XUTHUS),PIERIS_RAPAE)),PLUTELLA_XYLOSTELLA),(((((((DROSOPHILA_OBSCURA,(DROSOPHILA_PERSIMILIS,DROSOPHILA_MIRANDA)),(((DROSOPHILA_SUZUKII,(((DROSOPHILA_EUGRACILIS,(DROSOPHILA_BIARMIPES,DROSOPHILA_TAKAHASHII)),((DROSOPHILA_YAKUBA,DROSOPHILA_ERECTA),(DROSOPHILA_MELANOGASTER,(DROSOPHILA_SIMULANS,DROSOPHILA_SECHELLIA)))),DROSOPHILA_FICUSPHILA),(DROSOPHILA_RHOPALOA,DROSOPHILA_ELEGANS)),(DROSOPHILA_KIKKAWAI,DROSOPHILA_SERRATA)),(DROSOPHILA_BIPECTINATA,DROSOPHILA_ANANASSAE))),DROSOPHILA_WILLISTONI),(DROSOPHILA_BUSCKII,(DROSOPHILA_GRIMSHAWI,(DROSOPHILA_VIRILIS,(DROSOPHILA_ARIZONAE,DROSOPHILA_NAVOJOA,DROSOPHILA_MOJAVENSIS)))),(MUSCA_DOMESTICA,STOMOXYS_CALCITRANS))"5-16",((CERATITIS_CAPITATA,((BACTROCERA_LATIFRONS,BACTROCERA_DORSALIS),BACTROCERA_OLEAE,BACTROCERA_CUCURBITAE)),RHAGOLETIS_ZEPHYRIA))"17-19",((CULEX_QUINQUEFASCIATUS,(AEDES_ALBOPICTUS,AEDES_AEGYPTI)),ANOPHELES_GAMBIAE))"57-68")"69-114",((((AETHINA_TUMIDA,(ANOPLOPHORA_GLABRIPENNIS,DENDROCTONUS_PONDEROSAE)),TRIBOLIUM_CASTANEUM),AGRILUS_PLANIPENNIS),NICROPHORUS_VESPILLOIDES)))"115-122",(((NILAPARVATA_LUGENS,(HALYOMORPHA_HALYS,CIMEX_LECTULARIUS)),((DIURAPHIS_NOXIA,ACYRTHOSIPHON_PISUM,MYZUS_PERSICAE)"3",(BEMISIA_TABACI,DIAPHORINA_CITRI))),PEDICULUS_HUMANUS))"123",ZOOTERMOPSIS_NEVADENSIS),FOLSOMIA_CANDIDA),HYALELLA_AZTECA),((PARASTEATODA_TEPIDARIORUM,(TETRANYCHUS_URTICAE,(METASEIULUS_OCCIDENTALIS,IXODES_SCAPULARIS))),LIMULUS_POLYPHEMUS)),(((((CAENORHABDITIS_REMANEI,CAENORHABDITIS_BRIGGSAE,CAENORHABDITIS_ELEGANS),NECATOR_AMERICANUS),(BRUGIA_MALAYI,LOA_LOA)),TRICHINELLA_SPIRALIS),PRIAPULUS_CAUDATUS)),((HELOBDELLA_ROBUSTA,((OC

TOPUS_BIMACULOIDES,(((APLYSIA_CALIFORNICA,BIOMPHALARIA_GLABRATA),LOTTIA_GIGANTEA),(MIZUHOPECTEN_YESSOENSIS,CRASSOSTREA_GIGAS))),LINGULA_ANATINA)),(OPISTHORCHIS_VIVERRINI,(SCHISTOSOMA_HAEMATOBIUM,SCHISTOSOMA_MANSONI)))),(((ACANTHASTER_PLANCI,STRONGYLOCENTROTUS_PURPURATUS)"1",SACCOGLOSSUS_KOWALEVSKII),(CIONA_INTESTINALIS,(BRANCHIOSTOMA_FLORIDAE,BRANCHIOSTOMA_BELCHERI)))),(HYDRA_VULGARIS,((AIPTASIA_PALLIDA,NEMATOSTELLA_VECTENSIS),(ORBICELLA_FAVEOLATA,ACROPORA_DIGITIFERA)"2")))"124",TRICHOPLAX_ADHAERENS),AMPHIMEDON_QUEENSLANDICA),(MONOSIGA_BREVICOLLIS,SALPINGOECA_ROSETTA)),(SPHAEROFORMA_ARCTICA,CAPSASPORA_OWCZARZAKI)),FONTICULA_ALBA);

Supplementary File 6, Chapter 3: Plants: See description for Supplementary File 1
((((((((((((((((VIGNA_RADIATA,PHASEOLUS_VULGARIS,VIGNA_ANGULARIS)"820-822",GLYCINE_MAX)"825-828",CAJANUS_CAJAN),(CICER_ARIETINUM,MEDICAGO_TRUNCATULA)"634-637")"910-911",(ARACHIS_IPAENSIS,ARACHIS_DURANENSIS)),LUPINUS_ANGUSTIFOLIUS),((((CUCUMIS_SATIVUS,CUCUMIS_MELO)"627-633",MOMORDICA_CHARANTIA),JUGLANS_REGIA),((ZIZIPHUS_JUJUBA,MORUS_NOTABILIS),(((PRUNUS_PERSICA,PRUNUS_AVIUM,PRUNUS_MUME)"817-819",MALUS_DOMESTICA),FRAGARIA_VESCA)"901"))),((RICINUS_COMMUNIS,JATROPHA_CURCAS)"399-400",(POPULUS_EUPHRATICA,POPULUS_TRICHOCARPA)"144-145")),(((((THEOBROMA_CACAO,HERRANIA_UMBRATICA),((GOSSYPIUM_ARBOREUM,GOSSYPIUM_HIRSUTUM),GOSSYPIUM_RAIMONDII)"823-824"),(((((BRASSICA_NAPUS,BRASSICA_RAPA,BRASSICA_OLERACEA)"659-816",EUTREMA_SALSUGINEUM),RAPHANUS_SATIVUS)"902-909",((ARABIDOPSIS_LYRATA,ARABIDOPSIS_THALIANA)"1-141",(CAMELINA_SATIVA,CAPSELLA_RUBELLA)))"1180-1214",TARENAYA_HASSLERIANA))"1254-1255",(CITRUS_SINENSIS,CITRUS_CLEMENTINA)),EUCALYPTUS_GRANDIS)"1256-1268")"1269-1302",VITIS_VINIFERA)"1303-1315",((((((CAPSICUM_ANNUUM,((SOLANUM_LYCOPERSICUM,SOLANUM_PENNELLII),SOLANUM_TUBEROSUM)"650-658"),((NICOTIANA_ATTENUATA,NICOTIANA_SYLVESTRIS),(NICOTIANA_TOMENTOSIFORMIS,NICOTIANA_TABACUM)))"919-982",IPOMOEA_NIL),(ERYTHRANTHE_GUTTATA,SESAMUM_INDICUM))"1224-1232",DAUCUS_CAROTA)"1233-1235",BETA_VULGARIS)"1236-1252")"1316-1461",NELUMBO_NUCIFERA)"1462",(((DENDROBIUM_CATENATUM,PHALAENOPSIS_EQUESTRIS)"648-649",ASPARAGUS_OFFICINALIS),((((((SORGHUM_BICOLOR,ZEA_MAYS)"143",SETARIA_ITALICA),(ORYZA_BRACHYANTHA,ORYZA_SATIVA)),(AEGILOPS_TAUSCHII,BRACHYPODIUM_DISTACHYON)"142")"912-918",ANANAS_COMOSUS),((PHOENIX_DACTYLIFERA,ELAEIS_GUINEENSIS)"401-402",MUSA_ACUMINATA))"1215-1223")"1253")"1463-1600",AMBORELLA_TRICHOPODA)"1601-1627",SELAGINELLA_MOELLENDORFFII)"1628-1629",PHYSCOMITRELLA_PATENS)"1630-1702",(((BATHYCOCCUS_PRASINOS,OSTREOCOCCUS_LUCIMARINUS)"638-647",(MICROMONAS_PUSILLA,MICROMONAS_COMMODA)"403-626")"829-900",(((CHLAMYDOMONAS_REINHARDTII,VOLVOX_CARTERI)"146-398",MONORAPHIDIUM_NEGLECTUM),(CHLORELLA_VARIABILIS,AUXENOCHLORELLA_PROTOTHECOIDES)))"983-1179"),((GALDIERIA_SULPHURARIA,CYANIDIOSCHYZON_MEROLAE),CHONDRUS_CRISPUS));

Supplementary File 7, Chapter 3: Protozoa: See description for Supplementary File 1

((((((ACYTOSTELIUM_SUBGLOBOSUM,POLYSPHONDYLIUM_PALLIDUM),(DICTYOSTELIUM_PURPUREUM,DICTYOSTELIUM_DISCOIDEUM,DICTYOSTELIUM_FASCICULATUM)"91-1076")"1609-2288",(ENTAMOEBA_NUTTALLI,(ENTAMOEBA_HISTOLYTICA,ENTAMOEBA_DISPAR),ENTAMOEBA_INVADENS)),ACANTHAMOEBA_CASTELLANII),(((((TRYPANOSOMA_GRAYI,TRYPANOSOMA_BRUCEI,TRYPANOSOMA_CRUZI),LEPTOMONAS_PYRRHOCORIS,(((LEISHMANIA_INFANTUM,LEISHMANIA_DONOVANI),LEISHMANIA_MAJOR,LEISHMANIA_MEXICANA)"1077-1608",(LEISHMANIA_BRAZILIENSIS,LEISHMANIA_PANAMENSIS)"1988")"2289-2319"),(TRICHOMONAS_VAGINALIS,NAEGLERIA_GRUBERI)),GIARDIA_INTESTINALIS),(((((PARAMECIUM_TETRAURELIA,(ICHTHYOPHTHIRIUS_MULTIFILIIS,TETRAHYMENA_THERMOPHILA)),(((((BABESIA_BOVIS,BABESIA_BIGEMINA,THEILERIA_EQUI,BABESIA_MICROTI),(THEILERIA_ORIENTALIS,THEILERIA_ANNULATA,THEILERIA_PARVA))"2320-2321",((((PLASMODIUM_BERGHEI,PLASMODIUM_YOELII),PLASMODIUM_CHABAUDI,PLASMODIUM_VINCKEI),(PLASMODIUM_KNOWLESI,PLASMODIUM_CYNOMOLGI)),(PLASMODIUM_GABONI,PLASMODIUM_REICHENOWI,PLASMODIUM_FALCIPARUM),PLASMODIUM_INUI,PLASMODIUM_FRAGILE,PLASMODIUM_COATNEYI))"2333-2405",(((EIMERIA_NECATRIX,EIMERIA_TENELLA,EIMERIA_MITIS,EIMERIA_MAXIMA,EIMERIA_ACERVULINA),TOXOPLASMA_GONDII,HAMMONDIA_HAMMONDI,NEOSPORA_CANINUM)"2322-2326",GREGARINA_NIPHANDRODES,(CRYPTOSPORIDIUM_MURIS,CRYPTOSPORIDIUM_PARVUM,CRYPTOSPORIDIUM_HOMINIS)))"2406-2417",PERKINSUS_MARINUS)),(((BLASTOCYSTIS_HOMINIS,BLASTOCYSTIS_SP._SUBTYPE_4),(((THALASSIOSIRA_PSEUDONANA,PHAEODACTYLUM_TRICORNUTUM)"1-18",(AUREOCOCCUS_ANOPHAGEFFERENS,NANNOCHLOROPSIS_GADITANA)"89-90"),(PHYTOPHTHORA_PARASITICA,(SAPROLEGNIA_PARASITICA,SAPROLEGNIA_DICLINA),(APHANOMYCES_ASTACI,APHANOMYCES_INVADANS),PHYTOPHTHORA_SOJAE,PHYTOPHTHORA_INFESTANS))),BIGELOWIELLA_NATANS)"2327-2332")"2418-2432",EMILIANIA_HUXLEYI)"2433-2448",GUILLARDIA_THETA)"2449")),THECAMONAS_TRAHENS);

Supplementary File 8, Chapter 3: Mammals: See description for Supplementary File 1
((((((((((((HOMO_SAPIENS,PAN_PANISCUS)"304-308",GORILLA_GORILLA)"3924-4559",PONGO_ABELII)"5883-5891",NOMASCUS_LEUCOGENYS)"7309-7313",((CHLOROCEBUS_SABAEUS,((MACACA_NEMESTRINA,(MACACA_FASCICULARIS,MACACA_MULATTA))"3652",((CERCOCEBUS_ATYS,MANDRILLUS_LEUCOPHAEUS)"259-303",PAPIO_ANUBIS)"5302-5349")"7355-7438")"7491-7501",((RHINOPITHECUS_BIETI,RHINOPITHECUS_ROXELLANA)"590-701",COLOBUS_ANGOLENSIS)"4775-4840")"7859-8006")"8194-8537",((CALLITHRIX_JACCHUS,AOTUS_NANCYMAAE)"176-207",(SAIMIRI_BOLIVIENSIS,CEBUS_CAPUCINUS)"2464-2509")"6563-6661")"8615-9045",CARLITO_SYRICHTA)"9046-9063",((PROPITHECUS_COQUERELI,MICROCEBUS_MURINUS)"702-794",OTOLEMUR_GARNETTII)"4893-4923")"9138-9194",GALEOPTERUS_VARIEGATUS)"9197-9218",(((((((((MERIONES_UNGUICULATUS,(RATTUS_NORVEGICUS,(MUS_PAHARI,(MUS_CAROLI,MUS_MUSCULUS)"1231-1496")"3333-3626")"6093-6562")"7183-7248",((MICROTUS_OCHROGASTER,(MESOCRICETUS_AURATUS,CRICETULUS_GRISEUS)"2866-2985")"3174-3184",PEROMYSCUS_MANICULATUS)"6856-6902")"7650-7715",NANNOSPALAX_GALILI)"7795-7858",JACULUS_JACULUS)"8007-8018",(CASTOR_CANADENSIS,DIPODOMYS_ORDII)"309-345")"8096-8123",((CAVIA_PORCELLUS,(OCTODON_DEGUS,CHINCHILLA_LANIGERA)"1560-1632")"4924-4971",(HETEROCEPHALUS_GLABER,FUKOMYS_DAMARENSIS)"1-143")"6960-7031")"8565-8590",(MARMOTA_MARMOTA,ICTIDOMYS_TRIDECEMLINEATUS)"1633-1921")"9064-9110",(OCHOTONA_PRINCEPS,ORYCTOLAGUS_CUNICULUS)"346-476")"9111-9134",TUPAIA_CHINENSIS)"9135-9137")"9276-9461",(((((((PTEROPUS_VAMPYRUS,PTEROPUS_ALECTO)"2986-3171",ROUSETTUS_AEGYPTIACUS)"3185-3332",(RHINOLOPHUS_SINICUS,HIPPOSIDEROS_ARMIGER)"2744-2865")"7249-7260",(((MYOTIS_BRANDTII,MYOTIS_DAVIDII)"1922-2297",EPTESICUS_FUSCUS)"4972-5301",MINIOPTERUS_NATALENSIS)"5790-5882")"7636-7649",(((((((LIPOTES_VEXILLIFER,(ORCINUS_ORCA,TURSIOPS_TRUNCATUS)"1497-1559")"5350-5383",PHYSETER_CATODON)"6903-6959",BALAENOPTERA_ACUTOROSTRATA)"7116-7182",(((BISON_BISON,(BOS_INDICUS,BOS_MUTUS,BOS_TAURUS)"3627-3651",BUBALUS_BUBALIS)"7032-7115",PANTHOLOPS_HODGSONII,(CAPRA_HIRCUS,OVIS_ARIES)"477-589")"7508-7635",ODOCOILEUS_VIRGINIANUS)"7716-7794")"8124-8193",SUS_SCROFA)"8538-8564",((CAMELUS_FERUS,CAMELUS_BACTRIANUS,CAMELUS_DROMEDARIUS)"3653-3787",VICUGNA_PACOS)"5892-6079")"8591-8614",(CERATOTHERIUM_SIMUM,(EQUUS_ASINUS,(EQUUS_PRZEWALSKII,EQUUS_CABALLUS)"3172-3173")"4560-4774"))"9195-9196")"9219-

9234",((((((URSUS_MARITIMUS,AILUROPODA_MELANOLEUCA)"2329-
2400",(ODOBENUS_ROSMARUS,(NEOMONACHUS_SCHAUINSLANDI,LEPTONYCHOT
ES_WEDDELLII)"2401-2463")"4841-4892")"7261-7308",MUSTELA_PUTORIUS)"7314-
7329",CANIS_LUPUS)"7439-7490",((FELIS_CATUS,ACINONYX_JUBATUS)"144-
175",(PANTHERA_PARDUS,PANTHERA_TIGRIS)"208-258")"6662-6855")"8019-
8085",MANIS_JAVANICA)"8086-8095")"9235-
9275",((CONDYLURA_CRISTATA,SOREX_ARANEUS)"2510-
2743",ERINACEUS_EUROPAEUS)"3788-3902")"9462-9465")"9466-
9539",(((TRICHECHUS_MANATUS,LOXODONTA_AFRICANA),((ELEPHANTULUS_ED
WARDII,(ECHINOPS_TELFAIRI,CHRYSOCHLORIS_ASIATICA)"2298-2328")"3903-
3923",ORYCTEROPUS_AFER)"6080-6092")"7330-
7354",DASYPUS_NOVEMCINCTUS)"7502-7507")"9540-
10029",(MONODELPHIS_DOMESTICA,(SARCOPHILUS_HARRISII,PHASCOLARCTOS_
CINEREUS)"795-1230")"5384-5789"),ORNITHORHYNCHUS_ANATINUS);

Supplementary File 9, Chapter 3: Other Vertebrates: See description for Supplementary File 1

((((((((((((((((POECILIA_RETICULATA,POECILIA_FORMOSA,POECILIA_LATIPINNA,POECILIA_MEXICANA)"6928-7352",XIPHOPHORUS_MACULATUS)"7883-8180",(CYPRINODON_VARIEGATUS,FUNDULUS_HETEROCLITUS)"2997-3090")"8469-8581",(NOTHOBRANCHIUS_FURZERI,(AUSTROFUNDULUS_LIMNAEUS,KRYPTOLEBIAS_MARMORATUS)"2292-2473")"5535-5540")"9058-9110",ORYZIAS_LATIPES)"9111-9126",(((HAPLOCHROMIS_BURTONI,(PUNDAMILIA_NYEREREI,MAYLANDIA_ZEBRA)"1974-2029")"5632-5805",NEOLAMPROLOGUS_BRICHARDI),OREOCHROMIS_NILOTICUS)"7667-7882")"9580-9627",((PARALICHTHYS_OLIVACEUS,CYNOGLOSSUS_SEMILAEVIS)"13-55",MONOPTERUS_ALBUS)"5541-5563")"9689-9730",((LABRUS_BERGYLTA,TAKIFUGU_RUBRIPES)"187-192",NOTOTHENIA_CORIICEPS)"5312-5353")"9731-9788",HIPPOCAMPUS_COMES)"9789-9806",LARIMICHTHYS_CROCEA)"9807-9826",BOLEOPHTHALMUS_PECTINIROSTRIS)"9827-9870",(((ONCORHYNCHUS_KISUTCH,ONCORHYNCHUS_MYKISS)"2474-2672",SALMO_SALAR),ESOX_LUCIUS)"6568-6675")"9962-9993",(((DANIO_RERIO,((SINOCYCLOCHEILUS_RHINOCEROUS,SINOCYCLOCHEILUS_ANSHUIENSIS,SINOCYCLOCHEILUS_GRAHAMI)"5354-5534",CYPRINUS_CARPIO)"6837-6927")"7573-7666",(ICTALURUS_PUNCTATUS,PYGOCENTRUS_NATTERERI)"3688-3831")"8582-8652",CLUPEA_HARENGUS)"8655-8701")"9996-10158",SCLEROPAGES_FORMOSUS)"10159-10216",LEPISOSTEUS_OCULATUS),((((((((((((((((((CORVUS_CORNIX,CORVUS_BRACHYRHYNCHOS)"1742-1973",(((((GEOSPIZA_FORTIS,ZONOTRICHIA_ALBICOLLIS)"56-116",SERINUS_CANARIA)"5261-5311",(LONCHURA_STRIATA,TAENIOPYGIA_GUTTATA)"3669-3687")"7353-7439",(FICEDULA_ALBICOLLIS,STURNUS_VULGARIS)"410-485")"8381-8468",(PARUS_MAJOR,PSEUDOPODOCES_HUMILIS)"2137-2291")"8702-9048")"9127-9289",(MANACUS_VITELLINUS,LEPIDOTHRIX_CORONATA)"2673-2996")"9311-9554",ACANTHISITTA_CHLORIS)"9555-9579",(NESTOR_NOTABILIS,MELOPSITTACUS_UNDULATUS))"9628-9645",(FALCO_PEREGRINUS,FALCO_CHERRUG)"1473-1694")"9646-9685",CARIAMA_CRISTATA)"9686-9688",(((((BUCEROS_RHINOCEROS,(MEROPS_NUBICUS,PICOIDES_PUBESCENS)"2096-2122")"4445-4459",APALODERMA_VITTATUM)"5817-5827",LEPTOSOMUS_DISCOLOR)"7440-7443",COLIUS_STRIATUS)"8181",TYTO_ALBA)"8362-8366")"9871-9911",((HALIAEETUS_ALBICILLA,HALIAEETUS_LEUCOCEPHALUS)"1695-1741",AQUILA_CHRYSAETOS)"4748-4933")"9912-9961",OPISTHOCOMUS_HOAZIN)"9994-9995",(((((((EGRETTA_GARZETTA,PELECANUS_CRISPUS)"860-

866",NIPPONIA_NIPPON)"5241-5260",PHALACROCORAX_CARBO)"5806-
5816",(FULMARUS_GLACIALIS,(APTENODYTES_FORSTERI,PYGOSCELIS_ADELIAE)
"193-233")"5237-5240")"8367-8380",GAVIA_STELLATA)"8653-
8654",(PHAETHON_LEPTURUS,EURYPYGA_HELIAS)"177-186")"9049-
9057",(CALIDRIS_PUGNAX,CHARADRIUS_VOCIFERUS)"2128-2136")"9290-
9310")"10217-10379",BALEARICA_REGULORUM)"10380-
10385",((TAURACO_ERYTHROLOPHUS,(CHLAMYDOTIS_MACQUEENII,CUCULUS_C
ANORUS)"2123-
2127")"4747",((PTEROCLES_GUTTURALIS,MESITORNIS_UNICOLOR)"396-
409",COLUMBA_LIVIA)"5231-5236")"8332-8361")"10386-
10452",((CALYPTE_ANNA,CHAETURA_PELAGICA)"1160-
1200",CAPRIMULGUS_CAROLINENSIS))"10453-
10499",((((GALLUS_GALLUS,MELEAGRIS_GALLOPAVO)"234-
238",COTURNIX_JAPONICA)"4743-4746",NUMIDA_MELEAGRIS)"6676-
6836",(ANAS_PLATYRHYNCHOS,ANSER_CYGNOIDES)"3832-4020")"8279-
8331")"10500-10758",((APTERYX_AUSTRALIS,TINAMUS_GUTTATUS)"117-
176",STRUTHIO_CAMELUS)"5564-5631")"10759-
10861",((GAVIALIS_GANGETICUS,CROCODYLUS_POROSUS)"867-
1159",(ALLIGATOR_SINENSIS,ALLIGATOR_MISSISSIPPIENSIS)"3091-3668")"5828-
6567")"10862-11095",((CHRYSEMYS_PICTA,CHELONIA_MYDAS)"486-
859",PELODISCUS_SINENSIS)"4934-5230")"11096-
11484",(((PYTHON_BIVITTATUS,(THAMNOPHIS_SIRTALIS,PROTOBOTHROPS_MUCR
OSQUAMATUS)"1201-1472")"4021-
4444",(POGONA_VITTICEPS,ANOLIS_CAROLINENSIS)"239-395")"7444-
7572",GEKKO_JAPONICUS)"8182-8278")"11485-
11696",(NANORANA_PARKERI,(XENOPUS_TROPICALIS,XENOPUS_LAEVIS)"1-
12")"4460-4742")"11697-11789",LATIMERIA_CHALUMNAE)"11790-11822")"11823-
11877",(RHINCODON_TYPUS,CALLORHINCHUS_MILII)"2030-
2095"),LATES_CALCARIFER);

Supplementary File 10, Chapter 3: All Species: Accompanying character state change file for Supplementary File 1. The first 30 lines out of 25,771 lines are shown in this dissertation. The complete file is available upon request.

FPPS_ACA 1->0
FPPS_TTT 1->0
GPI14_CTA 1->0
EIF4G3_TTA 1->0
PGFS_CGA 1->0
PGFS_TGT 1->0
CYP4_AAA 1->0
CYP4_AAT 1->0
CYP4_TCC 1->0
CYP5_ACT 1->0
CYP5_CAA 1->0
CYP5_GGA 1->0
CYP2_AAA 1->0
CYP2_AGT 1->0
CYP2_CCC 1->0
CYP2_TTG 1->0
CYP2_GTT 1->0
CYP3_ATA 1->0
CYP3_AGA 1->0
CYP3_TTA 1->0
CYP9_AGG 1->0
CD_CAT 1->0
CD_GTA 1->0
ARL-1_AGA 1->0
ARL-1_CGG 1->0
ARL-1_GTT 1->0
CYP14_TAG 0->1
CYP10_CCC 1->0
CYP10_CCT 1->0
CYP11_CCT 1->0

Supplementary File 11, Chapter 3: Archaea: Accompanying character state change file for Supplementary File 2. The first 30 lines out of 925 lines are shown in this dissertation. The complete file is available upon request.

RPL34E_ATC 0->1
RPL34E_ATT 0->1
RPL34E_CAG 1->0
RPL34E_CTC 1->0
RPL34E_CGA 0->1
RPL34E_TAA 0->1
RPL34E_TTA 0->1
RPL34E_GCC 1->0
RPL34E_GTC 1->0
RPL34E_GGC 1->0
RPS19P_AAC 1->0
RPS19P_CAC 1->0
RPS19P_TAA 0->1
RPL39E_AAC 1->0
RPL39E_CAT 0->1
RPL39E_GCC 1->0
RPL37E_ACA 0->1
RPL37E_ATC 0->1
RPL37E_AGT 0->1
RPL37E_CAA 0->1
RPL37E_CAT 0->1
RPL37E_AAT 0->1
RPL37E_ACG 0->1
RPL37E_ATT 0->1
RPL37E_CGC 0->1
RPL37E_GCG 0->1
RPL37E_GTC 0->1
RPL37E_GTT 0->1
RPL34E_ACA 1->0
RPL34E_ACC 0->1

Supplementary File 12, Chapter 3: Bacteria: Accompanying character state change file for Supplementary File 3. The first 30 lines out of 6,639 lines are shown in this dissertation. The complete file is available upon request.

RCSA_TTT 1->0
SPAK_GTC 1->0
FKP_GCC 1->0
BCME_GCT 1->0
BCME_GTT 1->0
PLCR_AGA 0->1
AGUB_GGC 1->0
UDP_GGC 1->0
TORT_CTG 1->0
HYCD_GCC 1->0
PHNR_CAG 1->0
HYFE_GCC 1->0
QOXB_CCG 1->0
MYCP_TGG 1->0
MTR_TCG 1->0
CLPA_GAA 1->0
ACNB_GAA 1->0
FDHE_ATG 1->0
FDHE_GAA 1->0
NIFJ_GAA 1->0
PGAB_GAA 1->0
FDXH_GAA 1->0
BUK_GAA 1->0
FLHF_GAA 1->0
RHLP_AAA 1->0
RTCA_TTG 1->0
SOXC_GAA 1->0
GSPE_GAA 1->0
ACEF_GAA 1->0
PEPF_GAA 1->0

Supplementary File 13, Chapter 3: Fungi: Accompanying character state change file for Supplementary File 4. The first 30 lines out of 2,019 lines are shown in this dissertation. The complete file is available upon request.

HTP2_AGT 1->0
HTP2_CAA 1->0
HTP4_AAT 0->1
HTP4_ATT 0->1
HTP4_GTT 0->1
HTP2_ACT 1->0
HTP4_CGT 0->1
HTP4_TTT 0->1
BIP1_CAT 0->1
CPC2_CCT 0->1
CPC2_GTA 0->1
RHO3_CAT 1->0
RHO3_TAT 1->0
YPT1_AAT 1->0
YPT1_ACA 1->0
YPT1_AGA 1->0
YPT1_CAA 1->0
YPT1_CCC 0->1
YPT1_CGC 0->1
YPT1_TCA 1->0
YPT1_TTA 1->0
BIP1_GGG 0->1
CPC2_AAA 0->1
RAS2_AAA 1->0
RAS2_ACT 1->0
NAG1_AGA 1->0
NAG1_TGT 1->0
PDI1_CCA 1->0
COX9_CTC 1->0
ERG7_CGC 1->0

Supplementary File 14, Chapter 3: Invertebrates: Accompanying character state change file for Supplementary File 5. The first 30 lines out of 124 lines are shown in this dissertation. The complete file is available upon request.

ND4_GAT 1->0
ATP6_TGA 1->0
ATP6_GGA 1->0
ND5_GGG 1->0
SXL_CGG 0->1
SXL_TAG 0->1
SXL_GGG 0->1
PDF_CCA 0->1
PDF_CGC 0->1
PDF_AGA 0->1
PDF_AGC 0->1
PDF_CAC 0->1
PDF_CCT 0->1
PDF_CGT 0->1
PDF_GCA 0->1
PDF_GTT 0->1
RAB7_AGG 0->1
RAB7_GGG 0->1
RAB7_TAA 0->1
RPL41_GCT 1->0
RPL39_CAG 1->0
RPL39_CCG 0->1
RPL39_GTC 0->1
CDK5_GGG 1->0
UQCR11_CTC 1->0
PTEN_TAA 0->1
IMD_TAA 0->1
IMD_TGT 0->1
GSTD1_TCG 0->1
CPR16_TTA 0->1

Supplementary File 15, Chapter 3: Plants: Accompanying character state change file for Supplementary File 6. The first 30 lines out of 1,702 lines are shown in this dissertation. The complete file is available upon request.

AGD1_CCG 0->1
PSBT_TCC 0->1
WRKY36_CGC 0->1
WRKY36_CGG 0->1
WRKY36_TGA 0->1
WRKY20_CCG 0->1
WRKY20_CTC 0->1
WRKY36_TAA 0->1
WRKY29_TAG 0->1
SDP1_TAG 0->1
PAL3_ATA 0->1
TBP1_TAG 0->1
NMT1_CCC 0->1
NMT1_TGA 0->1
NIK1_CCC 0->1
APC8_CTA 0->1
TUB1_TAA 0->1
APC11_AAG 0->1
APC11_ATC 0->1
APC11_CCC 0->1
APC11_CCG 0->1
APC11_CTG 0->1
APC11_CGC 0->1
APC11_TCC 0->1
APC11_GGC 0->1
SBP1_TGA 0->1
SBP1_TGT 0->1
ISU1_AAT 0->1
ISU1_ACA 0->1
ISU1_TCA 0->1

Supplementary File 16, Chapter 3: Protozoa: Accompanying character state change file for Supplementary File 7. The first 30 lines out of 2,449 lines are shown in this dissertation. The complete file is available upon request.

CLPP_ACA 0->1
CLPP_CGG 0->1
PCNA_AGA 1->0
PCNA_CCG 0->1
PCNA_TCA 1->0
ACT1_CTA 0->1
ARF1_CGG 0->1
ARF1_GCA 0->1
CAM1_CTG 0->1
CAM1_CGC 0->1
CAM1_TCG 0->1
CAM1_TGC 0->1
CAM1_GTG 0->1
ARF1_TTA 0->1
SDH1_TCA 0->1
SDH1_TGA 0->1
CAM1_ACA 0->1
CHC_ATA 0->1
SIR2_AGA 1->0
SIR2_TAG 1->0
SIR2_TGA 0->1
RPA1_CCT 1->0
RPA1_TAT 1->0
PAH_AGG 0->1
PAH_TGT 1->0
CYSB_ACA 1->0
CYSB_TTG 1->0
ATG12_AAT 1->0
TAT_AGA 1->0
APRT_CAT 1->0

Supplementary File 17, Chapter 3: Mammals: Accompanying character state change file for Supplementary File 8. The first 30 lines out of 10,029 lines are shown in this dissertation. The complete file is available upon request.

PPT2_GCA 1->0
BRDT_CGG 1->0
SORD_GTA 1->0
MPP2_CTT 1->0
GRB7_TAT 1->0
CLEC4G_GAT 1->0
SLC27A3_ATT 1->0
DRAM1_AAA 1->0
PRSS38_TTG 1->0
LEPROT_CTT 1->0
MEIS3_ATT 1->0
KIAA1024L_CAA 1->0
HARBI1_ATA 1->0
PSMF1_AGG 1->0
PSMF1_TAT 1->0
ZNF692_CTT 1->0
CCND3_GCA 1->0
SOCS4_TTT 1->0
SOCS4_GCT 1->0
SOCS4_GTA 1->0
KRT35_TTG 1->0
KRT36_GAA 1->0
KLHL10_TAT 1->0
PDP1_CTA 1->0
GEM_GGA 1->0
DHX8_CTA 1->0
STBD1_TGC 1->0
FZD2_AAA 1->0
SBK3_TGT 1->0
ZNF398_AGG 1->0

Supplementary File 18, Chapter 3: Other Vertebrates: Accompanying character state change file for Supplementary File 9. The first 30 lines out of 11,877 lines are shown in this dissertation. The complete file is available upon request.

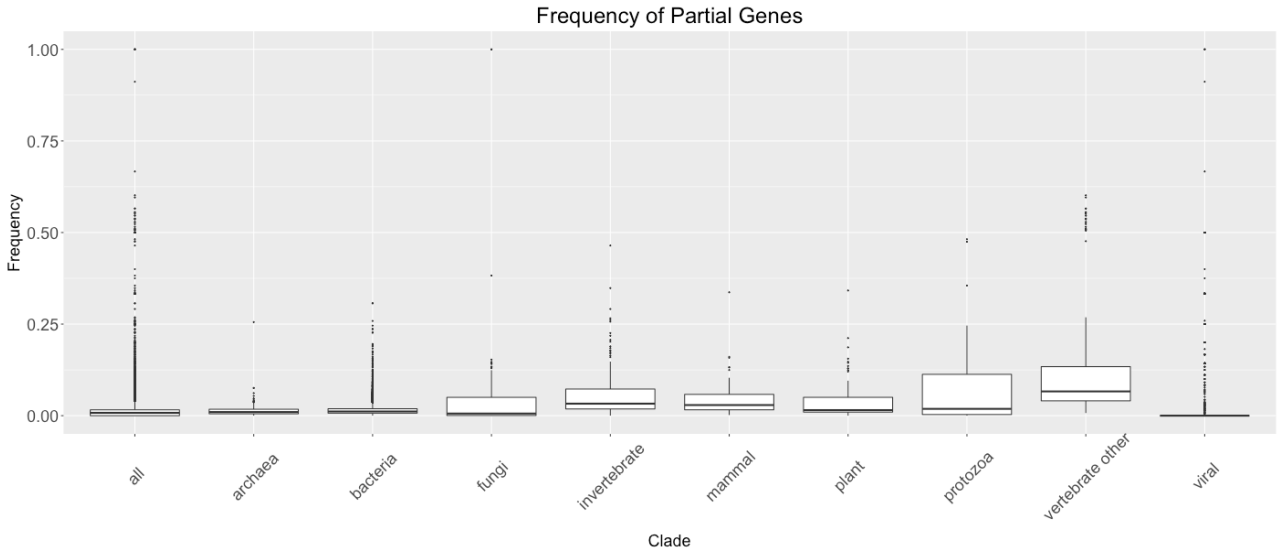
LHX2_CGC 1->0
NTF3_CTC 1->0
HNF1B_CGG 1->0
RSRC1_TAG 0->1
CES2_CGG 1->0
PTP4A2_GAC 1->0
RGS1_TCC 1->0
H3F3B_GTC 1->0
HMGB3_CCC 1->0
PDE6H_ACA 1->0
PSME3_TCT 1->0
ANXA7_CGG 1->0
PPP1R18_CGT 1->0
CAB39L_CGA 1->0
TMEM236_GAA 1->0
DMRT1_TTT 1->0
WASL_GTA 1->0
MAD2L1_GAT 1->0
APBB2_TGC 1->0
PTPRO_GTA 1->0
TBC1D4_CAA 1->0
TBC1D4_TTG 1->0
INPP4A_CTT 1->0
CYTH3_GTT 1->0
ERCC1_GGC 1->0
EMC6_GGG 1->0
ZFYVE9_TTA 1->0
ENOSF1_CTA 1->0
HACD1_CAT 1->0
PLXDC2_CTT 1->0

Appendix 3: Supplementary Figures and Tables for Chapter 4

Supplementary Table 1, Chapter 4. The advantages and disadvantages of each phylogenetic comparison metric are outlined. The first column is the name of the metric. The second column is a short description of how the metric works. The third and fourth columns explain the advantages and disadvantages of each method, respectively.

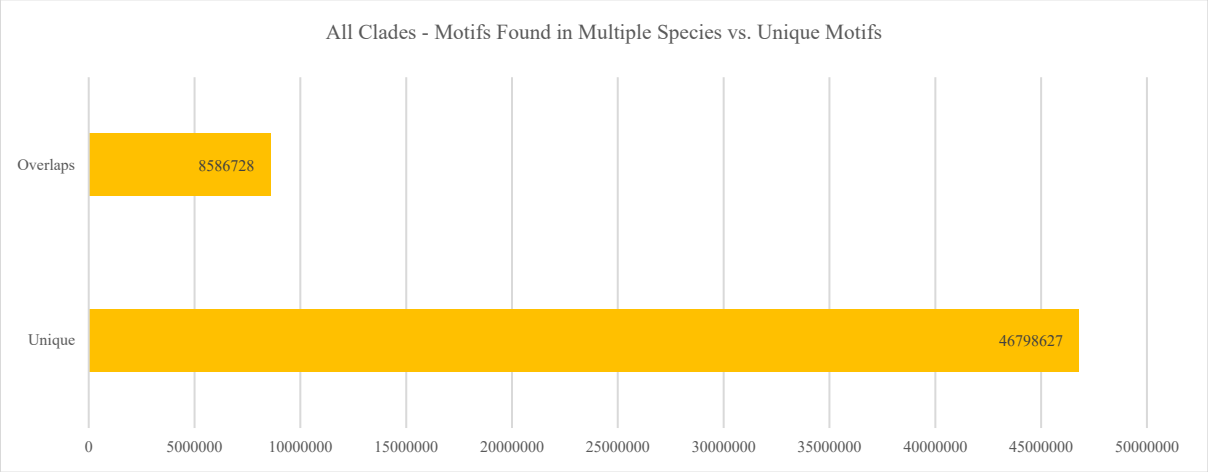
Metric	Description	Advantages	Shortcomings
Percentage of edge similarity[23]	Percentage of branches in both query and subject trees	- Useful for large trees - Useful for trees that contain polytomies	- Does not provide specific information of the location of differences.
Robinson Foulds Distance[26]	Counts number of edges present in one tree, but not the other tree	- Independent from any model of tree editing - Relies only on current characteristics of the two trees	- Sensitive to small changes in leaf nodes[25] - Provides low amount of discrimination for large trees[25]
Maximum Agreement SubTree[27]	Determines the smallest collection of leaves that, when removed, induce the same tree	- Useful for sizeable collections of trees - Useful for smaller trees with "rogue" taxa (taxa whose placement is unclear)	- Requiring exact agreement is computationally demanding and may lead to inaccurate results
Edit distance metrics (Nearest Neighbor Interchange, Subtree Prune and Regraft, Tree Bisection and Reconnection)[28]	Smallest number of allowed operations that will transform one tree into another	- Useful for smaller trees - Useful for when the change operations done on trees are known	- NP-hard - Unclear which operations to use

Supplementary Figure 1, Chapter 4. This figure shows the proportion of partial genes in each clade. A partial gene is defined as a gene in which we do not have the entire DNA sequence available. Each boxplot represents the distribution of the proportion of partial genes in each species of the clade.

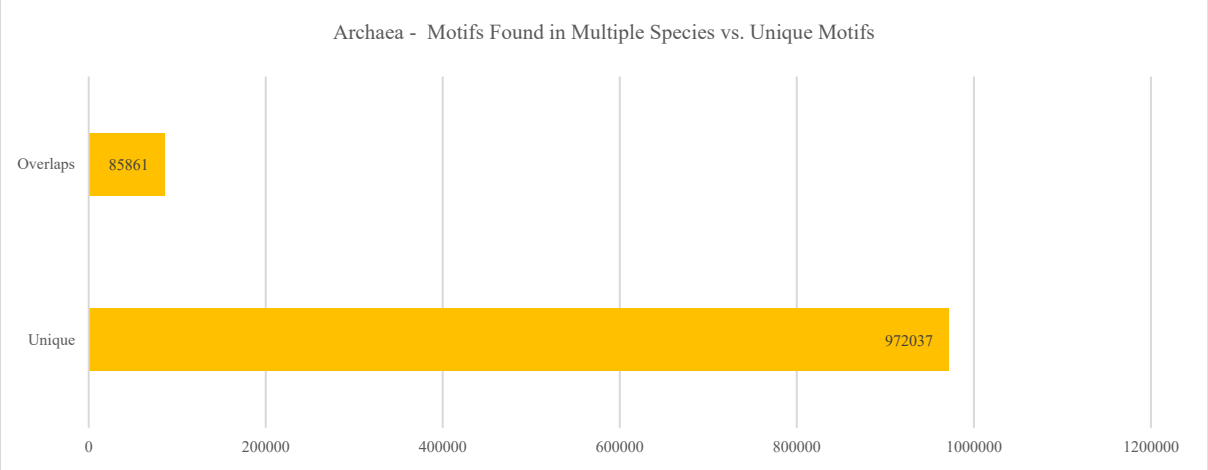


The following figures show how many motifs are unique versus motifs that are found in multiple genes of the clades.

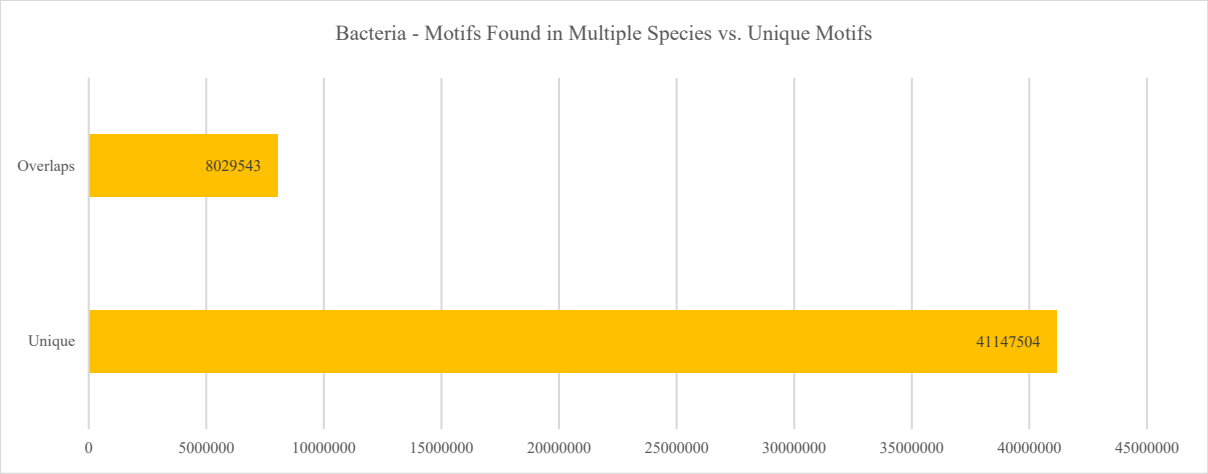
Supplementary Figure 2, Chapter 4.



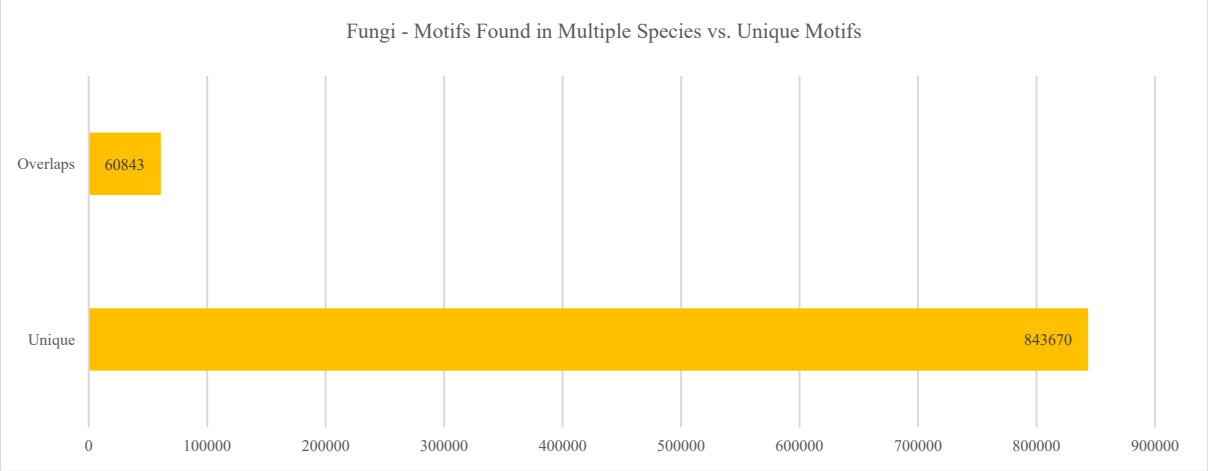
Supplementary Figure 3, Chapter 4.



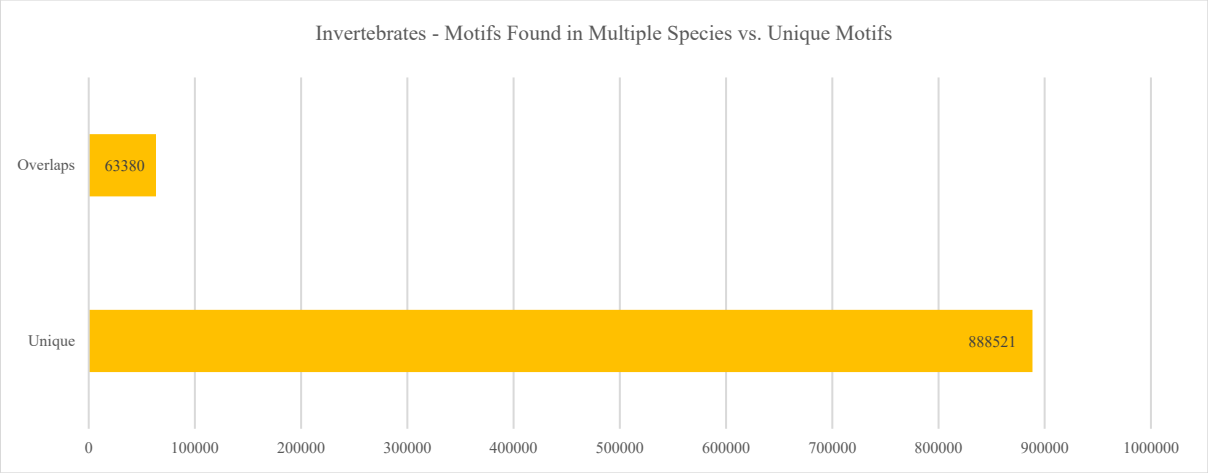
Supplementary Figure 4, Chapter 4.



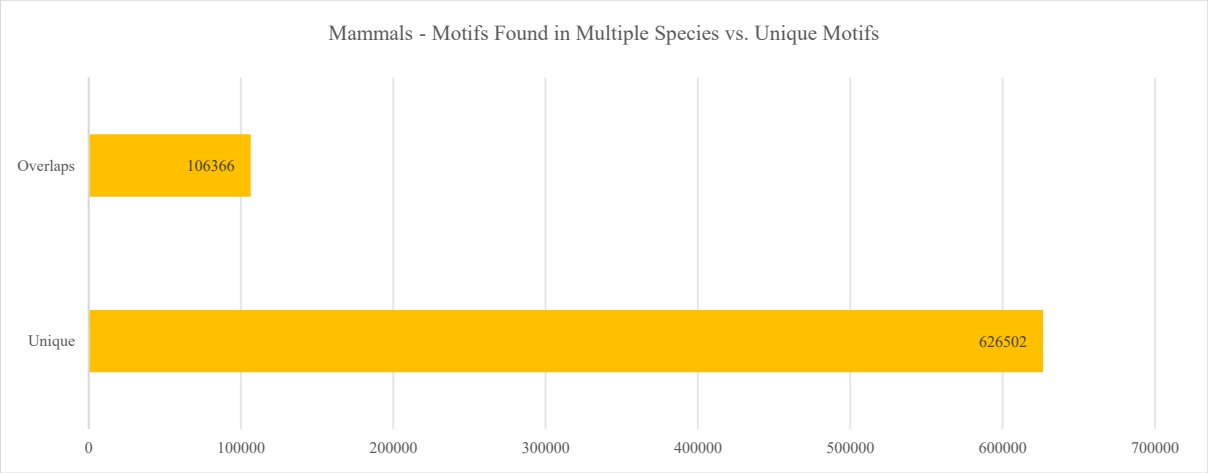
Supplementary Figure 5, Chapter 4.



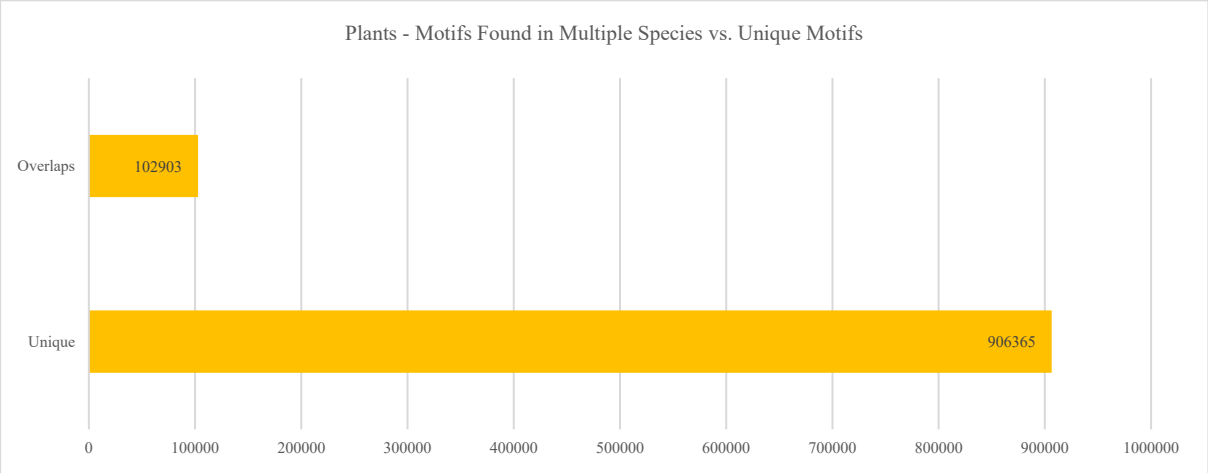
Supplementary Figure 6, Chapter 4.



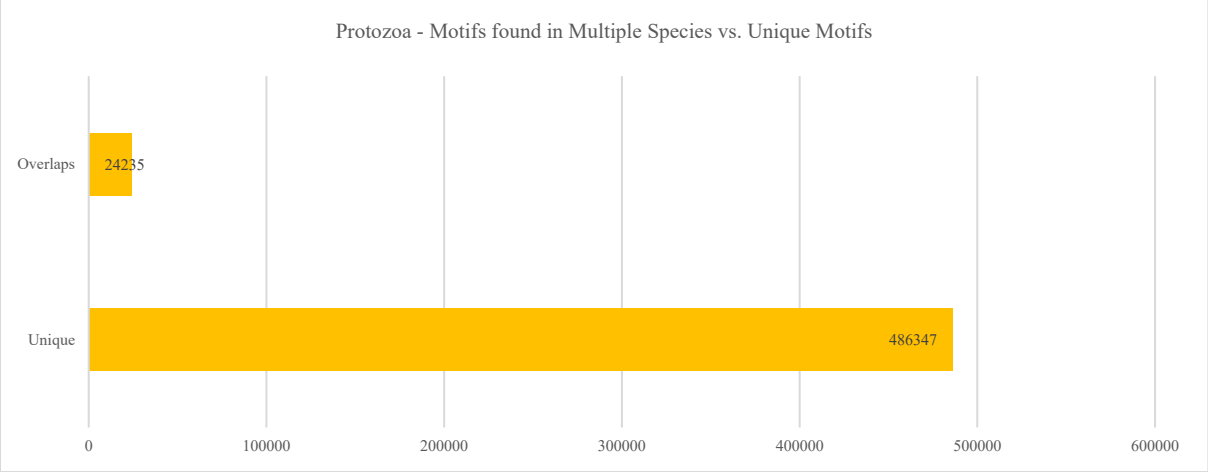
Supplementary Figure 7, Chapter 4.



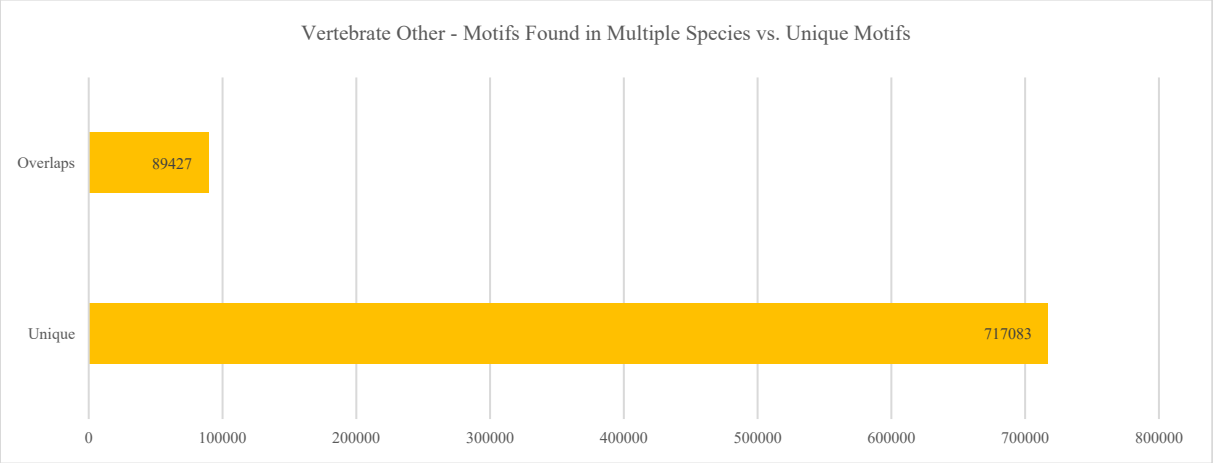
Supplementary Figure 8, Chapter 4.



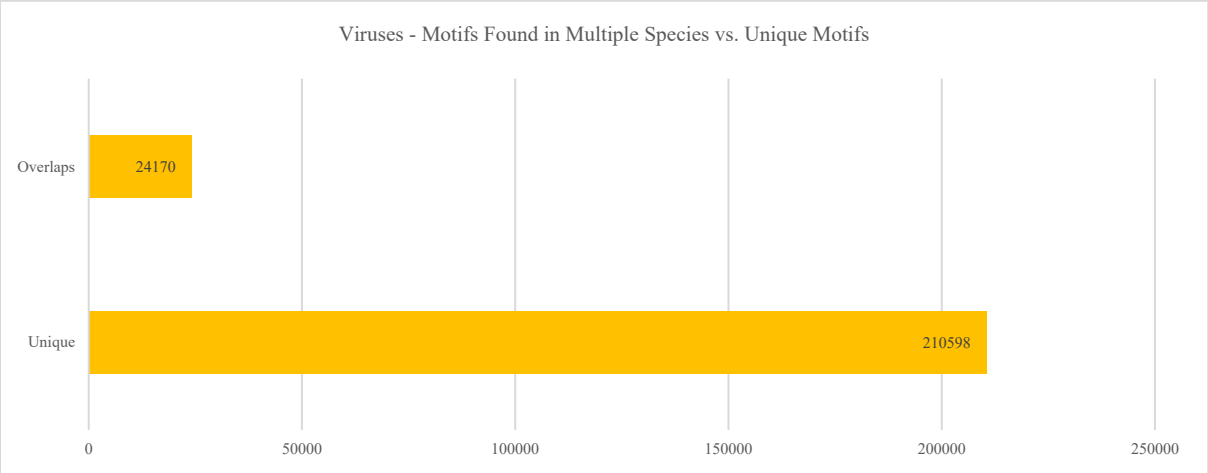
Supplementary Figure 9, Chapter 4.



Supplementary Figure 10, Chapter 4.

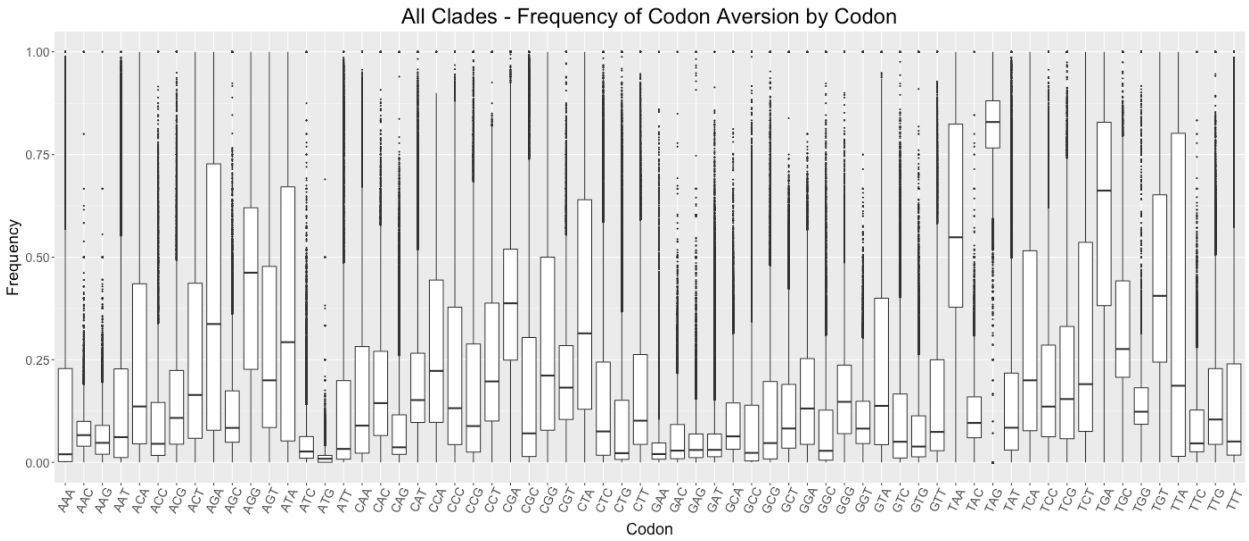


Supplementary Figure 11, Chapter 4.

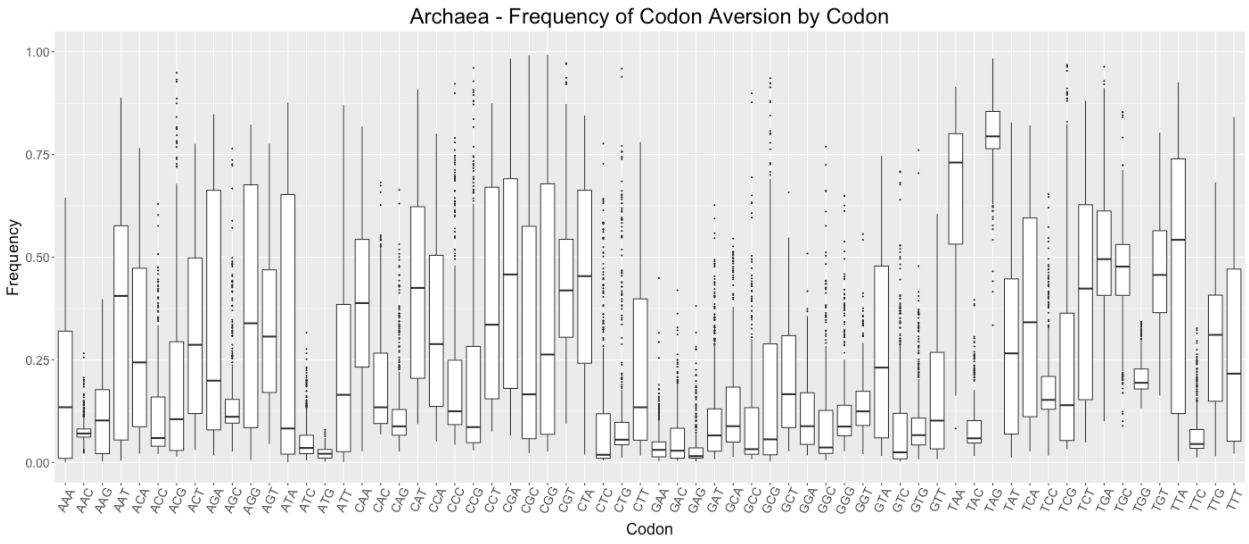


The following figures show the frequency of codons exclusion for each clade. Each box plot in the graphs represents the distribution that species in the clade exclude a certain codon.

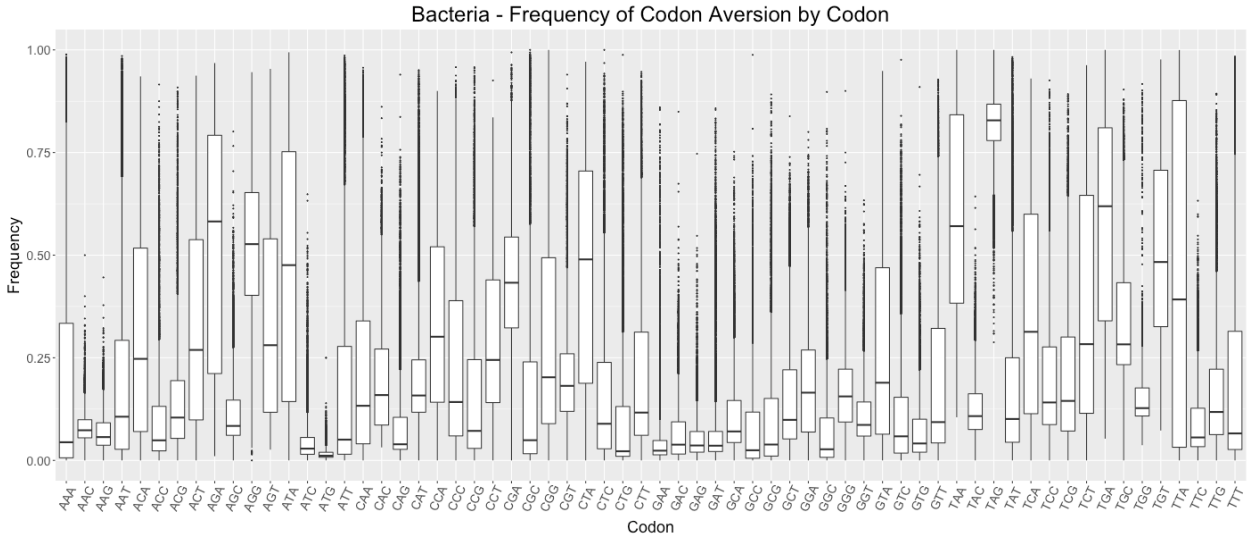
Supplementary Figure 12, Chapter 4.



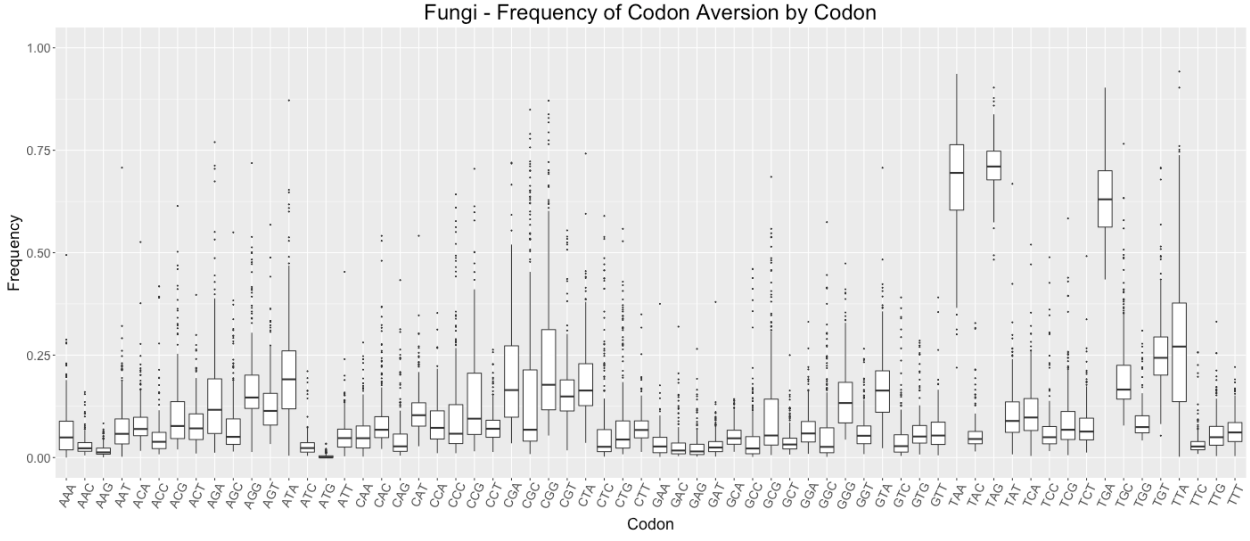
Supplementary Figure 13, Chapter 4.



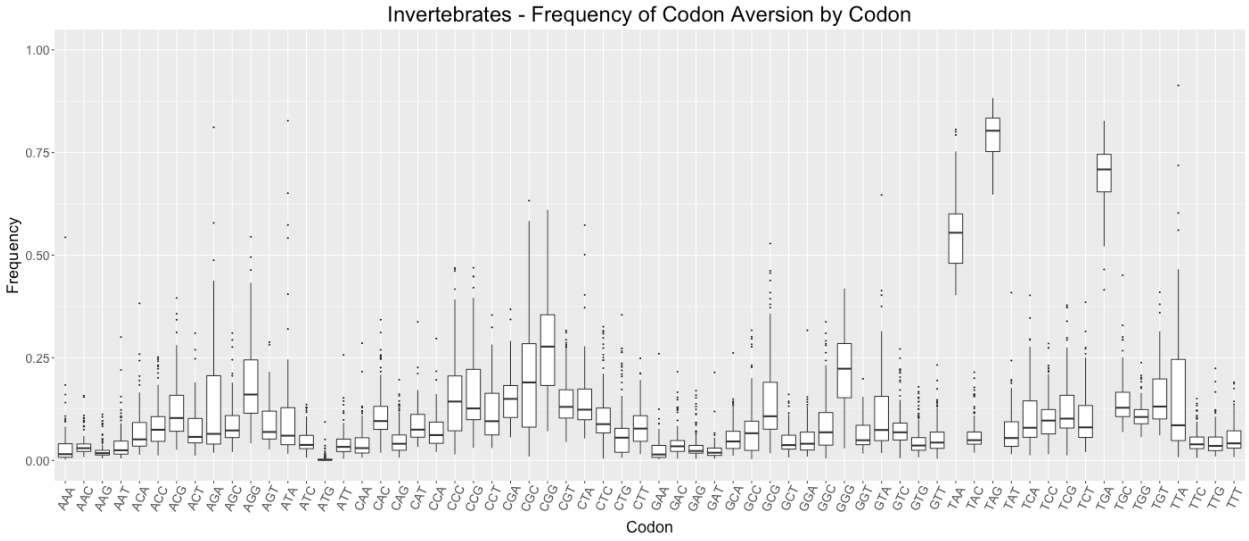
Supplementary Figure 14, Chapter 4.



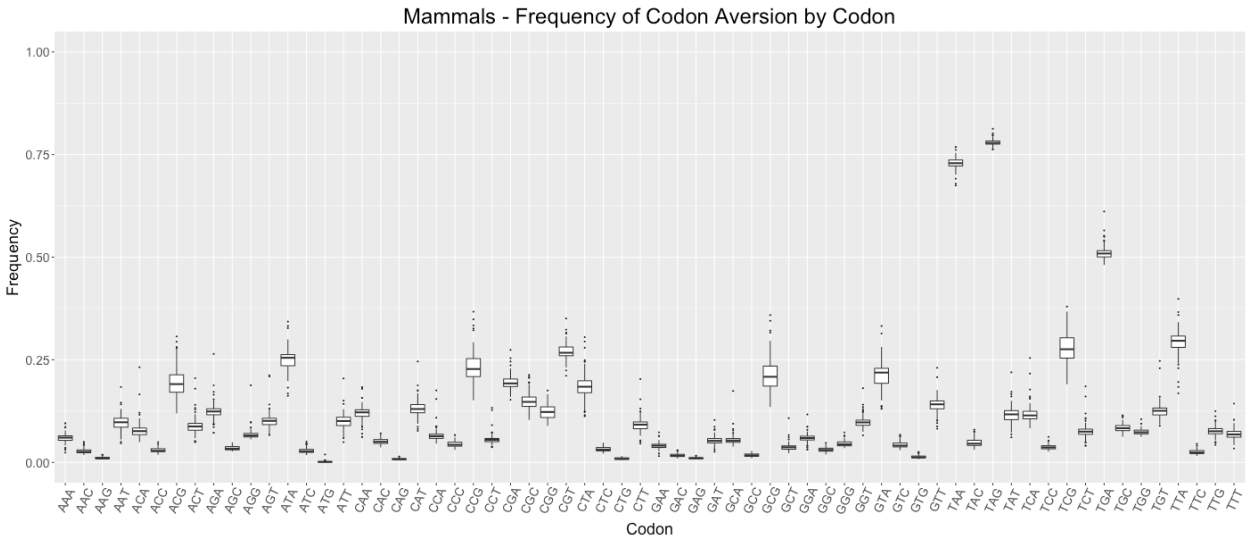
Supplementary Figure 15, Chapter 4.



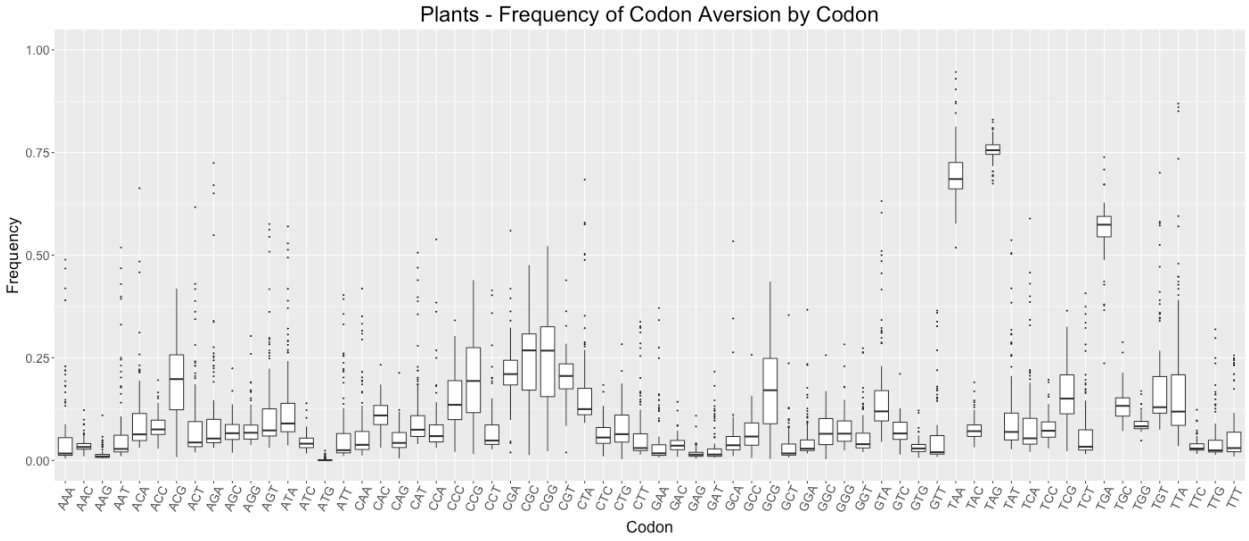
Supplementary Figure 16, Chapter 4.



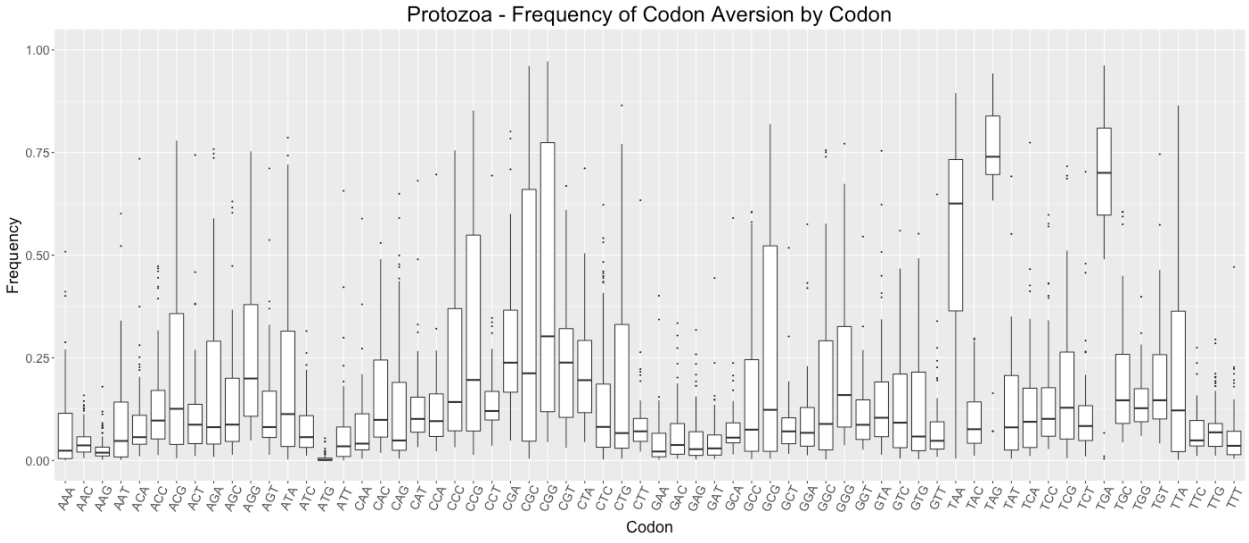
Supplementary Figure 17, Chapter 4.



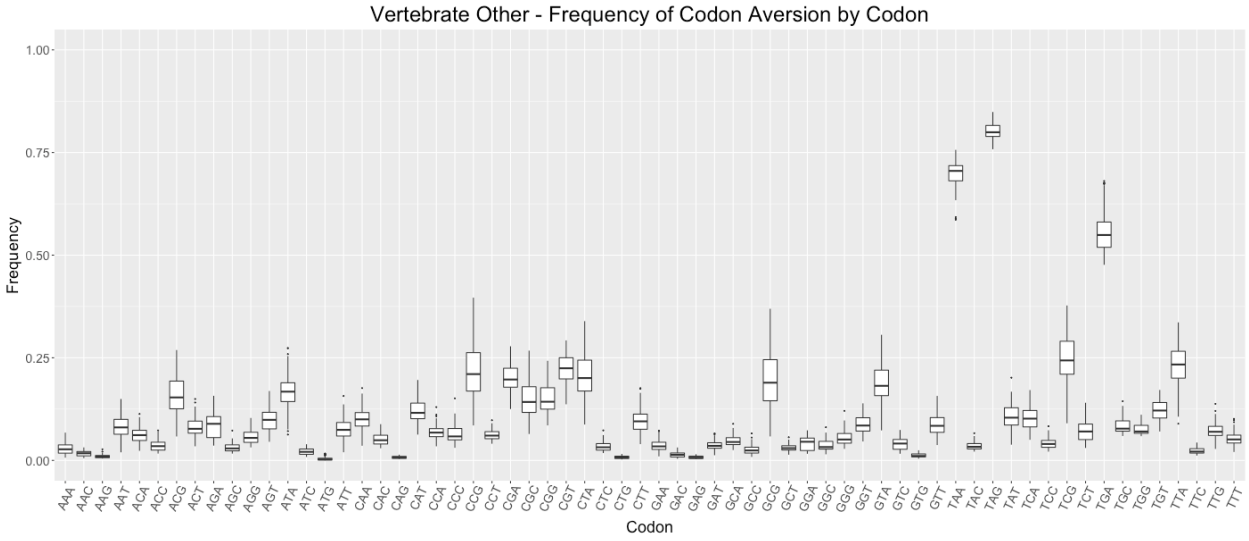
Supplementary Figure 18, Chapter 4.



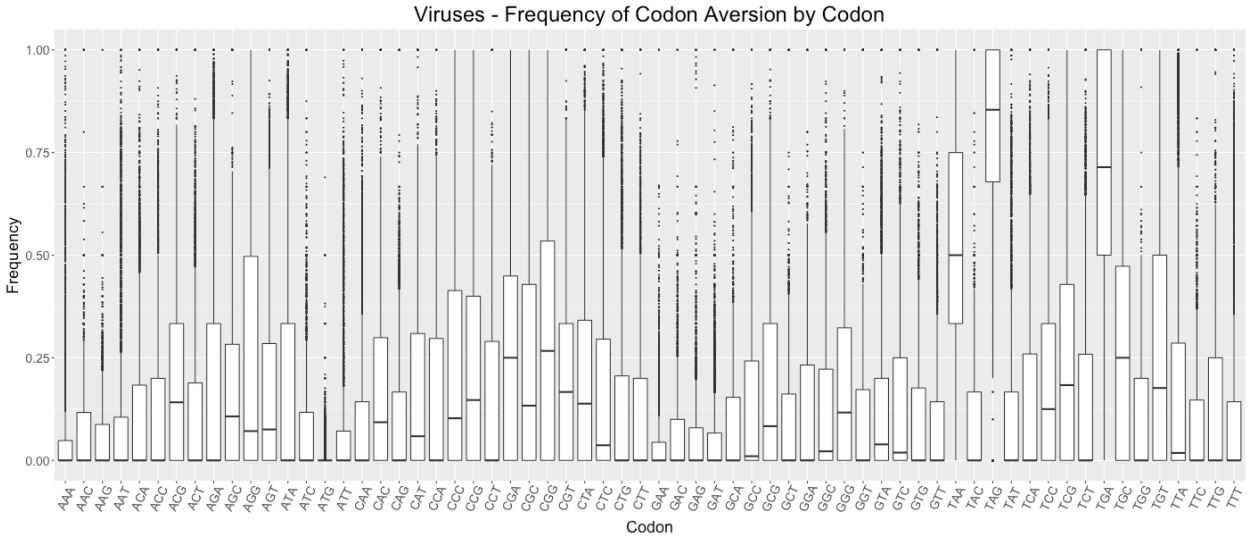
Supplementary Figure 19, Chapter 4.



Supplementary Figure 20, Chapter 4.

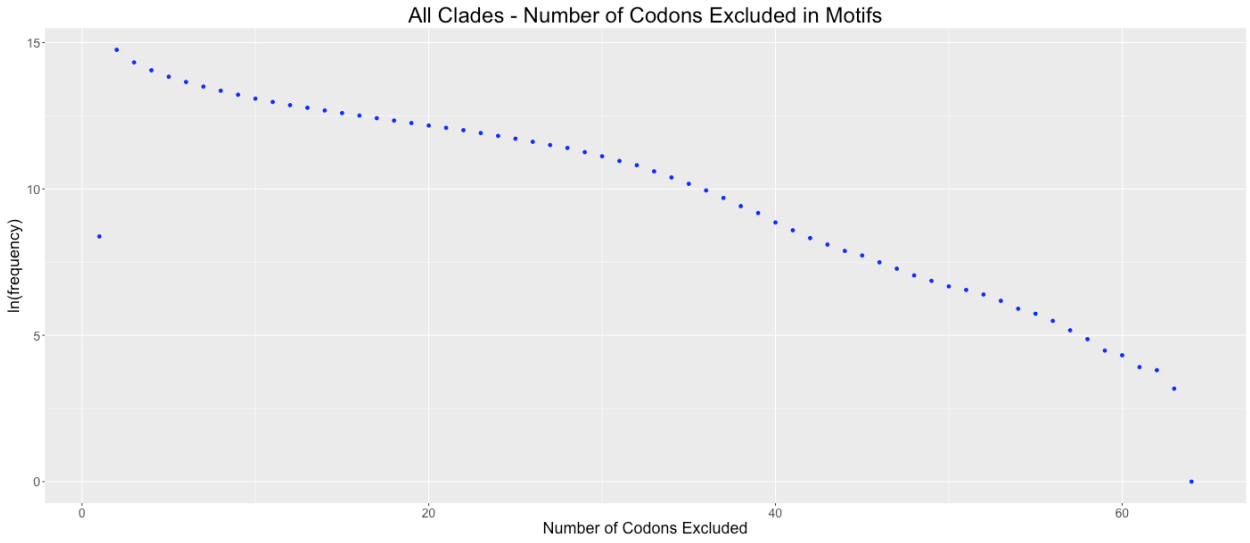


Supplementary Figure 21, Chapter 4.

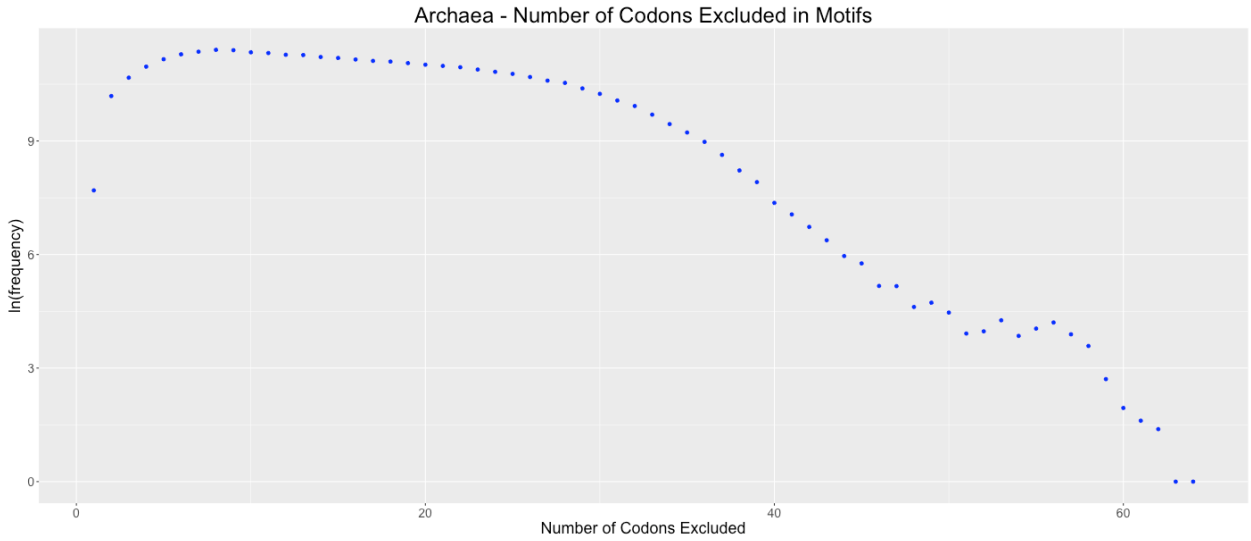


The following figures show how many codons are not used in each gene of the corresponding clades.

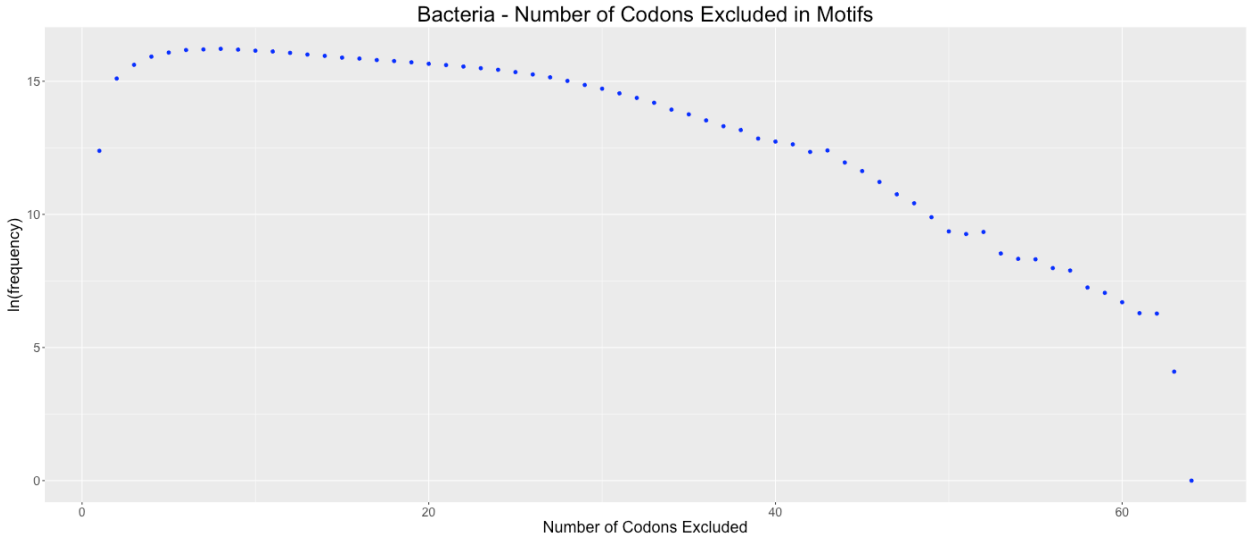
Supplementary Figure 22, Chapter 4.



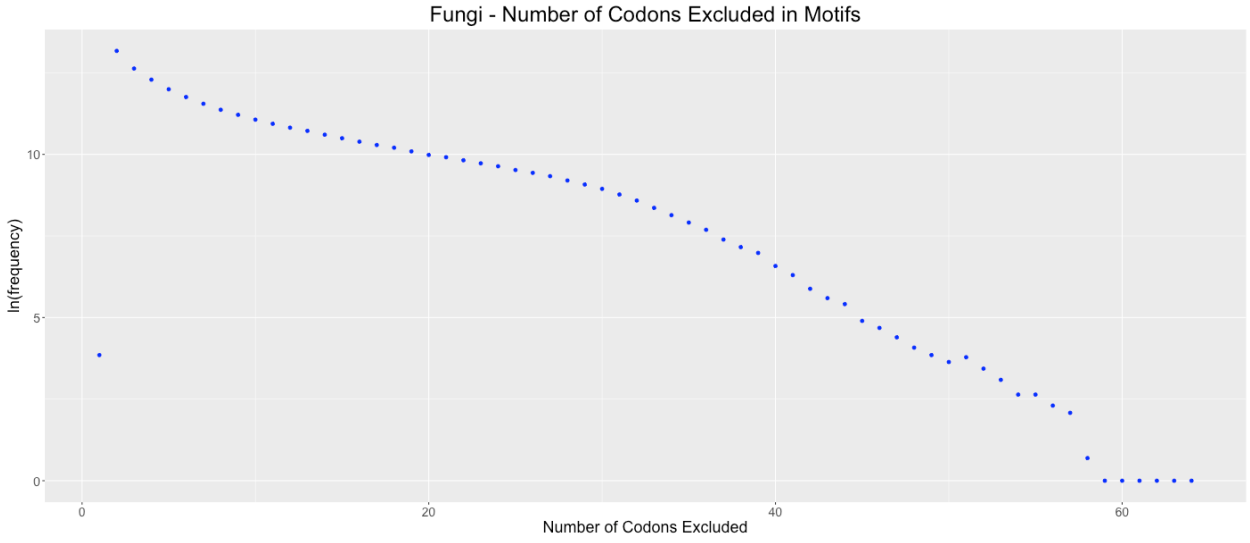
Supplementary Figure 23, Chapter 4.



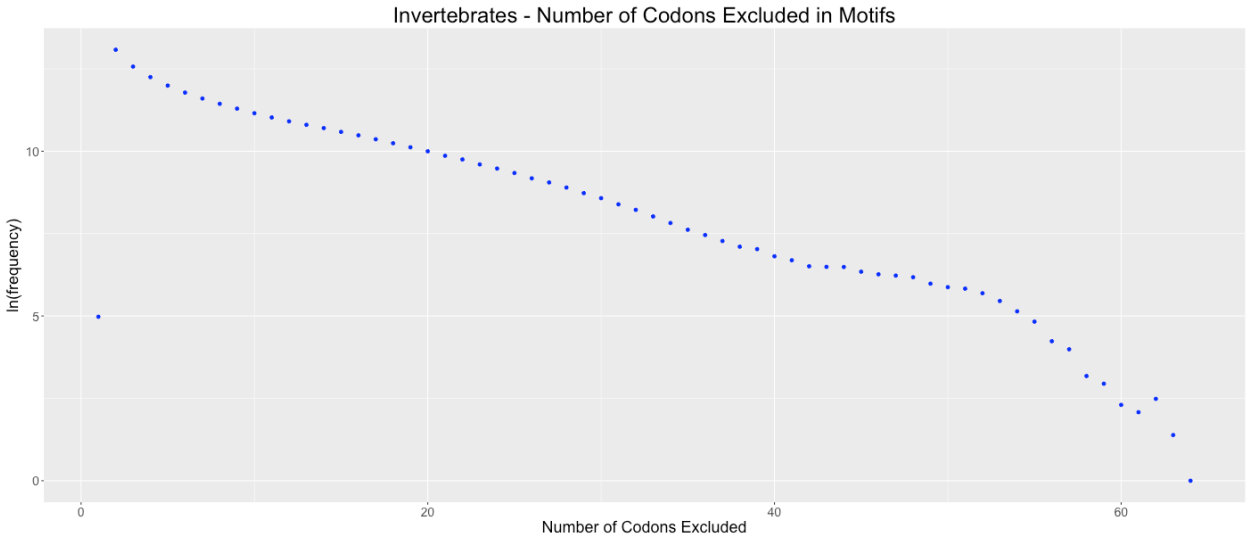
Supplementary Figure 24, Chapter 4.



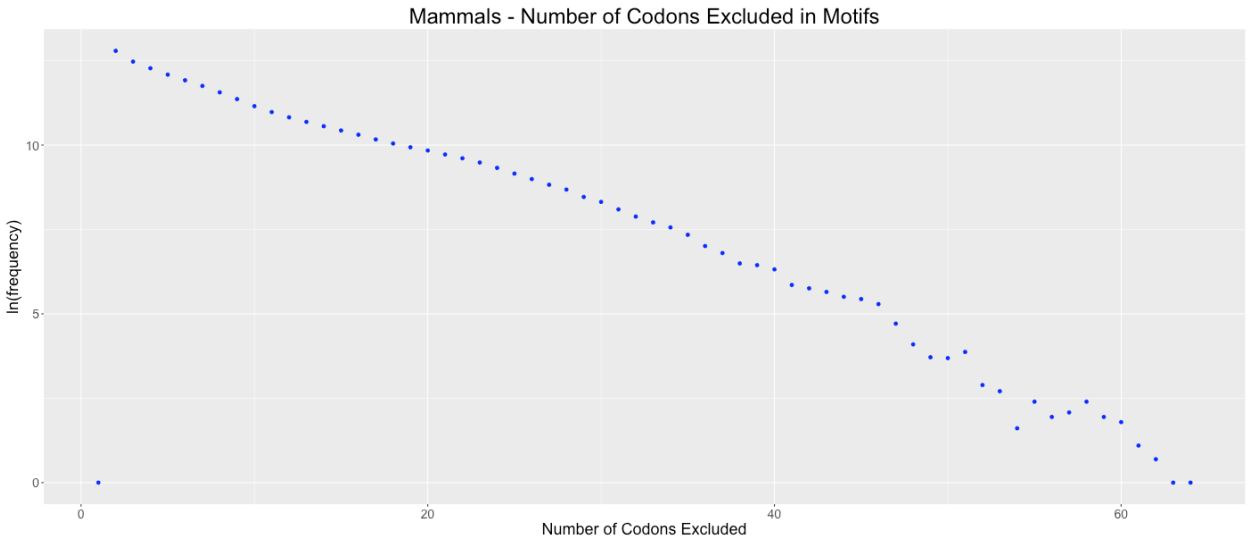
Supplementary Figure 25, Chapter 4.



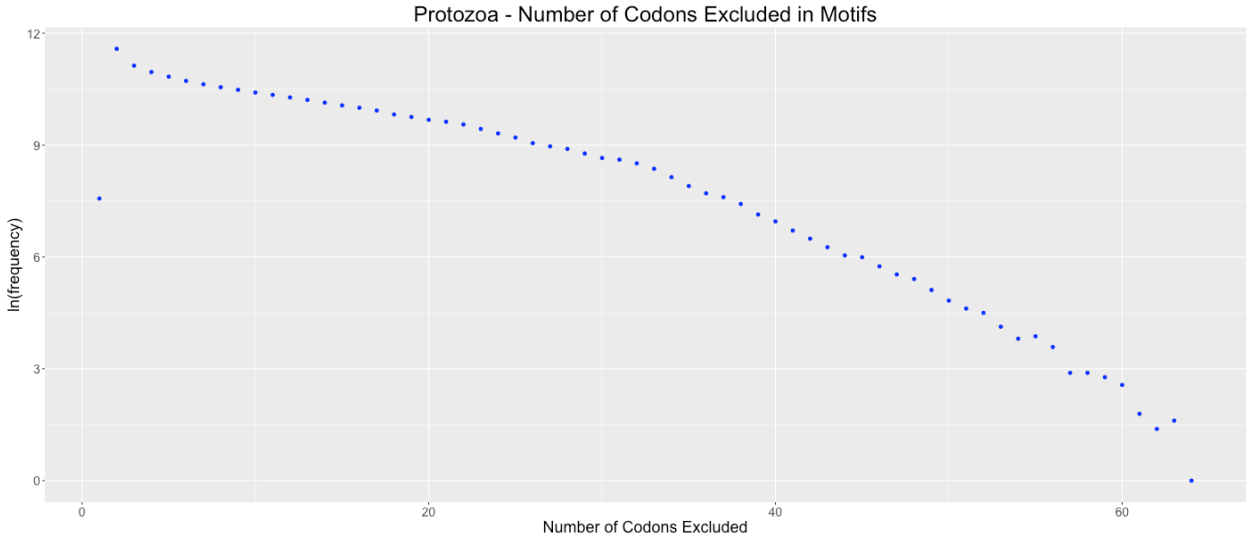
Supplementary Figure 26, Chapter 4.



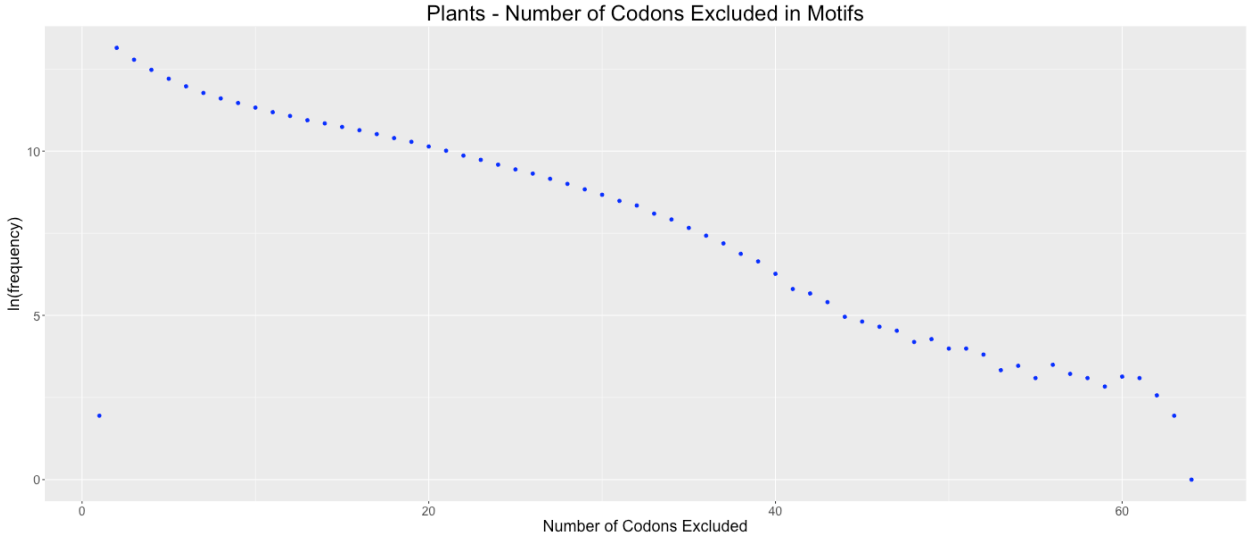
Supplementary Figure 27, Chapter 4.



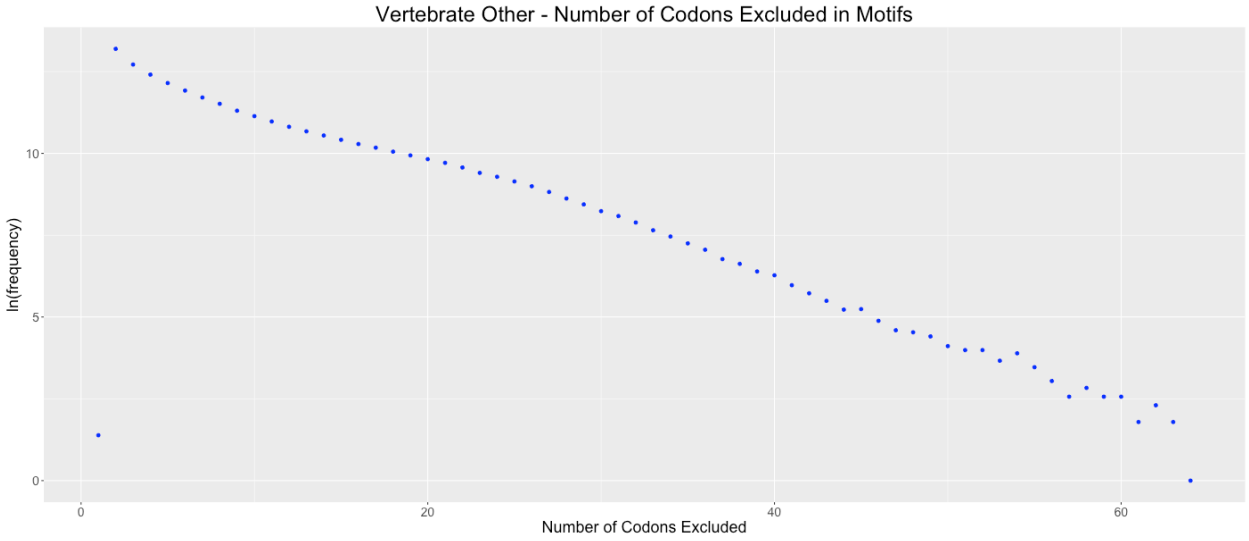
Supplementary Figure 28, Chapter 4.



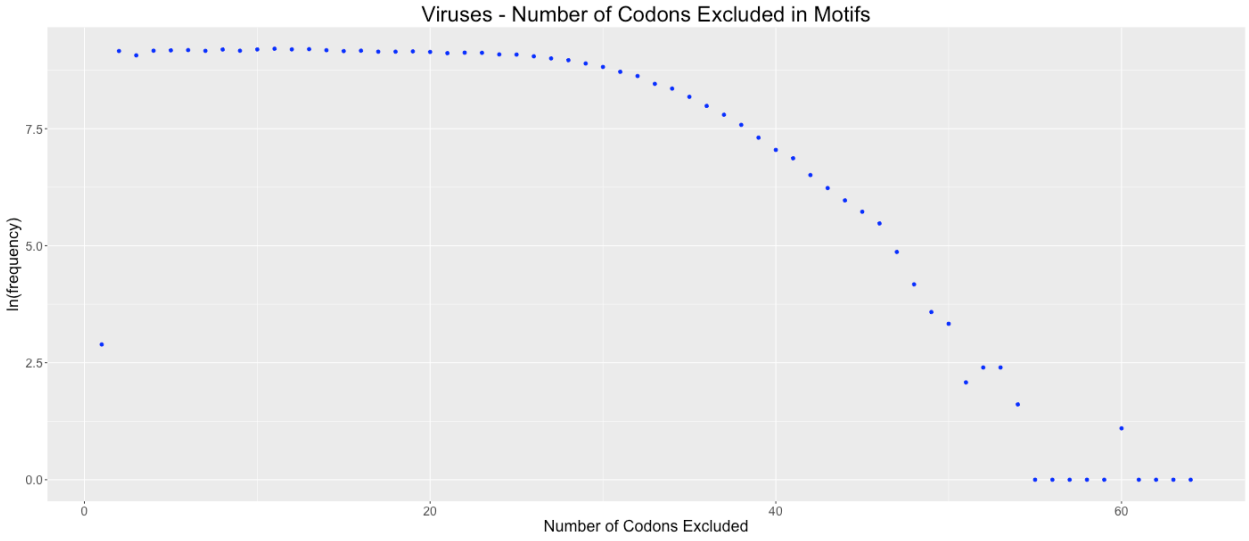
Supplementary Figure 29, Chapter 4.



Supplementary Figure 30, Chapter 4.

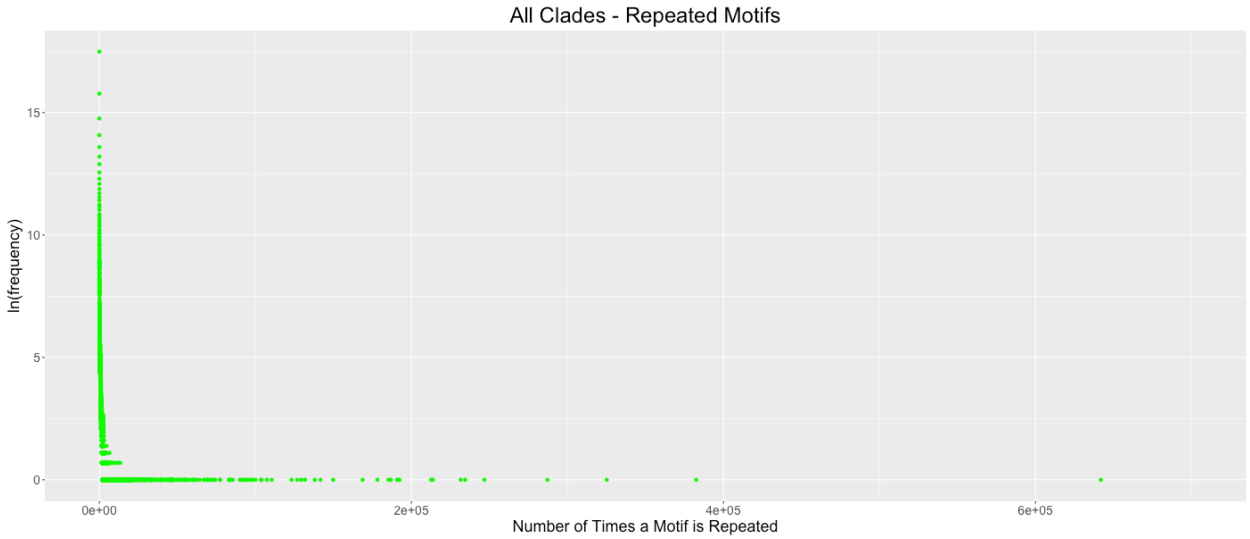


Supplementary Figure 31, Chapter 4.



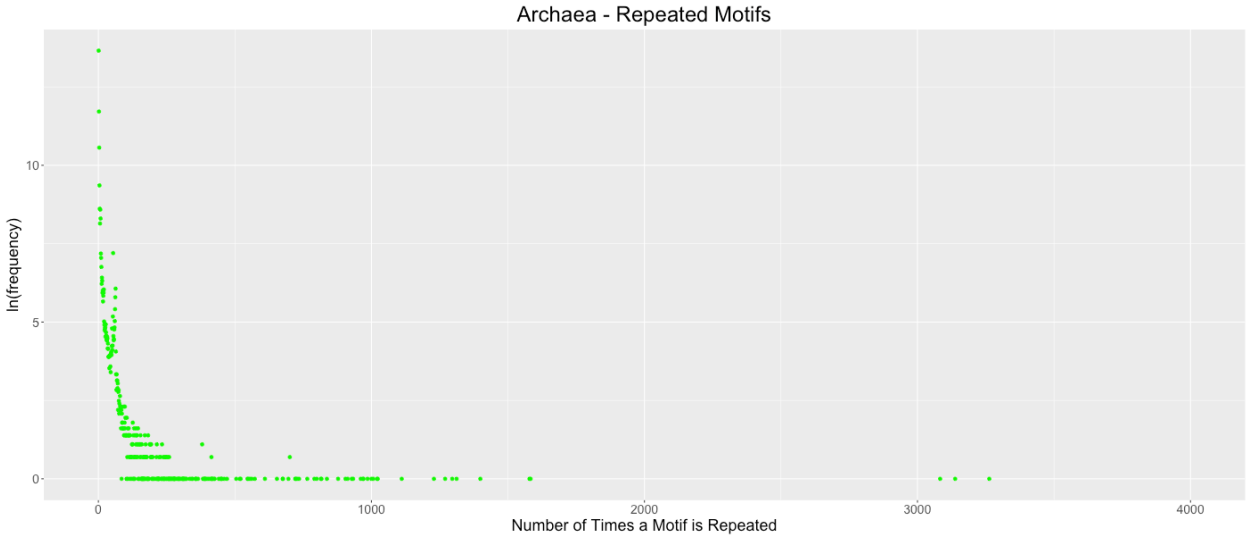
The following figures show the frequency of which codon motifs are repeated in each clade. The x- axis represents how many times a motif was repeated in all the genes in a clade. The y-axis represents how many motifs were repeated a given number of times (shown in the natural log). Some outliers were removed from each graph for clarity. These outliers represent the motifs in which only stop codons are excluded.

Supplementary Figure 32, Chapter 4.



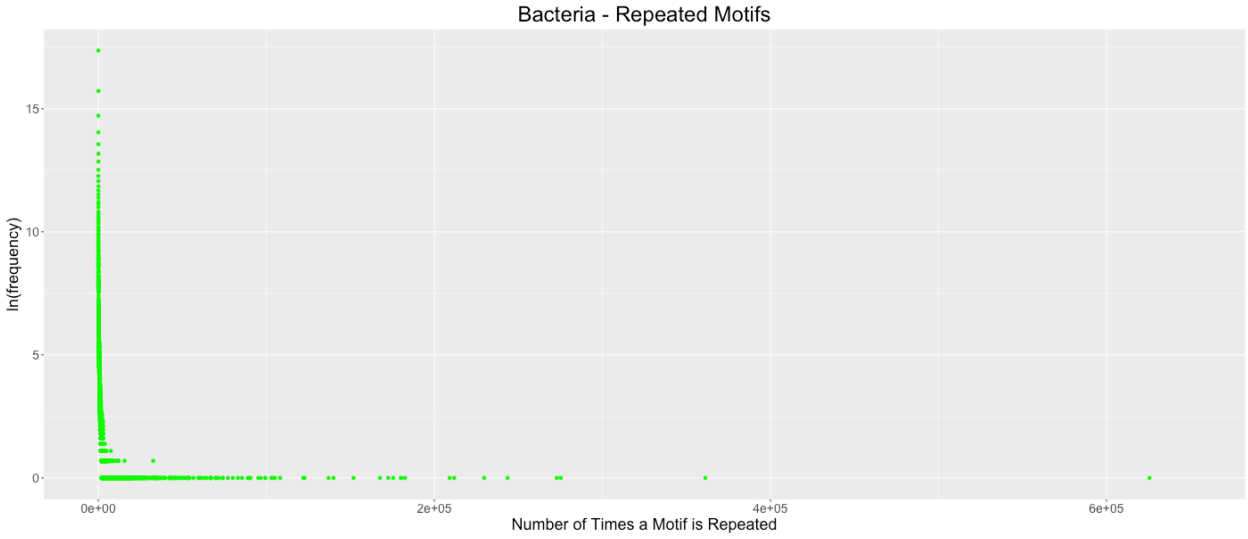
All clades outliers excluded: (1309911,1), (2185083,1), (2433089,1)

Supplementary Figure 33, Chapter 4.



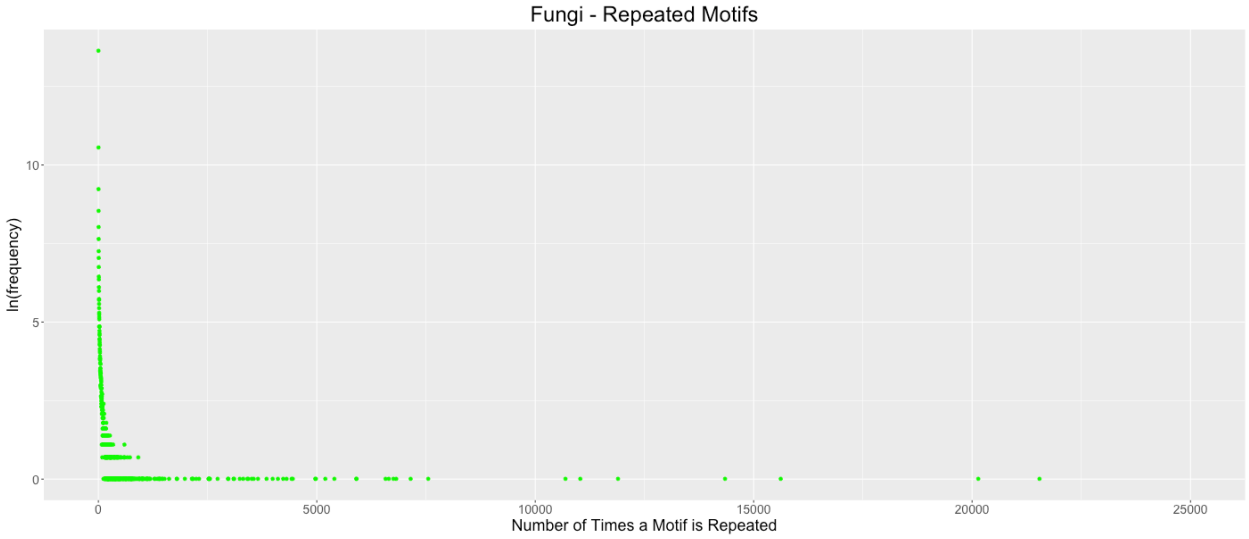
Archaea outliers excluded: (10360,1), (10564,1)

Supplementary Figure 34, Chapter 4.



Bacteria outliers excluded: (681998,1), (1085854,1), (1611727,1)

Supplementary Figure 35, Chapter 4.

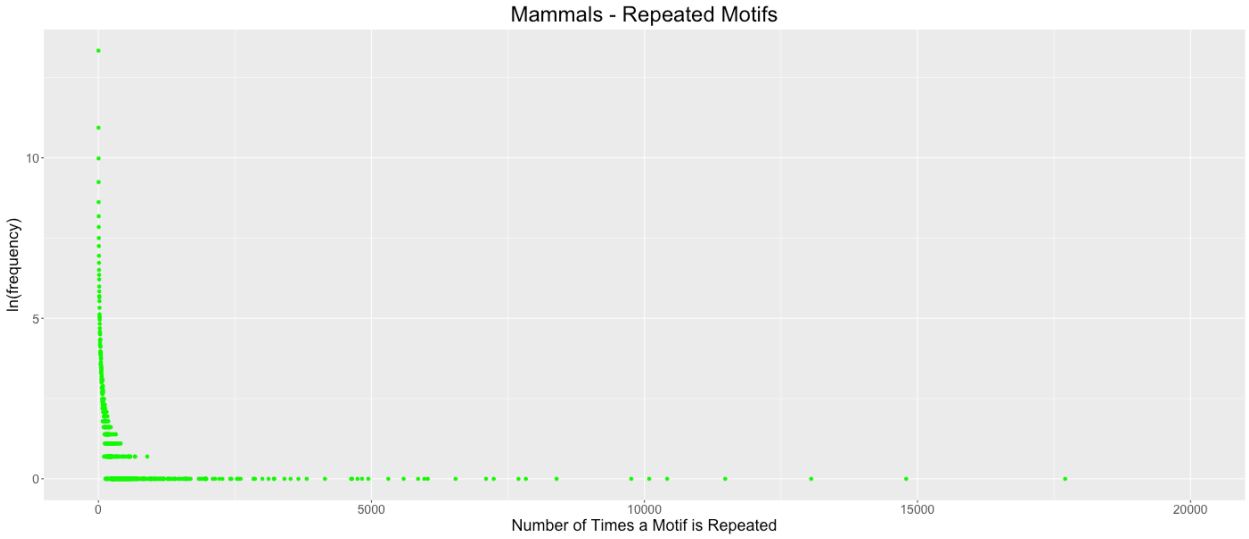


Fungi outliers excluded: (140907,0), (157884,1), (226451,1)

Supplementary Figure 36, Chapter 4.

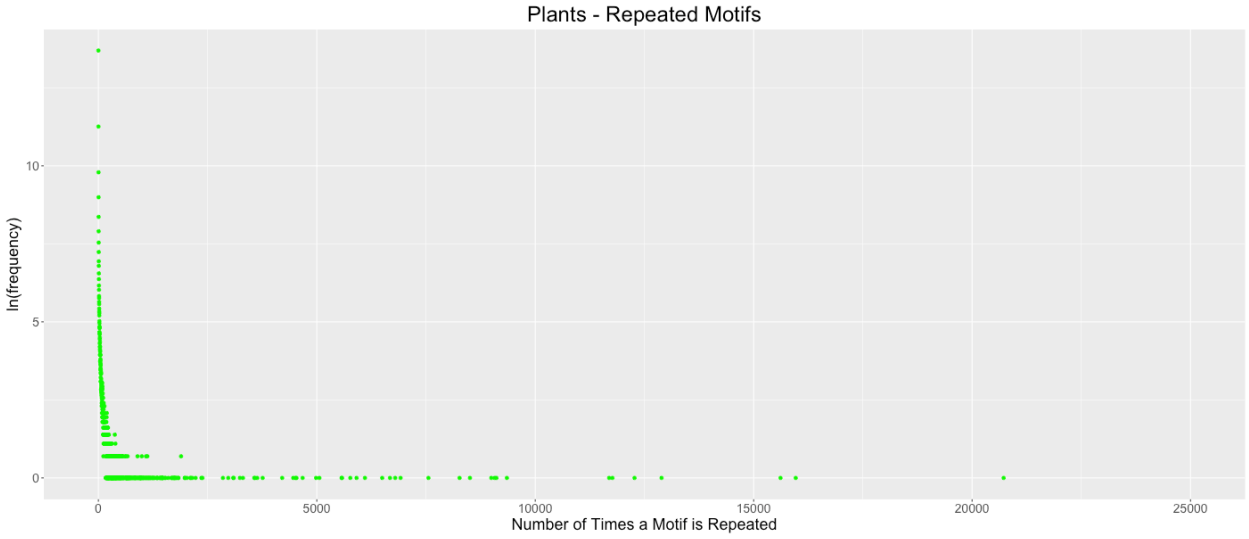
Invertebrates outliers excluded: (110662,1), (201864,1), (166597,1)

Supplementary Figure 37, Chapter 4.



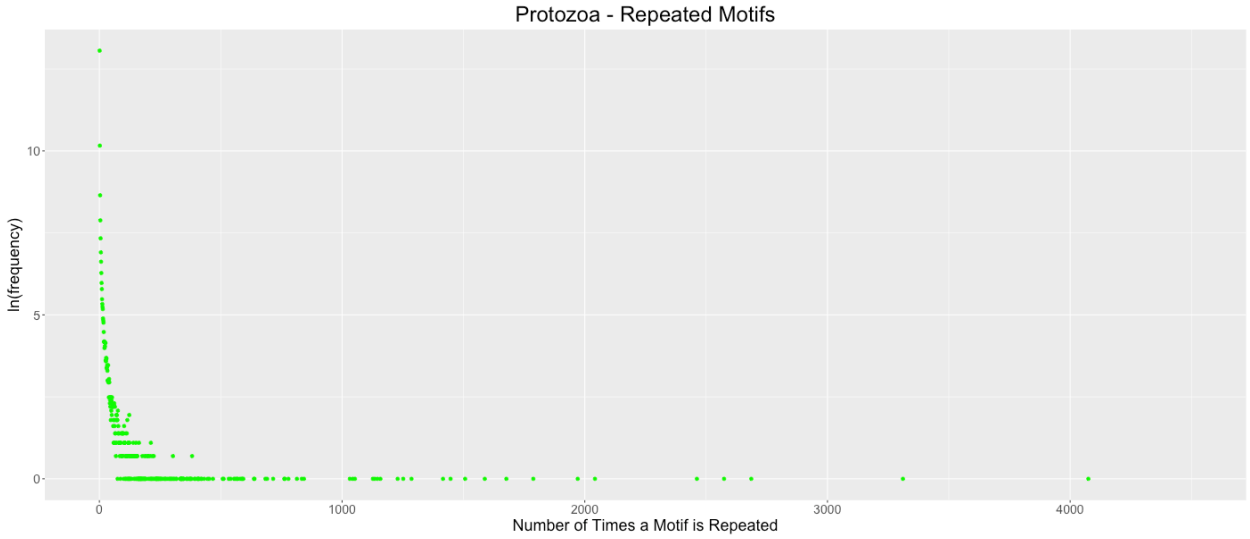
Mammal outliers excluded: (81051,1), (105156,1), (17812,1)

Supplementary Figure 38, Chapter 4.



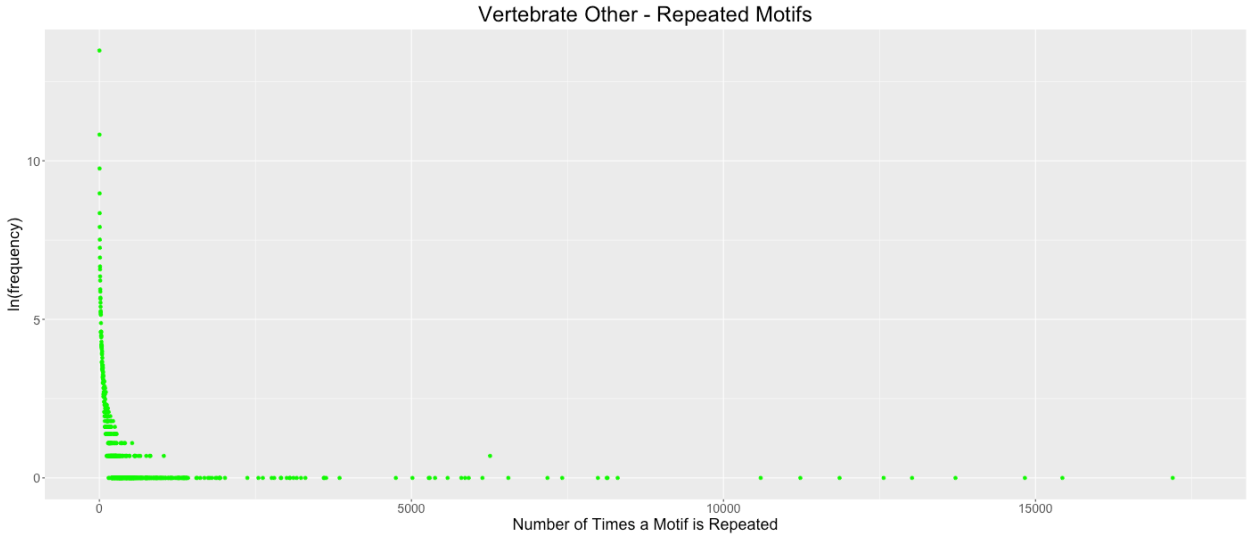
Plant outliers excluded: (158430,1), (127795,1), (224688,1)

Supplementary Figure 39, Chapter 4.



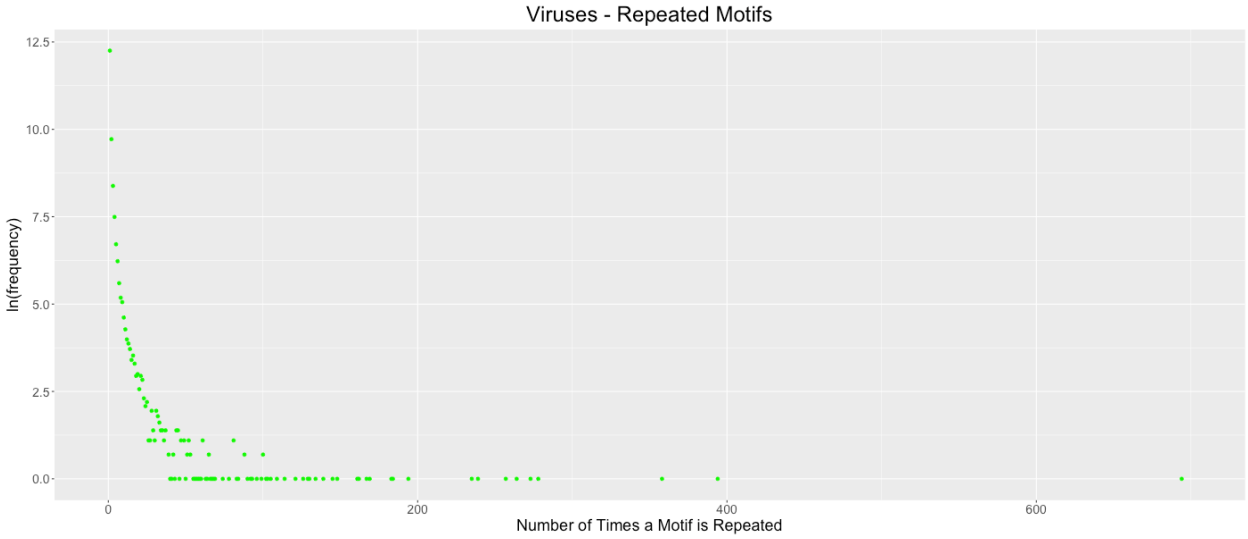
Protozoa Outliers excluded: (32139,1), (41048,1), (30539,1)

Supplementary Figure 40, Chapter 4.



Vertebrate other outliers excluded: (167892,1), (114746,1), (254804,1)

Supplementary Figure 41, Chapter 4.



Viruses outliers excluded: (2669,1), (2167,1), (4664,1)

Appendix 4: Supplementary Figures and Tables for Chapter 5

Supplementary Tables for Chapter 5

Supplementary Table 1, Chapter 5: Species Included in Identical Codon Pairing

Taxonomic Group	2	3	4	5	6	7	8	9	10	11	Average	Total Number of Species
Archaea	106	95	95	95	95	95	95	95	95	95	96.1	418
Fungi	19	9	13	9	9	19	9	13	9	9	11.8	234
Invertebrates	65	55	57	55	55	55	63	55	57	55	57.2	149
Plants	60	60	60	61	61	61	61	61	61	61	60.7	89
Protozoa	15	15	15	16	17	16	16	20	15	15	16	75
Mammals	97	97	97	97	97	97	97	97	97	97	97	107
Other vertebrates	114	114	114	114	114	114	114	114	114	114	114	123
Viruses	168	137	152	188	177	174	174	220	176	184	175	7 233

Each column is the length of the ribosome window, in codons, that was used and each cell is the number of species that included at least 5% of the parsimony-informative codons for each taxonomic group. Column 12 is the average number of species included in each ribosome window. Column 13 is the total number of species in each taxonomic group.

Supplementary Table 2, Chapter 5: Species Used in Co-tRNA Codon Pairing

Taxonomic Group	2	3	4	5	6	7	8	9	10	11	Average	Total Number of Species
Archaea	107	107	107	107	107	107	107	107	107	106	106.9	418
Fungi	9	10	10	20	9	20	9	14	13	10	12.7	234
Invertebrates	65	65	66	55	65	57	65	55	57	55	60.5	149
Plants	61	59	59	59	59	59	59	59	59	59	59.2	89
Protozoa	17	16	19	17	15	18	15	16	16	16	16.5	75
Mammals	97	97	97	97	97	97	97	97	97	97	97	107
Other vertebrates	114	114	114	114	114	114	114	114	114	114	114	123
Viruses	282	279	262	247	243	256	256	245	236	244	255	7 233

Each column is the length of the ribosome window, in codons, that was used and each cell is the number of species that included at least 5% of the parsimony-informative amino acids for each taxonomic group. Column 12 is the average number of species included in each ribosome window. Column 13 is the total number of species in each taxonomic group.

Supplementary Table 3, Chapter 5: Species Used in Combined Comparison

Taxonomic Group	2	3	4	5	6	7	8	9	10	11	Average	Total Number of Species
Archaea	96	100	96	100	101	101	101	95	105	106	100.1	418
Fungi	13	13	10	14	11	11	13	10	11	14	12	234
Invertebrates	55	57	57	69	93	58	57	65	65	58	63.4	149
Plants	61	61	59	59	59	59	59	59	59	59	59.4	89
Protozoa	15	15	25	27	17	18	21	20	20	21	19.9	75
Mammals	97	97	97	97	97	97	97	97	97	97	97	107
Other vertebrates	114	114	114	114	114	114	114	114	114	114	114	123
Viruses	199	224	180	190	190	180	278	262	261	257	222.1	7 233

Each column is the length of the ribosome window, in codons, that was used and each cell is the number of species that included at least 5% of the parsimony-informative amino acids for each taxonomic group. Column 12 is the average number of species included in each ribosome window. Column 13 is the total number of species in each taxonomic group.

Supplementary Table 4, Chapter 5: Informative Codons used in Identical Codon Pairing

Taxonomic Group	2	3	4	5	6	7	8	9	10	11	Average
Archaea	6151	8450	9902	10544	11035	11254	11518	11687	11664	11842	10404.7
Fungi	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Invertebrates	794	988	1081	1160	1236	1263	1329	1427	1353	1423	1205.4
Plants	6230	8033	9036	9842	10153	10517	10607	10691	10725	10693	9652.7
Protozoa	12449	14864	16051	16253	16103	16171	15837	15838	15764	15532	15486.2
Mammals	197074	311796	319490	381908	404078	335058	398896	386949	436474	380879	355260.2
Other vertebrates	228194	277024	347408	388789	355121	376798	400771	380813	390161	15532	316061.1
Viruses	16622	23528	28145	28776	30176	30768	32248	31580	33374	33082	28829.9

Supplementary Table 5, Chapter 5: Informative Codons used in Co-tRNA Codon Pairing

Taxonomic Group	2	3	4	5	6	7	8	9	10	11	Average
Archaea	3293	3294	3087	2921	2783	2725	2689	2661	2560	2579	2859.2
Fungi	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Invertebrates	418	461	475	455	450	429	422	428	410	382	433
Plants	3219	3188	3157	3082	2945	2940	2929	2819	2808	2723	2981
Protozoa	5415	5319	5020	4840	4589	4358	4198	4067	3976	2819	4460.1
Mammals	94018	10195	93587	93627	82729	74208	70805	71310	69678	55666	71582.3
Other vertebrates	90872	86618	82126	73534	74286	70704	63995	62945	60569	57241	72289
Viruses	11409	12103	11556	11421	11248	11068	10812	10550	10325	10115	11060.7

Supplementary Table 6, Chapter 5: Informative Codons used in the Combined Comparison

Taxonomic Group	2	3	4	5	6	7	8	9	10	11	Average
Archaea	2823	2568	2292	2105	1974	1851	1756	1577	1612	1527	2008.5
Fungi	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Invertebrates	463	464	417	411	359	351	319	302	293	272	365.1
Plants	3236	2813	2571	2347	2214	2085	1977	1849	1789	1709	2259
Protozoa	4488	3603	3021	2813	2381	2311	2122	1951	1862	1730	2628.2
Mammals	72029	64087	56125	44448	44984	37830	45077	41319	34711	35801	44089.2
Other vertebrates	6937	61754	48110	45789	43517	41519	34718	33856	30895	27919	37501.4
Viruses	12737	12003	11587	10946	10352	9924	9691	9247	8921	8578	10398.6

Supplementary Table 7, Chapter 5: Optimal Window Size and Options for Each Taxonomic Group

Taxonomic Group	Alignment-free identical codon pairing (I), co-tRNA codon pairing (C), or both (B)	Alignment-free minimum optimal window sizes	Maximum parsimony identical codon pairing (I), co-tRNA codon pairing (C), or both (B)	Maximum parsimony minimum optimal window sizes
All	B	2	N/A	N/A
Archaea	B	4	B	3
Bacteria*	B	2	N/A	N/A
Fungi	B	5	N/A	N/A
Invertebrates	B	2	B	4
Plants	I	4	C	10
Protozoa	B	2	B	2
Mammals	I	6	I	2
Other vertebrates	I	5	B	3
Viruses*	B	3	N/A	N/A

The first column indicates the taxonomic group. Column 2 shows if the best percent overlap for the alignment-free method came from identical codon pairing (I), co-tRNA codon pairing (C), or the combined method (B). Column 3 shows the window size used to recover the phylogeny most congruent with the NCBI Taxonomy and the OTL for the alignment-free method. If multiple window sizes recovered phylogenies with the same percent congruence, preference was given to smaller window sizes. Column 4 shows if the best percent overlap for the parsimony method came from identical codon pairing (I), co-tRNA codon pairing (C), or the combined method (B). Column 5 shows the window size used to recover the phylogeny most congruent with the NCBI Taxonomy and the OTL for the parsimony method. If multiple window sizes recovered phylogenies with the same percent congruence, preference was given to smaller window sizes.
*Indicates that some species overlap between the viruses and bacteria.

Supplementary Table 8, Chapter 5:

Taxonomic Group	2	3	4	5	6	7	8	9	10	11
All	90	90	90	90	90	90	90	90	90	90
Archaea	84	84	84	84	85	84	84	83	83	83
Bacteria	91	92	92	92	92	92	92	92	92	92
Fungi	74	73	74	74	74	75	75	74	74	74
Invertebrates	72	72	72	72	73	74	72	72	72	71
Plants	75	78	81	80	80	79	80	80	80	80
Protozoa	79	78	79	79	77	78	76	77	77	77
Mammals	92	92	94	94	95	94	93	91	90	91
Other Vertebrates	81	85	86	87	85	83	81	82	82	81
Viruses	89	89	89	90	90	91	91	91	90	90

Alignment-free: Identical codon pairing percent overlap with the NCBI Taxonomy. Column 1 shows the taxonomic groups analyzed. Columns 2-11 show the percent overlap with the NCBI Taxonomy from a phylogeny recovered using identical codon pairing with the respective window size 2-11. The highest percent overlap for each taxonomic group is highlighted.

Supplementary Table 9, Chapter 5:

Taxonomic Group	2	3	4	5	6	7	8	9	10	11
All	83	83	83	83	83	83	83	83	83	83
Archaea	78	78	77	78	78	77	77	77	77	76
Bacteria	85	85	85	85	85	85	85	85	85	85
Fungi	71	72	71	71	71	72	72	71	71	72
Invertebrates	63	64	63	63	65	65	64	63	62	61
Plants	68	71	73	72	72	71	74	72	74	74
Protozoa	69	70	70	70	68	70	69	69	69	68
Mammals	83	85	86	86	89	87	86	84	84	84
Other Vertebrates	68	71	72	73	72	71	69	70	70	69

Alignment-free: Identical codon pairing percent overlap with the OTL. Column 1 shows the taxonomic group analyzed. Columns 2-11 show the percent overlap with the OTL from a phylogeny recovered using identical codon pairing with the respective window size 2-11. The highest percent overlap for each taxonomic group is highlighted.

Supplementary Table 10, Chapter 5:

Taxonomic Group	2	3	4	5	6	7	8	9	10	11
All	91	91	91	91	91	91	91	91	91	91
Archaea	88	87	88	88	87	87	87	87	87	87
Bacteria	93	93	93	92	92	92	92	92	92	92
Fungi	77	77	77	78	77	77	76	75	76	75
Invertebrates	78	76	75	75	75	74	74	74	74	73
Plants	74	73	73	73	73	71	72	70	71	70
Protozoa	80	79	79	78	77	77	77	77	78	78
Mammals	89	87	87	87	85	85	85	85	84	84
Other Vertebrates	79	78	77	80	81	80	81	81	80	80
Viruses	90	91	91	91	91	91	91	91	91	91

Alignment-free: Both co-tRNA and identical codon pairing percent overlap with the NCBI Taxonomy. Column 1 shows the taxonomic group analyzed. Columns 2-11 show the percent overlap with the NCBI Taxonomy from a phylogeny recovered using both co-tRNA and identical codon pairing with the respective window size 2-11. The highest percent overlap for each taxonomic group is highlighted.

Supplementary Table 11, Chapter 5:

Taxonomic Group	2	3	4	5	6	7	8	9	10	11
All	84	84	84	84	84	84	84	84	84	84
Archaea	81	81	82	81	81	81	81	81	80	81
Bacteria	86	86	86	86	86	86	86	86	86	86
Fungi	74	75	74	76	74	74	73	73	74	73
Invertebrates	69	67	67	65	66	65	64	65	64	64
Plants	65	65	65	65	63	62	62	61	62	61
Protozoa	70	69	70	69	67	67	67	67	67	67
Mammals	79	76	78	78	76	76	76	77	75	75
Other Vertebrates	67	66	65	66	66	67	67	68	67	66

Alignment-free: Both co-tRNA and identical codon pairing percent overlap with the OTL. Column 1 shows the taxonomic group analyzed. Columns 2-11 show the percent overlap with the OTL from a phylogeny recovered using both identical and co-tRNA codon pairing with the respective window size 2-11. The highest percent overlap for each taxonomic group is highlighted.

Supplementary Table 12, Chapter 5:

Taxonomic Group	2	3	4	5	6	7	8	9	10	11
All	90	90	90	90	90	90	90	90	90	90
Archaea	83	83	83	83	83	82	82	82	83	82
Bacteria	91	91	91	91	91	91	91	91	91	91
Fungi	72	72	72	73	72	71	71	70	71	70
Invertebrates	70	70	70	70	70	70	69	69	69	69
Plants	70	69	69	68	68	68	68	69	67	68
Protozoa	76	75	77	75	73	74	73	74	73	75
Mammals	87	87	86	87	85	86	87	84	84	84
Other Vertebrates	76	76	76	76	76	76	75	74	77	75
Viruses	90	90	90	90	90	90	90	90	90	90

Alignment-free: Co-tRNA codon pairing percent overlap with the NCBI Taxonomy. Column 1 shows the taxonomic group analyzed. Columns 2-11 show the percent overlap with the NCBI Taxonomy from a phylogeny recovered using co-tRNA codon pairing with the respective window size 2-11. The highest percent overlap for each taxonomic group is highlighted.

Supplementary Table 13, Chapter 5:

Taxonomic Group	2	3	4	5	6	7	8	9	10	11
All	82	82	82	82	82	82	82	82	82	82
Archaea	75	76	75	76	76	75	75	75	76	76
Bacteria	84	84	85	85	85	85	85	85	85	85
Fungi	69	70	71	70	70	69	69	69	69	69
Invertebrates	62	61	61	61	60	60	60	61	60	59
Plants	62	60	60	60	59	59	59	60	59	59
Protozoa	70	68	69	68	65	66	65	66	65	66
Mammals	77	77	76	77	76	77	77	76	75	75
Other Vertebrates	66	65	63	64	64	65	65	64	65	63

Alignment-free: Co-tRNA codon pairing percent overlap with the OTL. Column 1 shows the taxonomic group analyzed. Columns 2-11 show the percent overlap with the OTL from a phylogeny recovered using co-tRNA codon pairing with the respective window size 2-11. The highest percent overlap for each taxonomic group is highlighted.

Parsimony

Supplementary Table 14, Chapter 5:

Taxonomic Group	2	3	4	5	6	7	8	9	10	11
All	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Archaea	84	83	87	88	86	85	86	84	85	84
Bacteria	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Fungi	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Invertebrates	58	63	67	65	67	64	61	65	62	63
Plants	80	80	82	82	81	82	84	79	79	79
Protozoa	81	81	81	77	73	77	77	68	81	81
Mammals	96	96	95	94	93	92	93	94	94	93
Other Vertebrates	91	91	90	90	91	90	90	91	90	91
Viruses	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Parsimony: Identical codon pairing percent overlap with the NCBI Taxonomy. Column 1 shows the taxonomic group analyzed. Columns 2-11 show the percent overlap with the NCBI Taxonomy from a phylogeny recovered using identical codon pairing with the respective window size 2-11. The highest percent overlap for each taxonomic group is highlighted.

Supplementary Table 15, Chapter 5:

Taxonomic Group	2	3	4	5	6	7	8	9	10	11
All	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Archaea	84	84	86	86	86	86	85	84	85	84
Bacteria	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Fungi	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Invertebrates	55	58	62	61	64	59	57	61	59	60
Plants	77	75	77	78	78	77	79	75	74	75
Protozoa	68	68	68	66	63	66	66	58	68	68
Mammals	90	89	87	86	86	86	86	88	87	87
Other Vertebrates	77	76	75	76	76	76	75	76	77	77

Parsimony: Identical codon pairing percent overlap with the OTL. Column 1 shows the taxonomic group analyzed. Columns 2-11 show the percent overlap with the OTL from a phylogeny recovered using identical codon pairing with the respective window size 2-11. The highest percent overlap for each taxonomic group is highlighted.

Supplementary Table 16, Chapter 5:

Taxonomic Group	2	3	4	5	6	7	8	9	10	11
All	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Archaea	87	90	89	89	88	87	88	88	89	87
Bacteria	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Fungi	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Invertebrates	63	65	71	63	63	63	63	63	62	64
Plants	77	77	79	77	84	81	78	80	77	78
Protozoa	87	85	67	66	76	75	67	67	61	66
Mammals	95	92	95	91	94	92	92	93	94	94
Other Vertebrates	93	94	92	93	93	93	91	91	91	91
Viruses	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Parsimony: Combined codon pairing percent overlap with the NCBI Taxonomy. Column 1 shows the taxonomic group analyzed. Columns 2-11 show the percent overlap with the NCBI Taxonomy from a phylogeny recovered using co-tRNA codon pairing with the respective window size 2-11. The highest percent overlap for each taxonomic group is highlighted.

Supplementary Table 17, Chapter 5:

Taxonomic Group	2	3	4	5	6	7	8	9	10	11
All	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Archaea	86	87	90	89	89	86	88	88	87	86
Bacteria	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Fungi	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Invertebrates	57	61	66	57	56	58	59	59	58	57
Plants	73	71	74	72	77	75	73	74	72	72
Protozoa	73	71	57	57	66	66	59	59	54	60
Mammals	87	86	86	85	87	85	84	85	85	85
Other Vertebrates	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Parsimony: Combined codon pairing percent overlap with the OTL. Column 1 shows the taxonomic group analyzed. Columns 2-11 show the percent overlap with the OTL from a phylogeny recovered using co-tRNA codon pairing with the respective window size 2-11. The highest percent overlap for each taxonomic group is highlighted.

Supplementary Table 18, Chapter 5:

Taxonomic Group	2	3	4	5	6	7	8	9	10	11
All	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Archaea	88	88	86	86	83	84	84	83	85	82
Bacteria	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Fungi	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Invertebrates	59	60	59	60	59	58	59	59	58	59
Plants	76	82	80	80	80	82	81	80	84	80
Protozoa	38	77	74	75	86	74	86	77	81	81
Mammals	94	92	91	92	93	90	94	92	92	92
Other Vertebrates	92	91	91	91	90	90	91	91	90	90
Viruses	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Parsimony: Co-tRNA codon pairing percent overlap with the NCBI Taxonomy. Column 1 shows the taxonomic group analyzed. Columns 2-11 show the percent overlap with the NCBI Taxonomy from a phylogeny recovered using co-tRNA codon pairing with the respective window size 2-11. The highest percent overlap for each taxonomic group is highlighted.

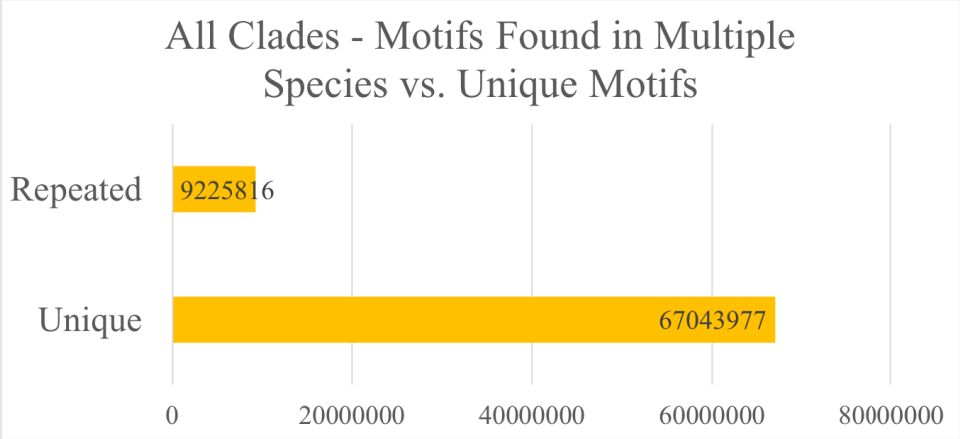
Supplementary Table 19, Chapter 5:

Taxonomic Group	2	3	4	5	6	7	8	9	10	11
All	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Archaea	89	88	86	87	83	85	85	82	83	84
Bacteria	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Fungi	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Invertebrates	53	54	54	54	53	55	54	53	53	53
Plants	73	75	77	76	79	79	79	79	80	78
Protozoa	59	66	63	62	72	65	72	66	69	69
Mammals	86	85	85	86	85	84	86	84	85	84
Other Vertebrates	76	76	77	76	75	76	76	76	76	76

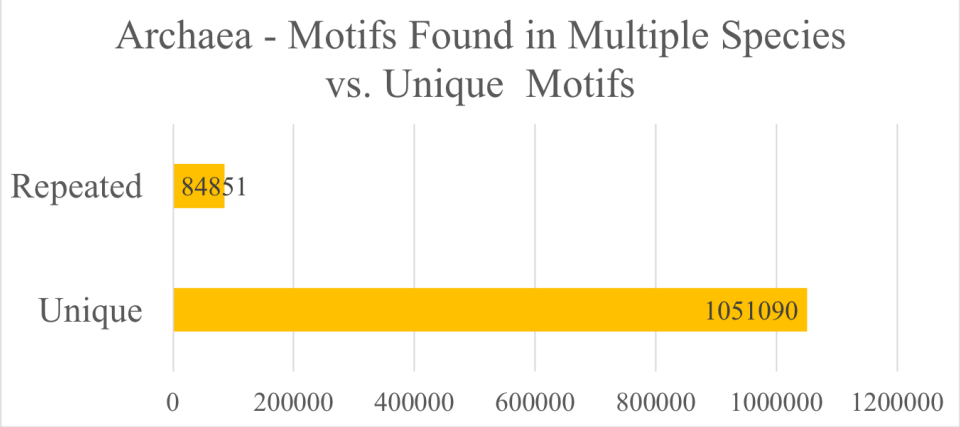
Parsimony: Co-tRNA codon pairing percent overlap with the OTL. Column 1 shows the taxonomic group analyzed. Columns 2-11 show the percent overlap with the OTL from a phylogeny recovered using co-tRNA codon pairing with the respective window size 2-11. The highest percent overlap for each taxonomic group is highlighted.

Supplementary Figures for Chapter 5

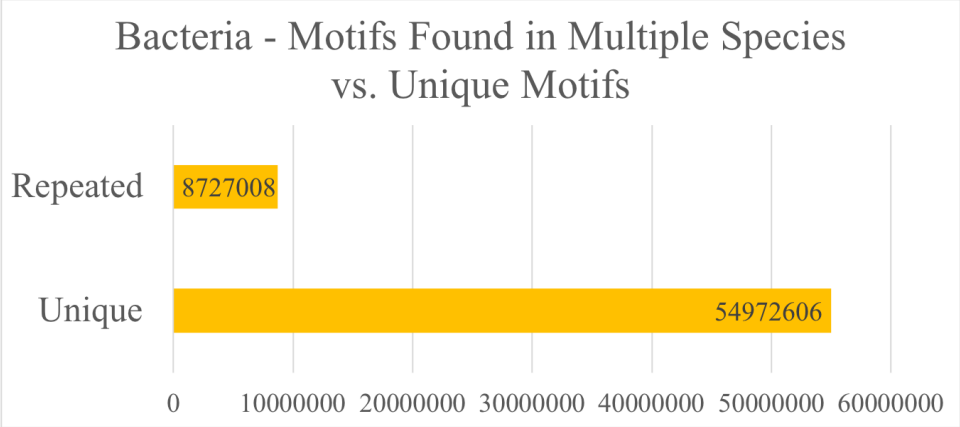
Supplementary Figure 1, Chapter 5:



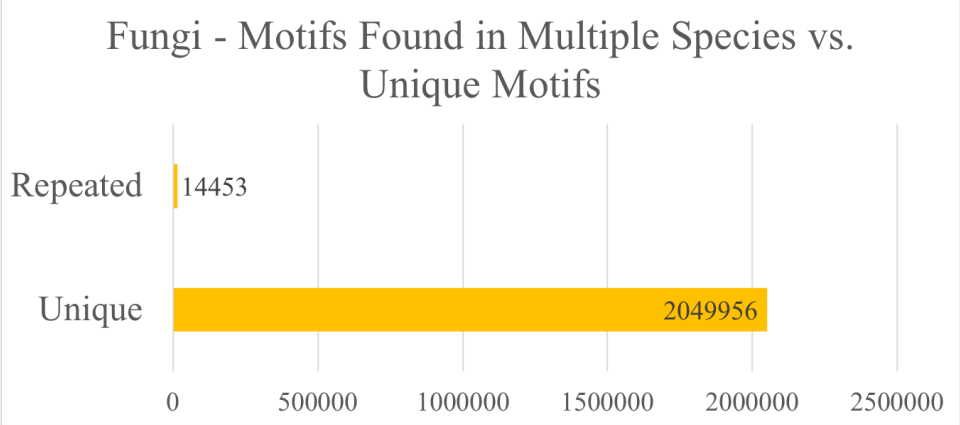
Supplementary Figure 2, Chapter 5:



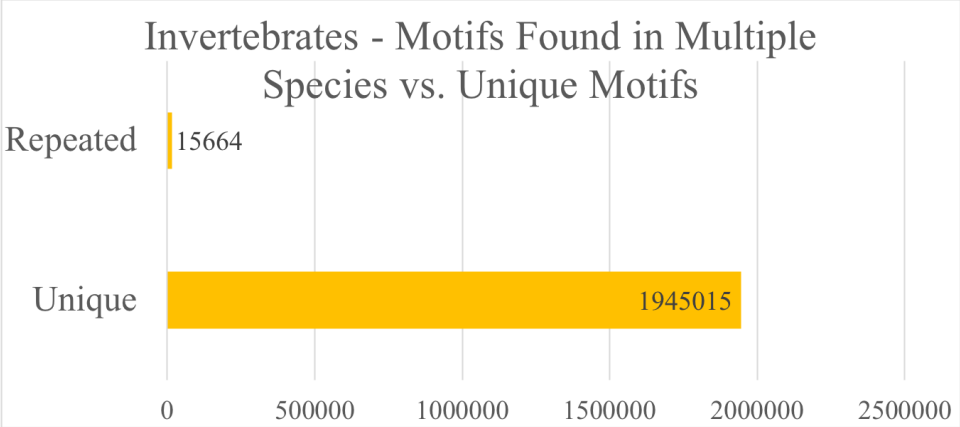
Supplementary Figure 3, Chapter 5:



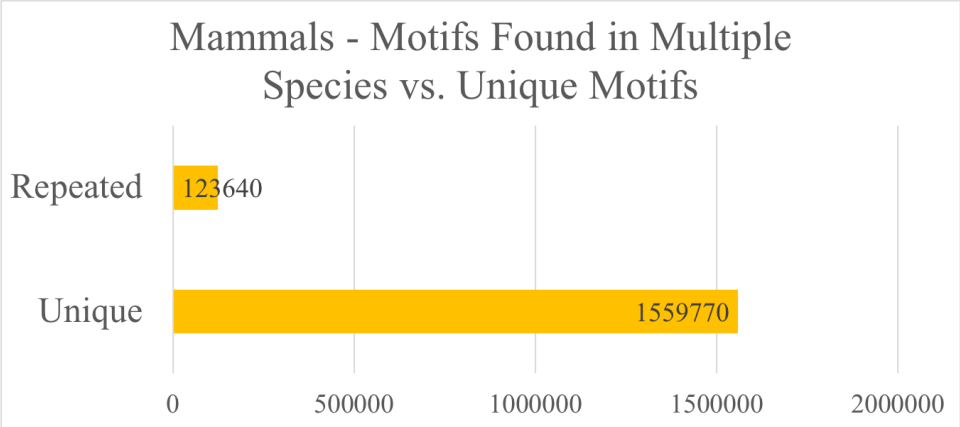
Supplementary Figure 4, Chapter 5:



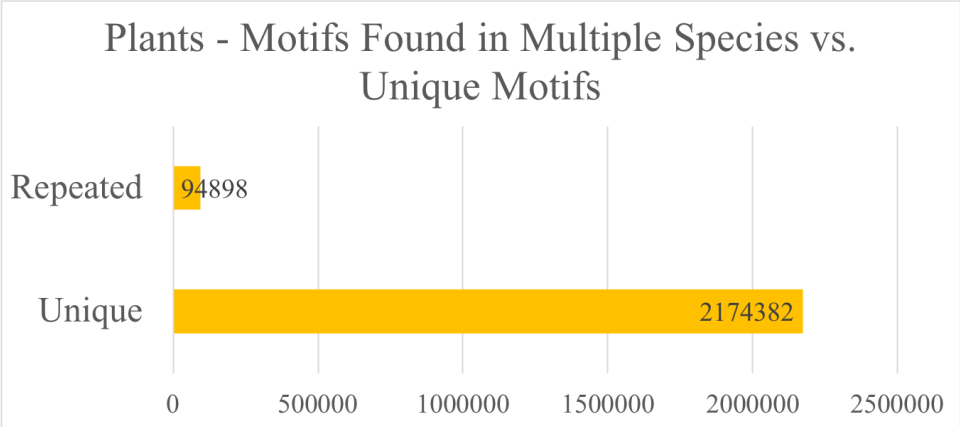
Supplementary Figure 5, Chapter 5:



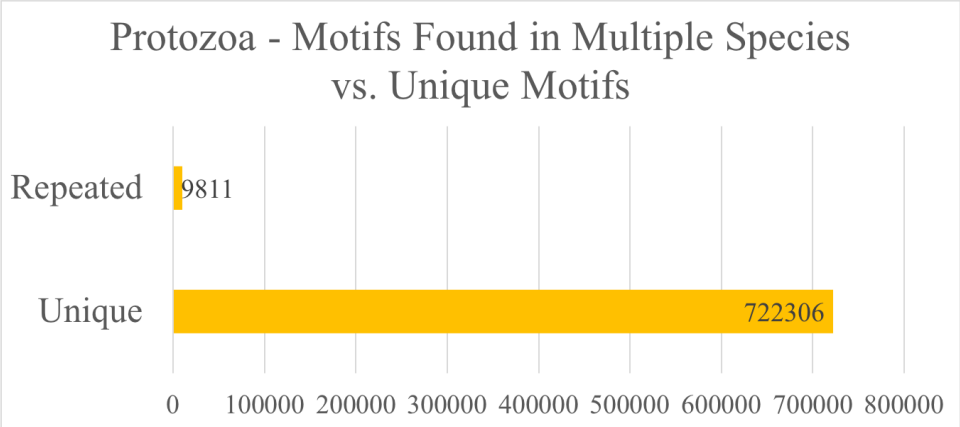
Supplementary Figure 6, Chapter 5:



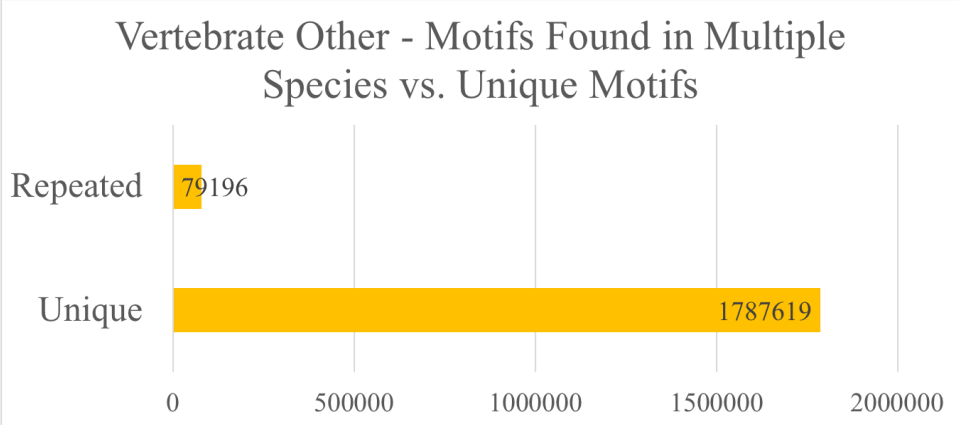
Supplementary Figure 7, Chapter 5:



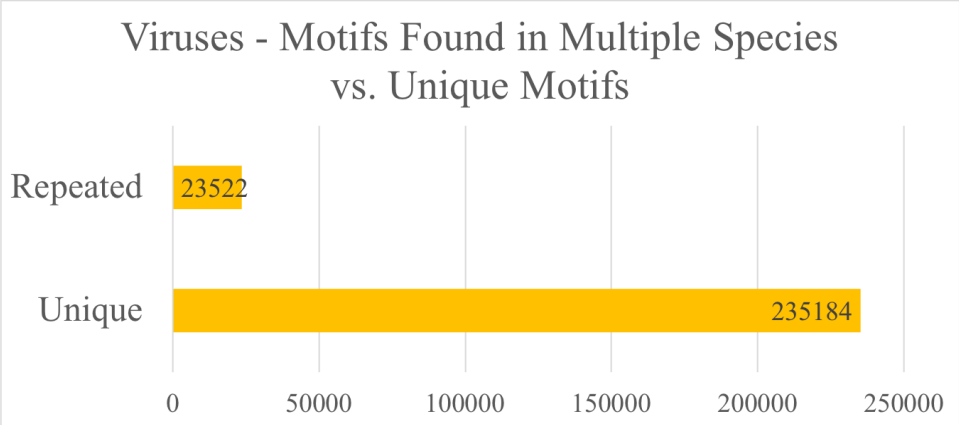
Supplementary Figure 8, Chapter 5:



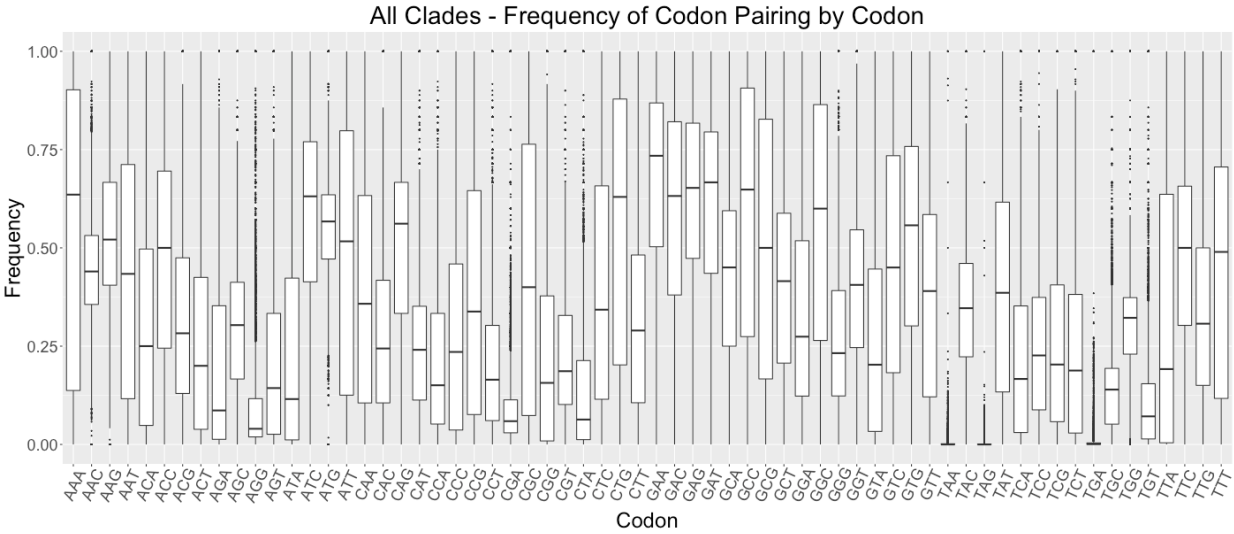
Supplementary Figure 9, Chapter 5:



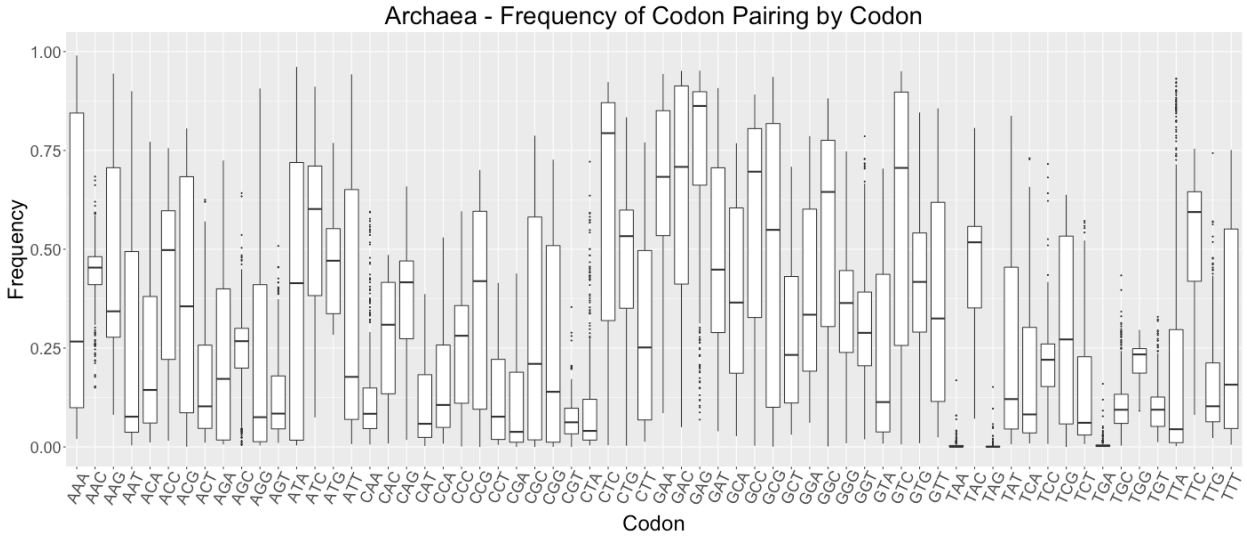
Supplementary Figure 10, Chapter 5:



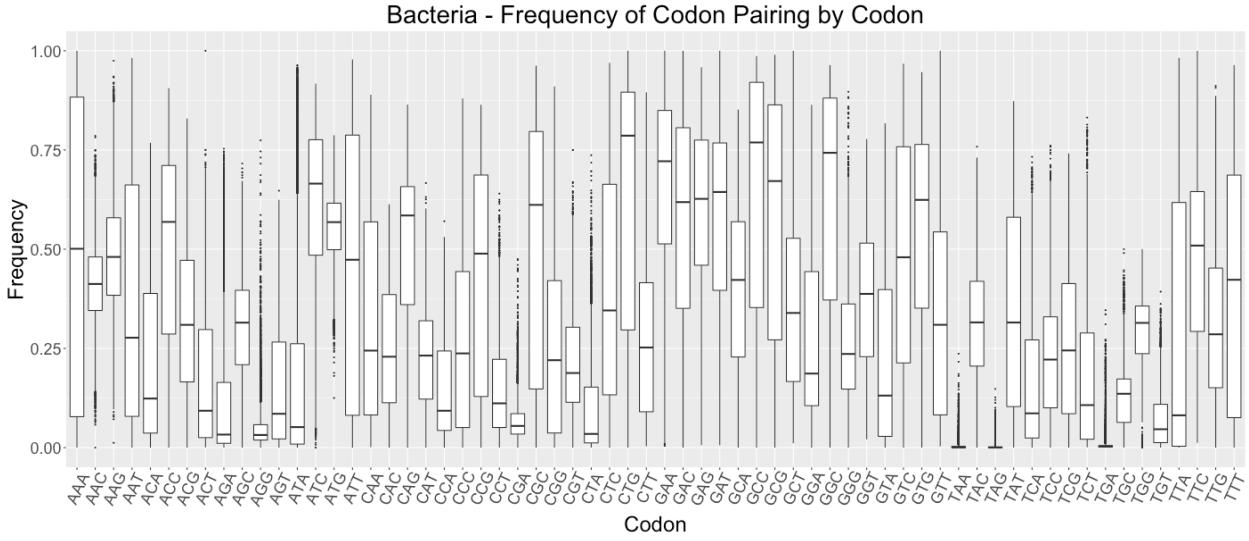
Supplementary Figure 11, Chapter 5:



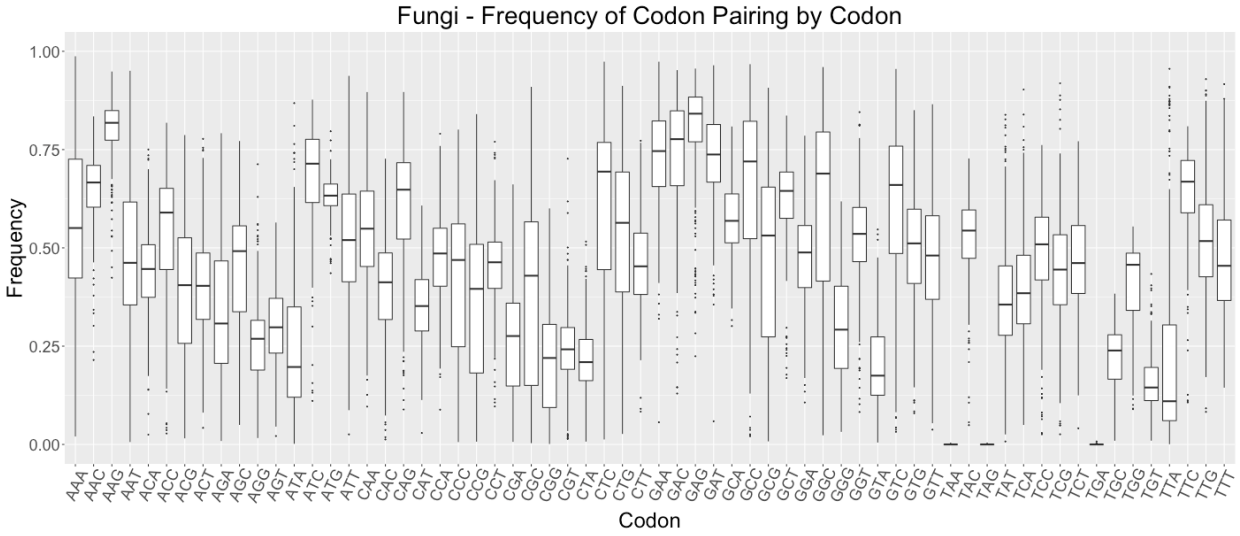
Supplementary Figure 12, Chapter 5:



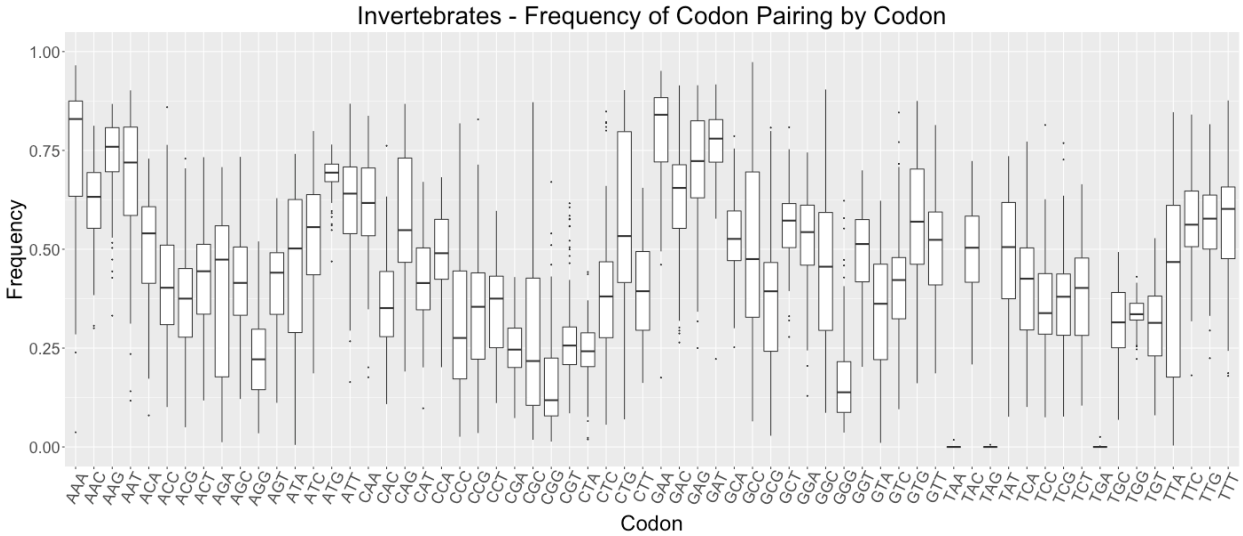
Supplementary Figure 13, Chapter 5:



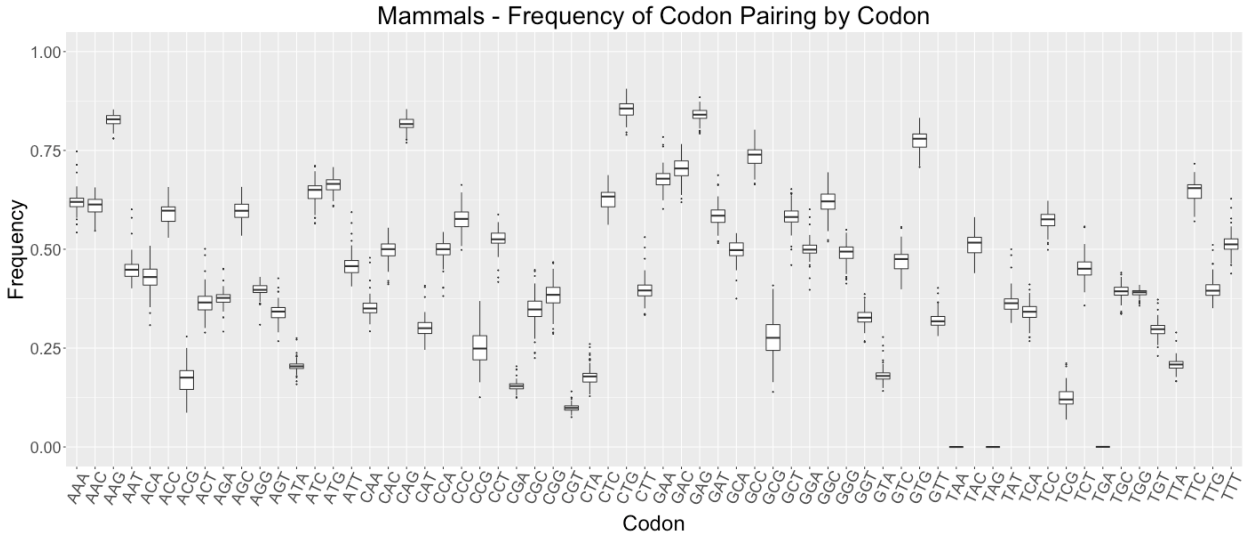
Supplementary Figure 14, Chapter 5:



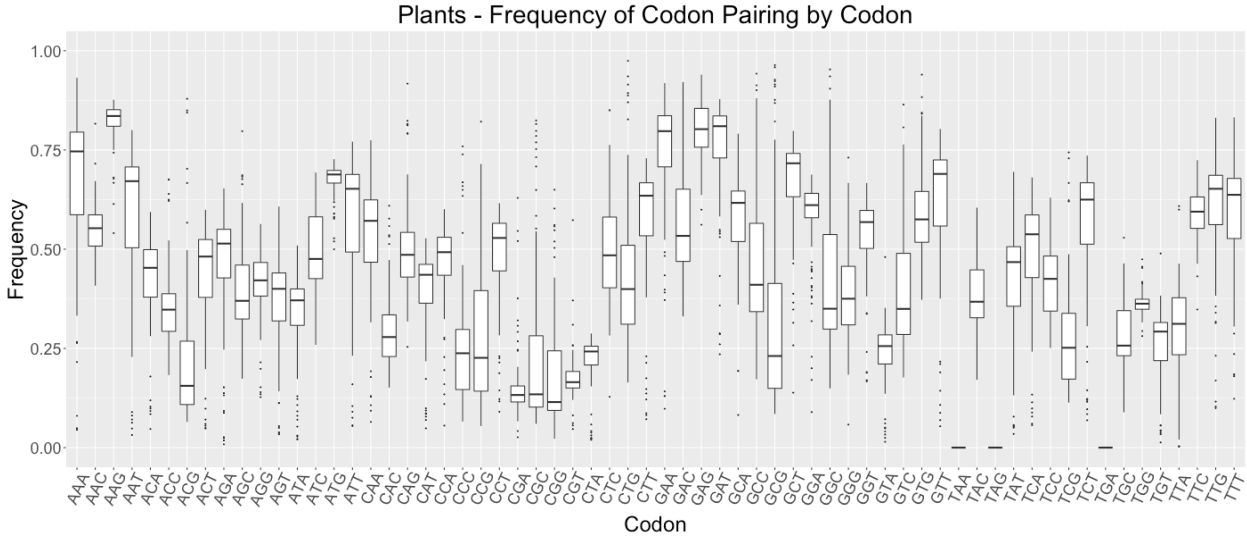
Supplementary Figure 15, Chapter 5:



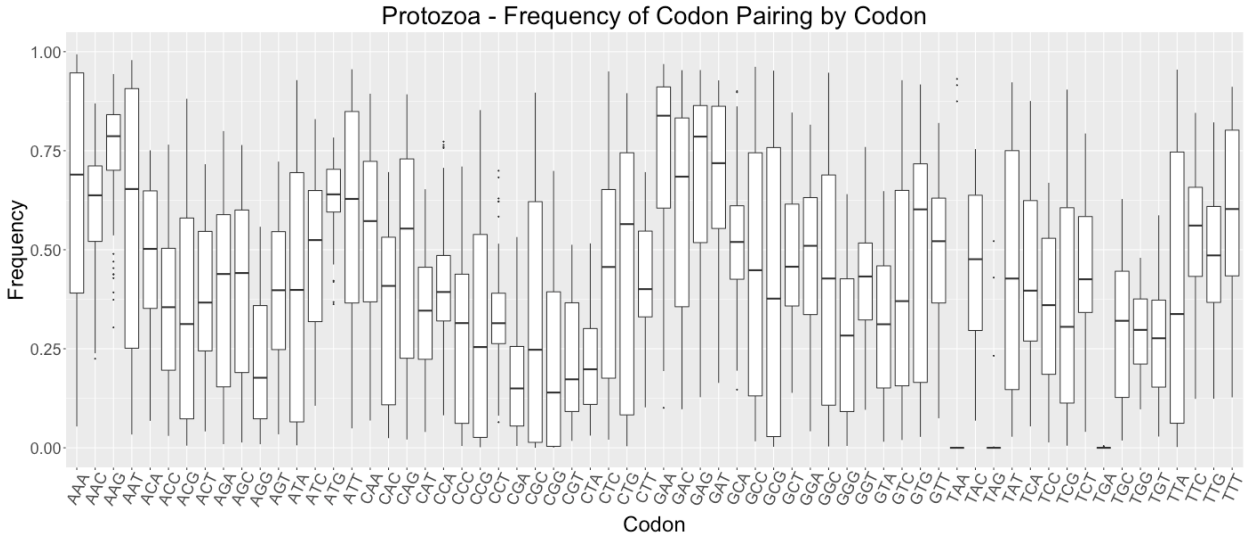
Supplementary Figure 16, Chapter 5:



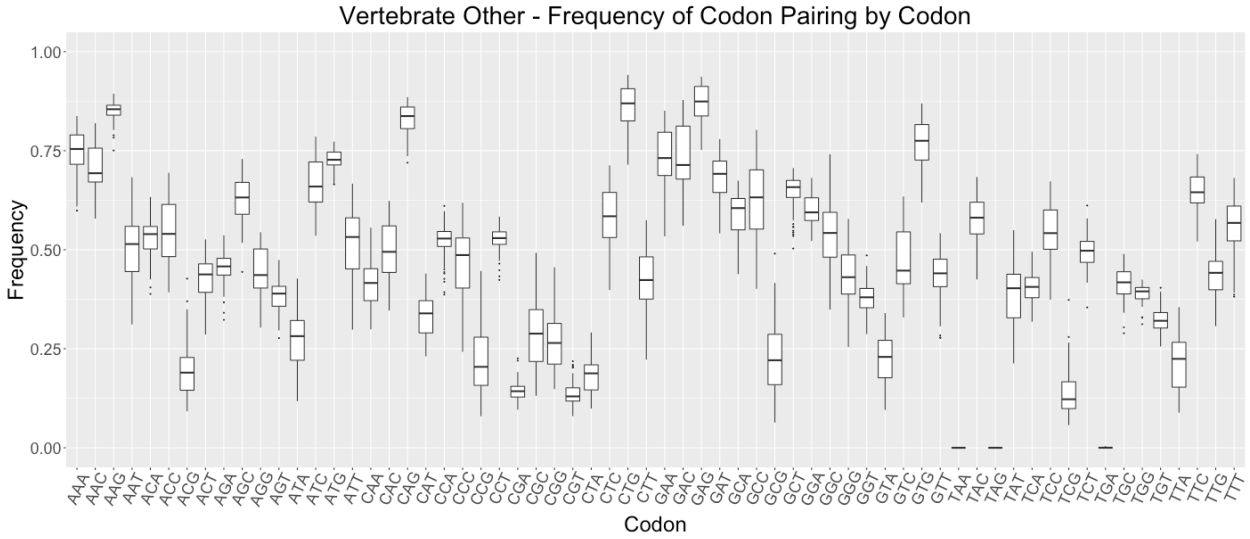
Supplementary Figure 17, Chapter 5:



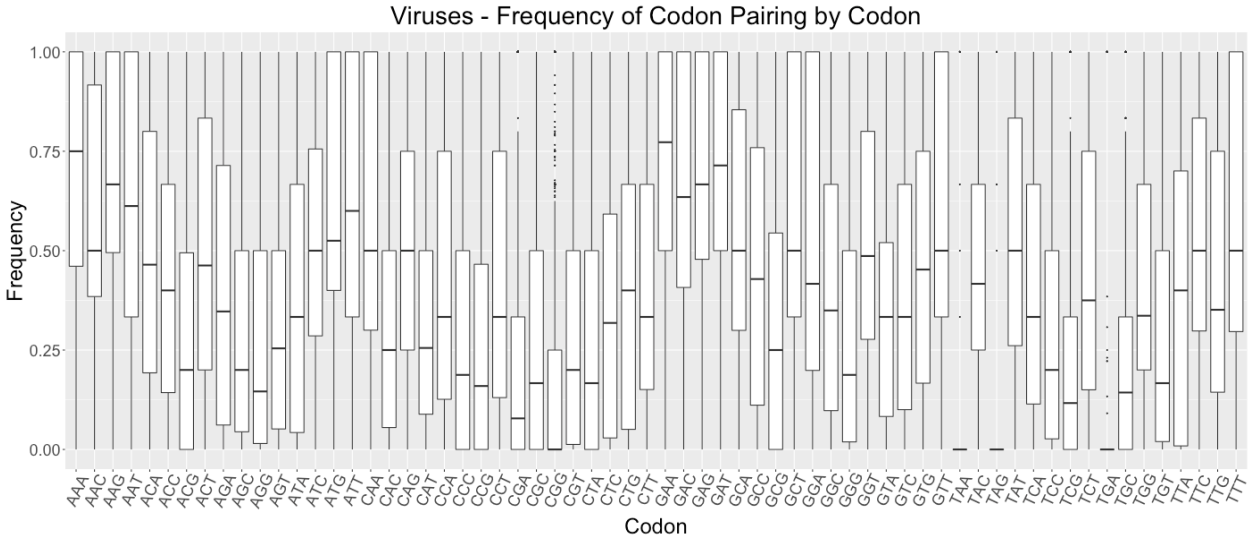
Supplementary Figure 18, Chapter 5:



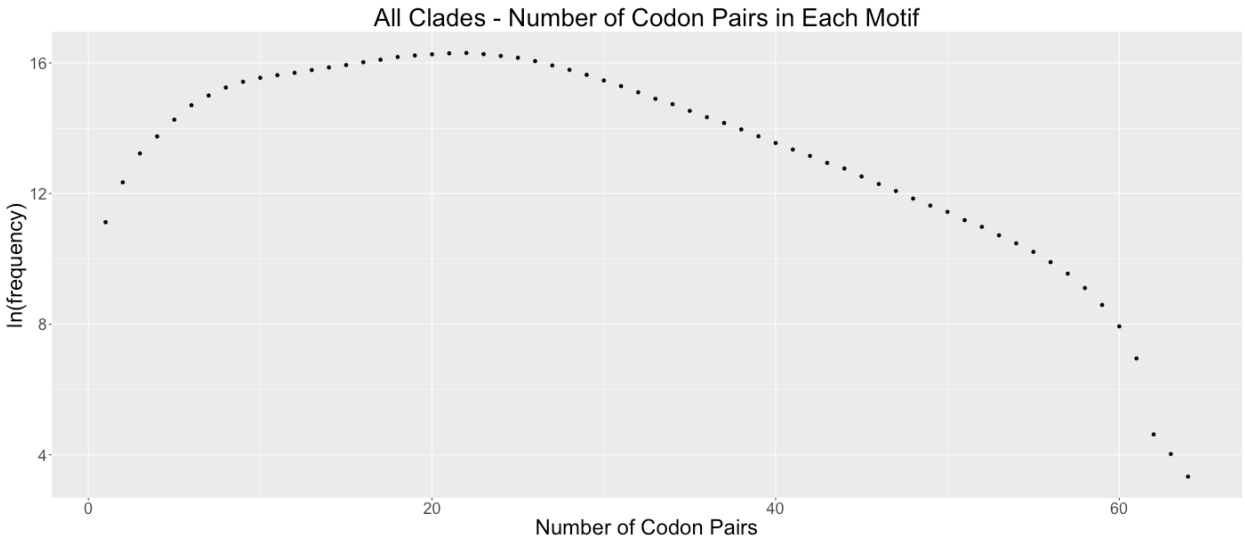
Supplementary Figure 19, Chapter 5:



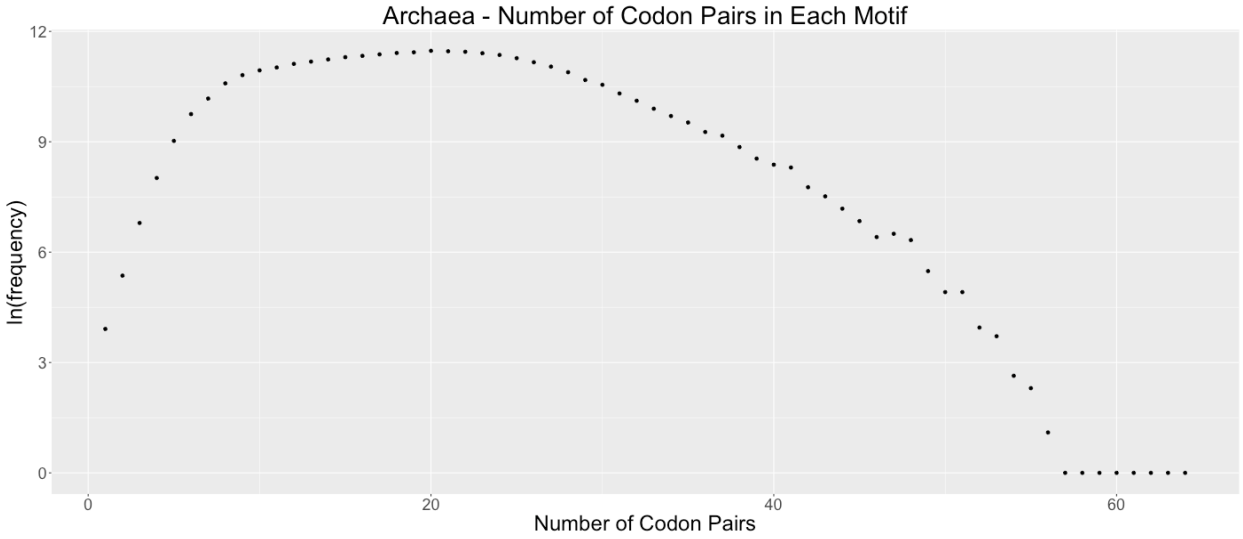
Supplementary Figure 20, Chapter 5:



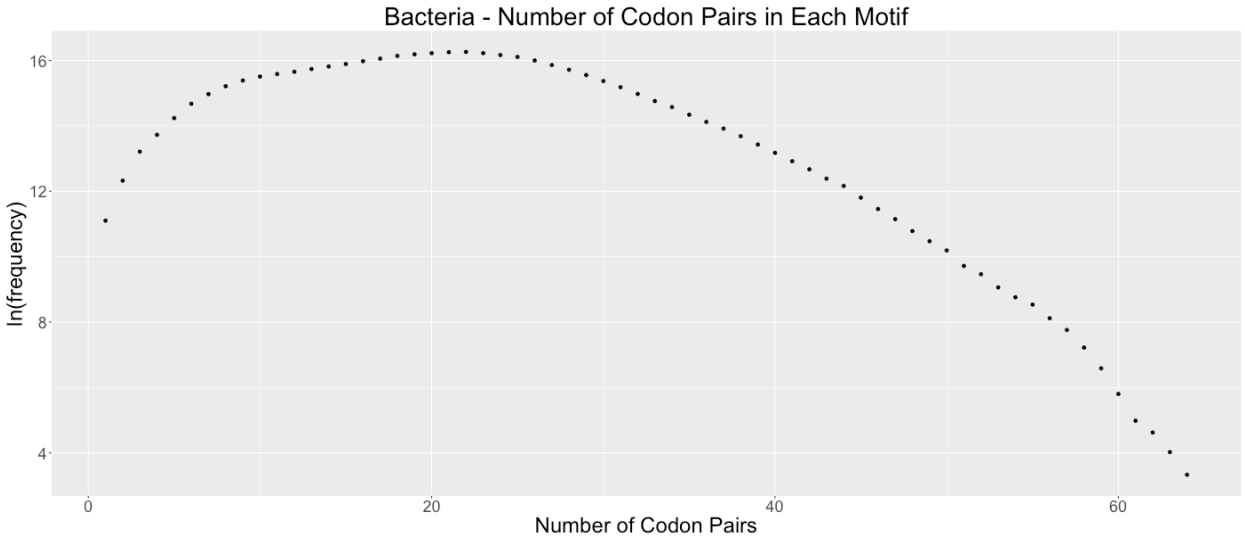
Supplementary Figure 21, Chapter 5:



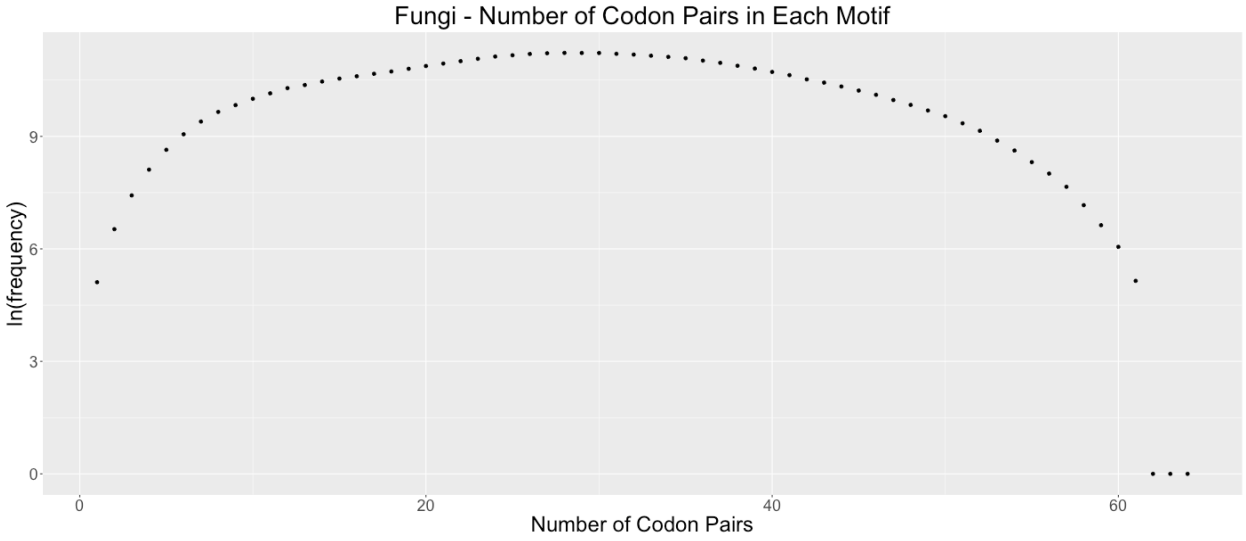
Supplementary Figure 22, Chapter 5:



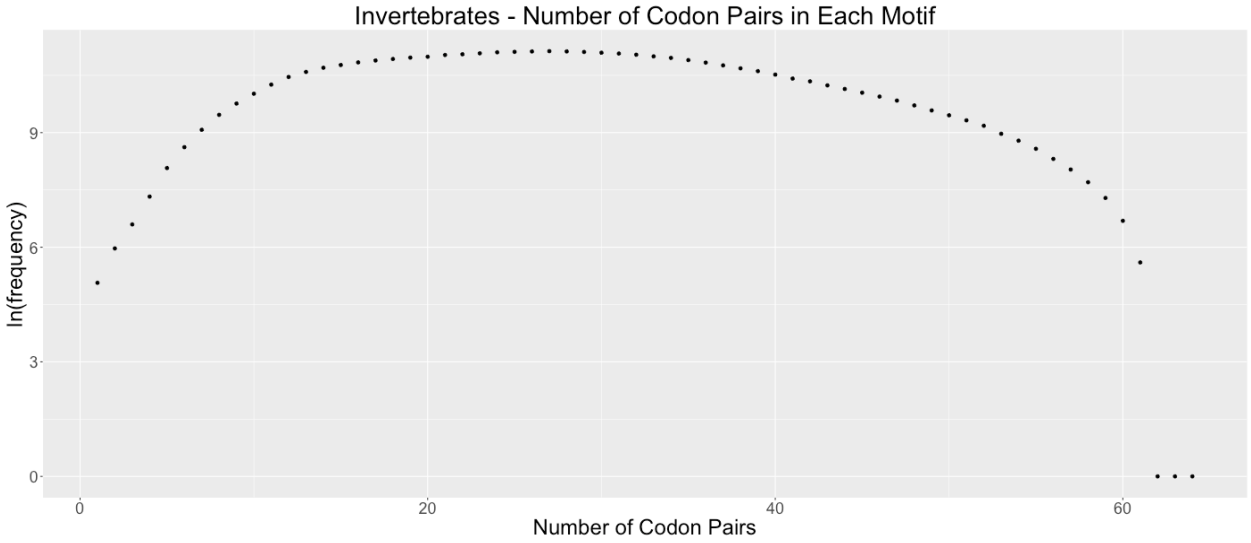
Supplementary Figure 23, Chapter 5:



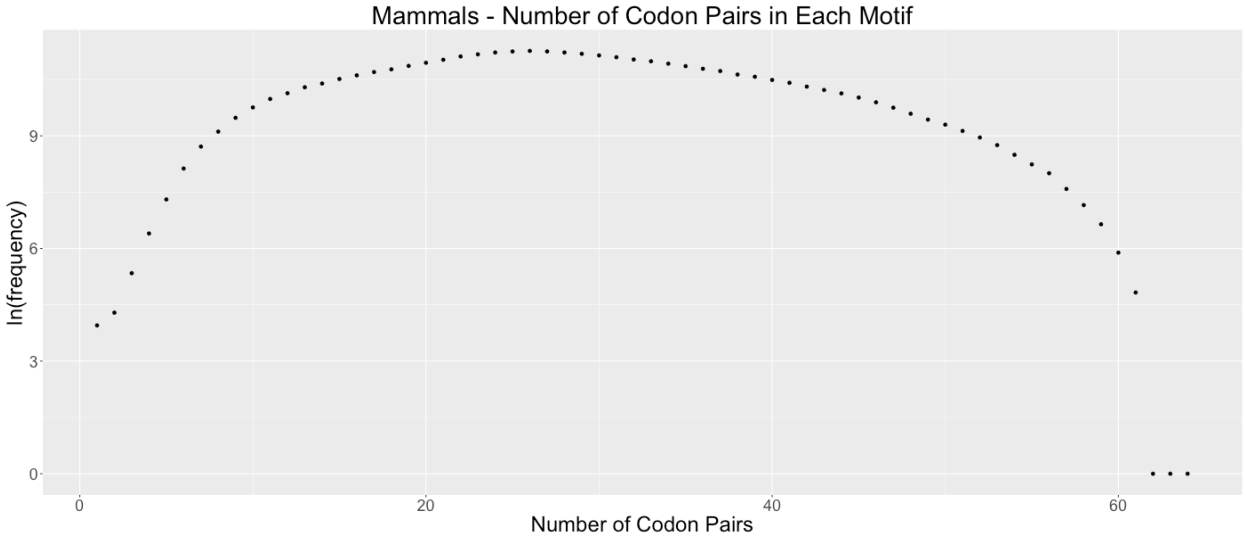
Supplementary Figure 24, Chapter 5:



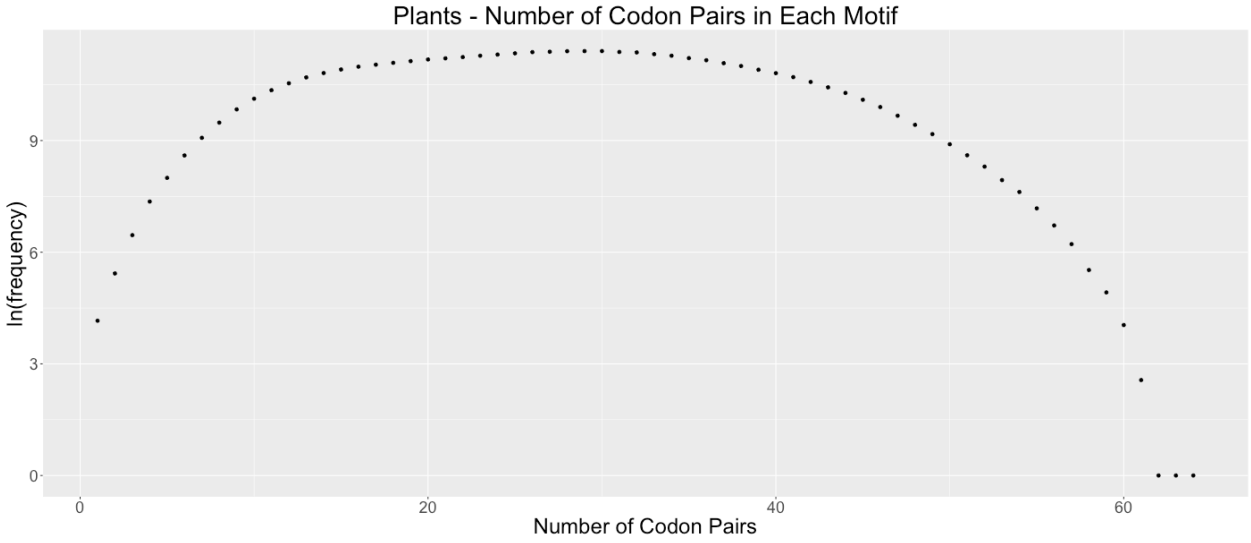
Supplementary Figure 25, Chapter 5:



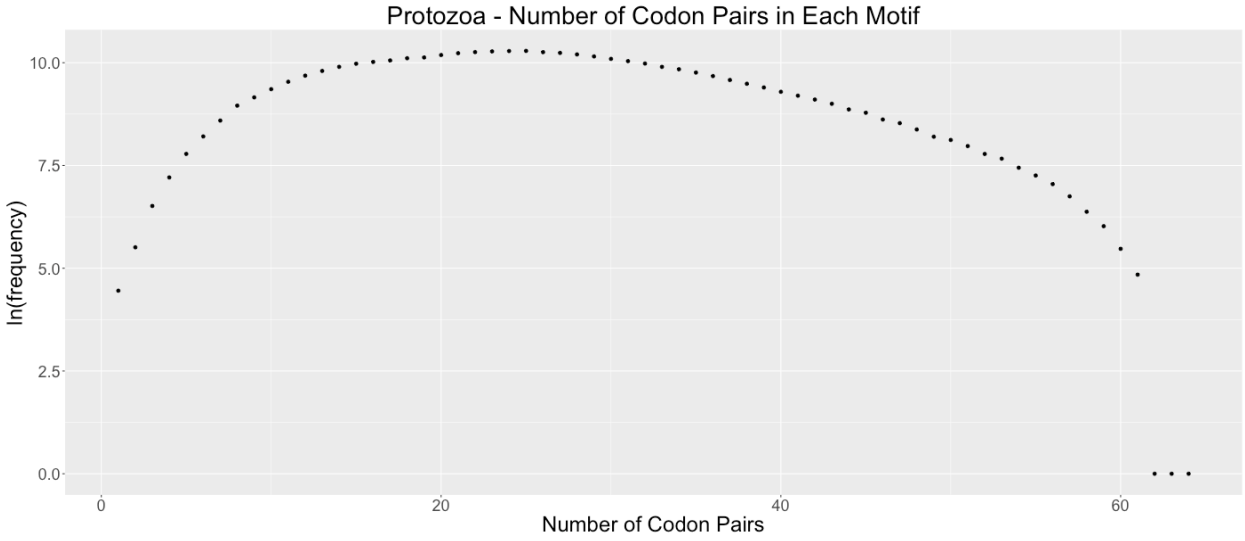
Supplementary Figure 26, Chapter 5:



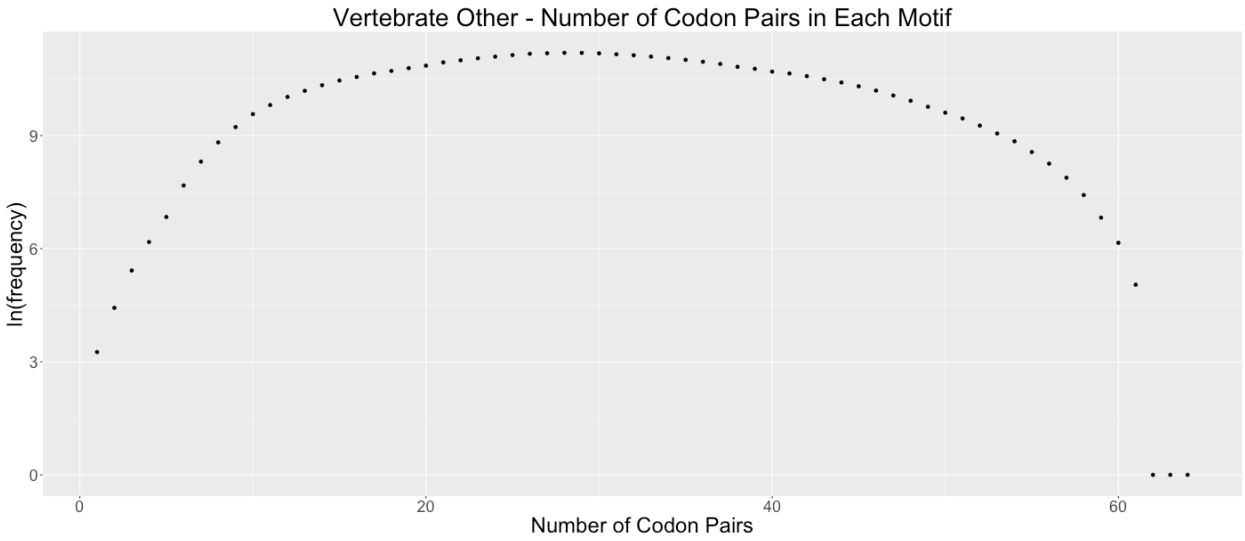
Supplementary Figure 27, Chapter 5:



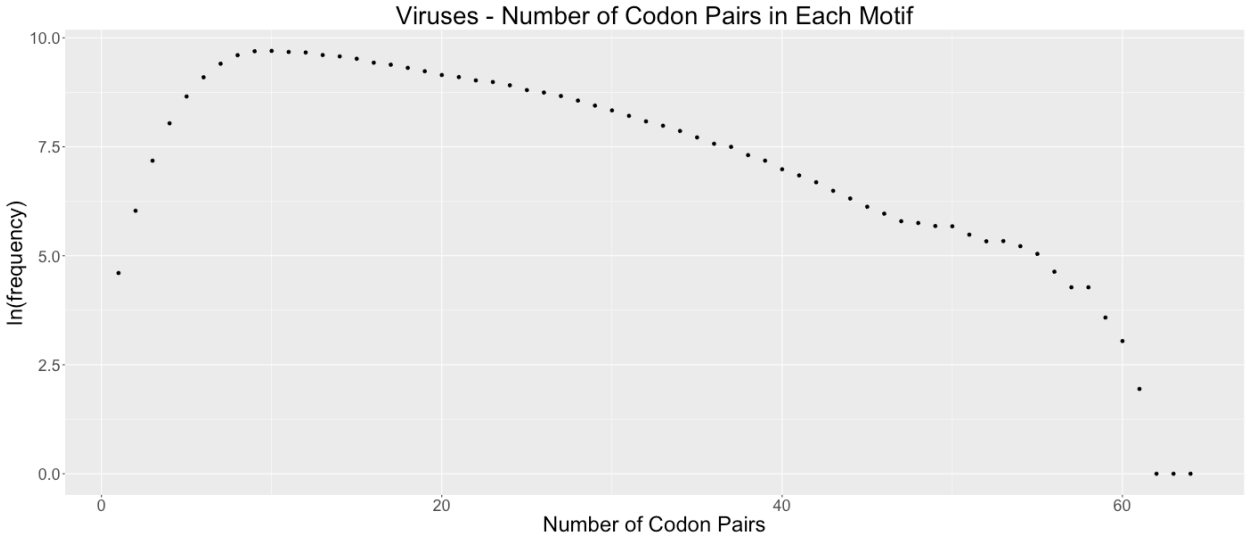
Supplementary Figure 28, Chapter 5:



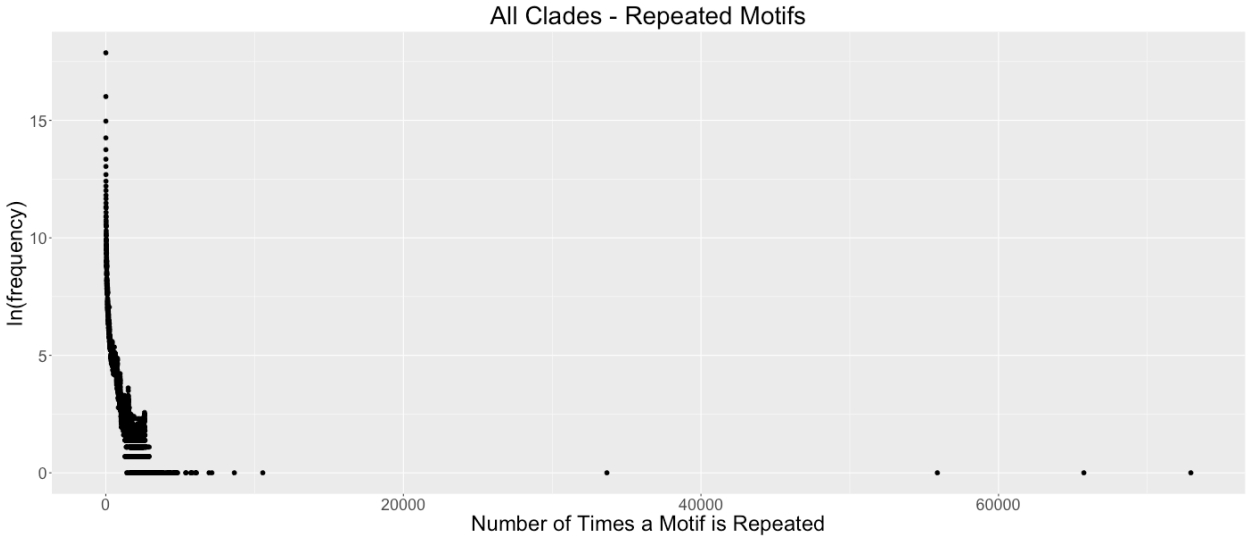
Supplementary Figure 29, Chapter 5:



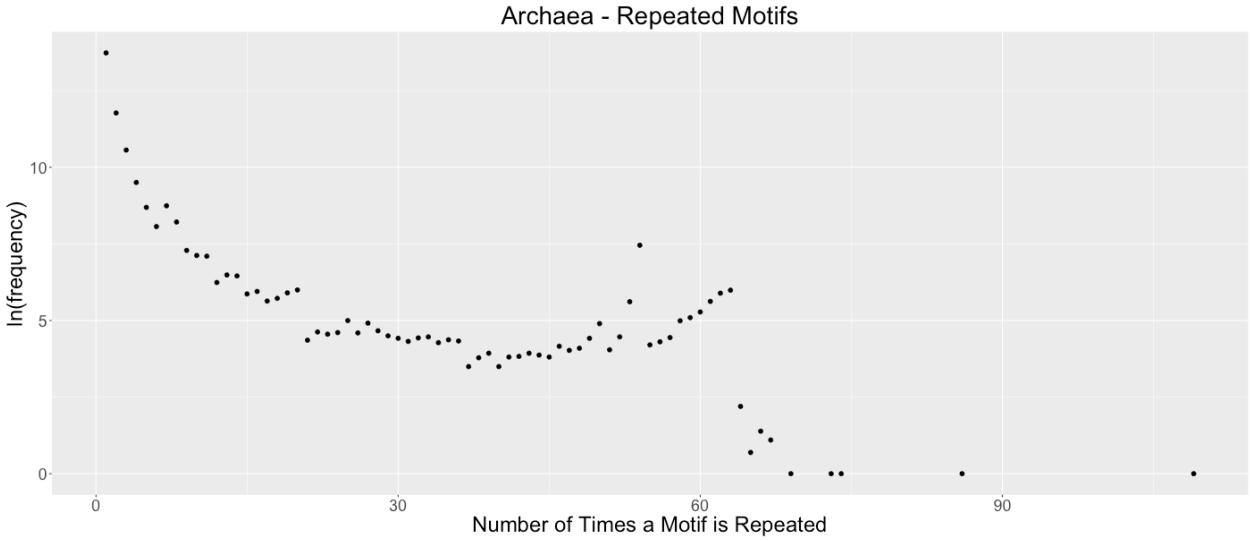
Supplementary Figure 30, Chapter 5:



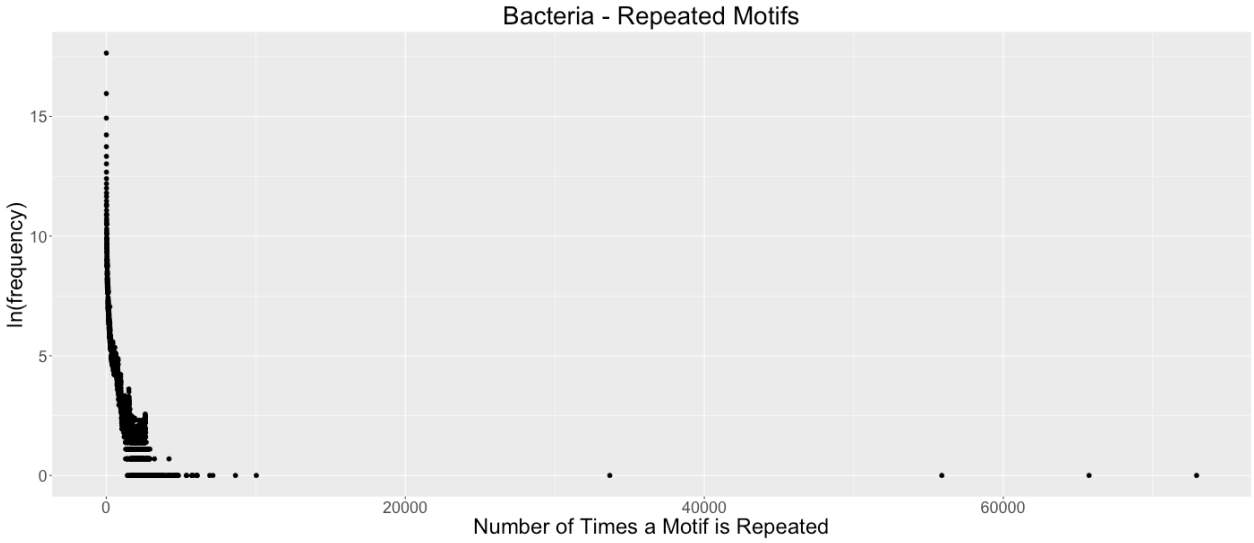
Supplementary Figure 31, Chapter 5:



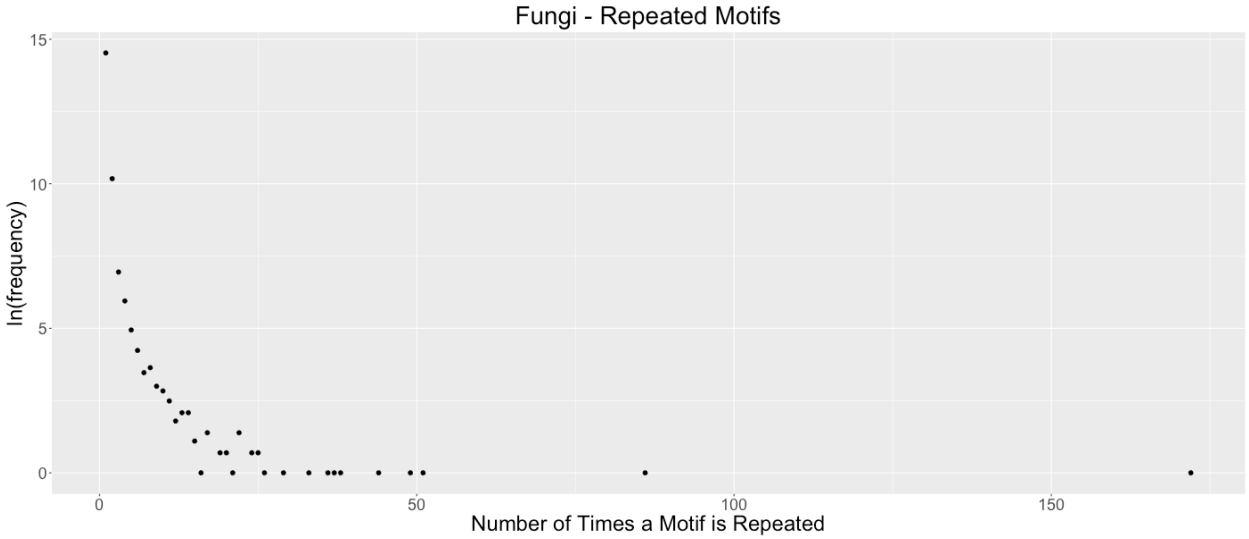
Supplementary Figure 32, Chapter 5:



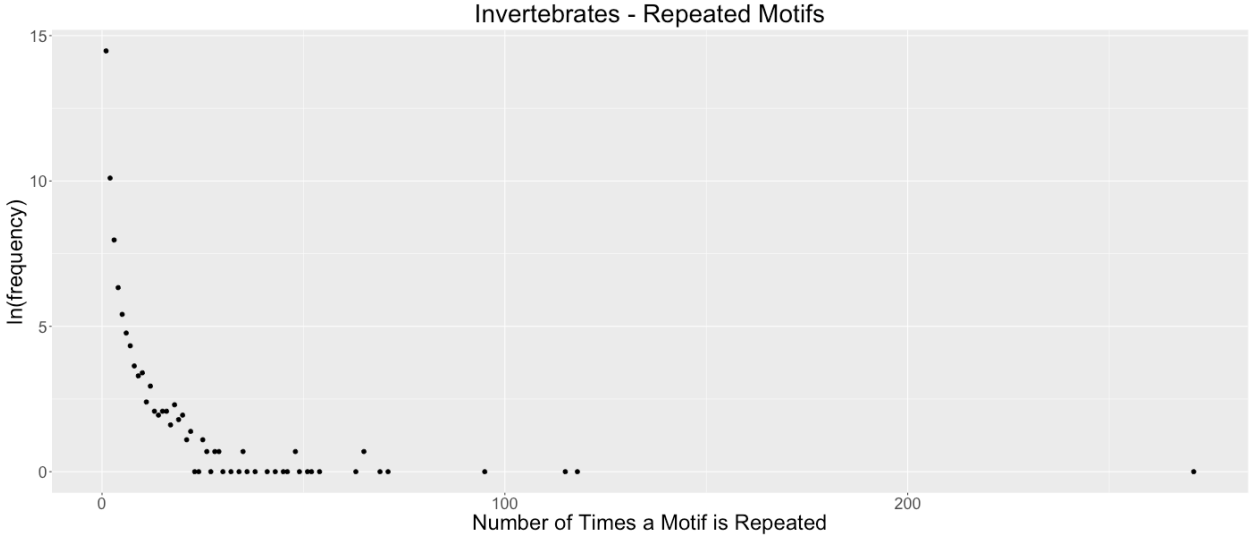
Supplementary Figure 33, Chapter 5:



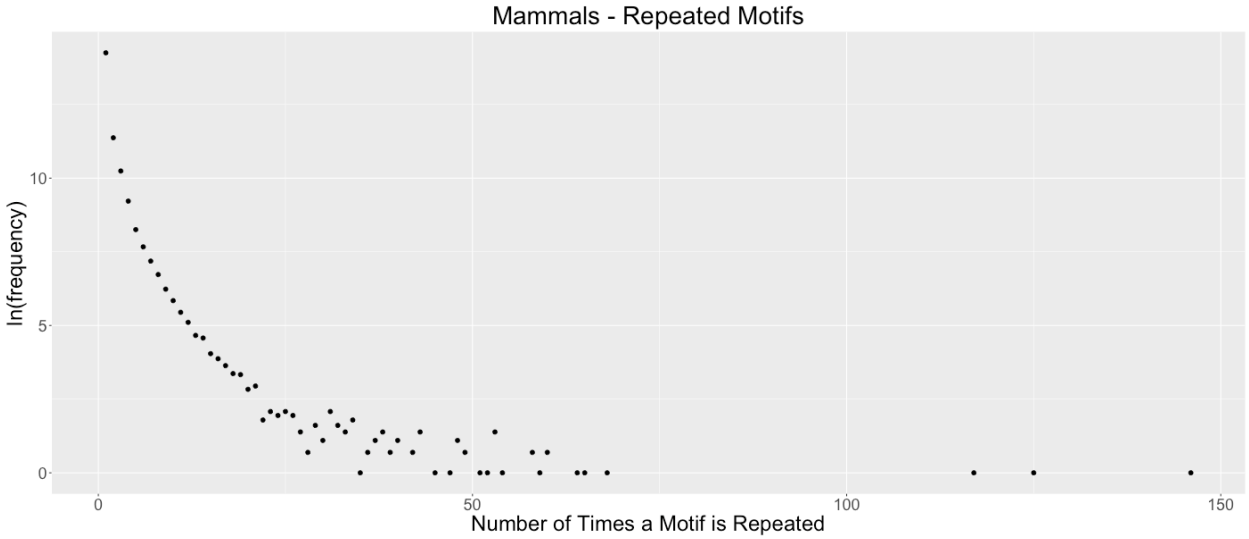
Supplementary Figure 34, Chapter 5:



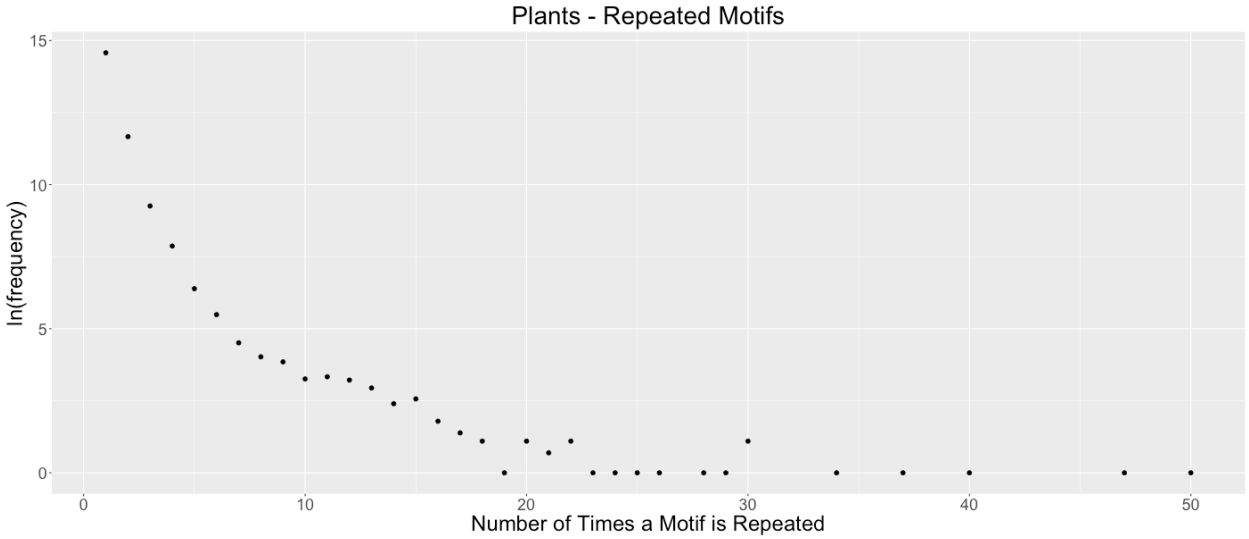
Supplementary Figure 35, Chapter 5:



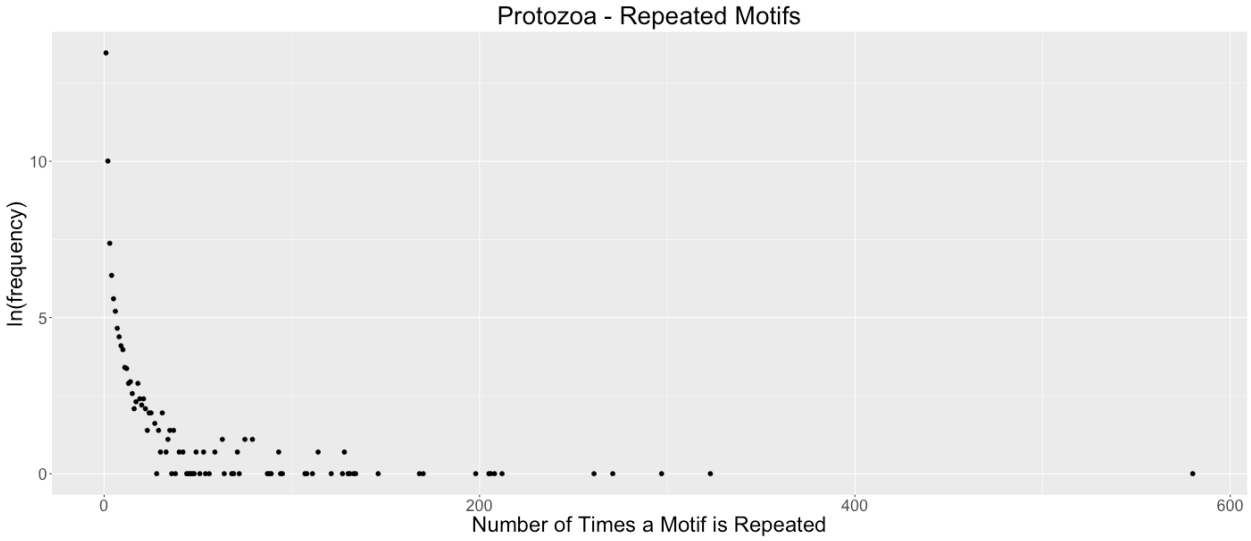
Supplementary Figure 36, Chapter 5:



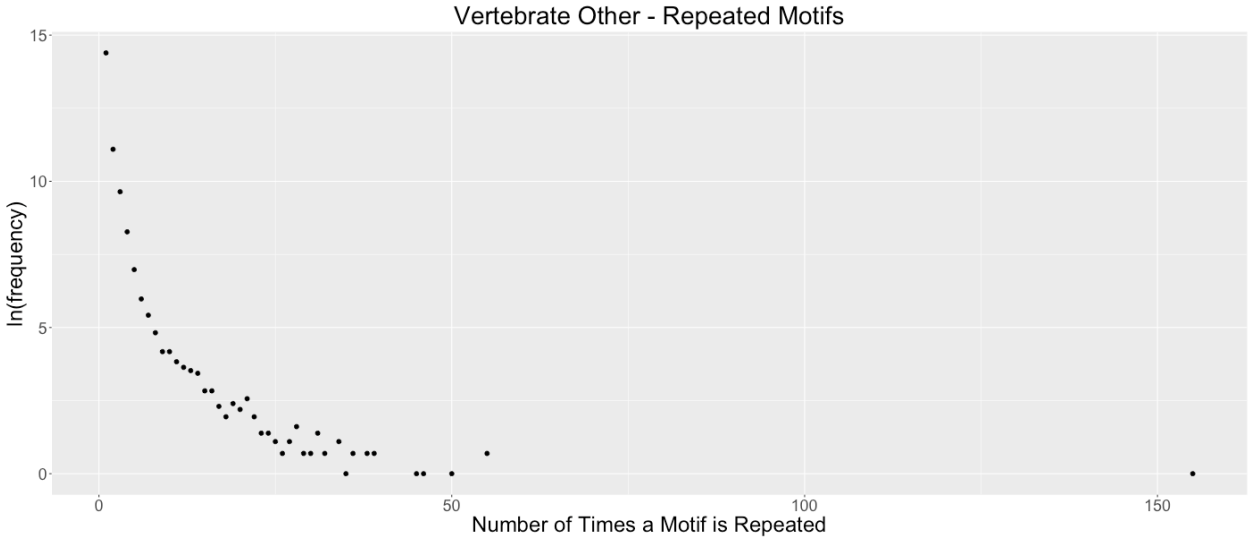
Supplementary Figure 37, Chapter 5:



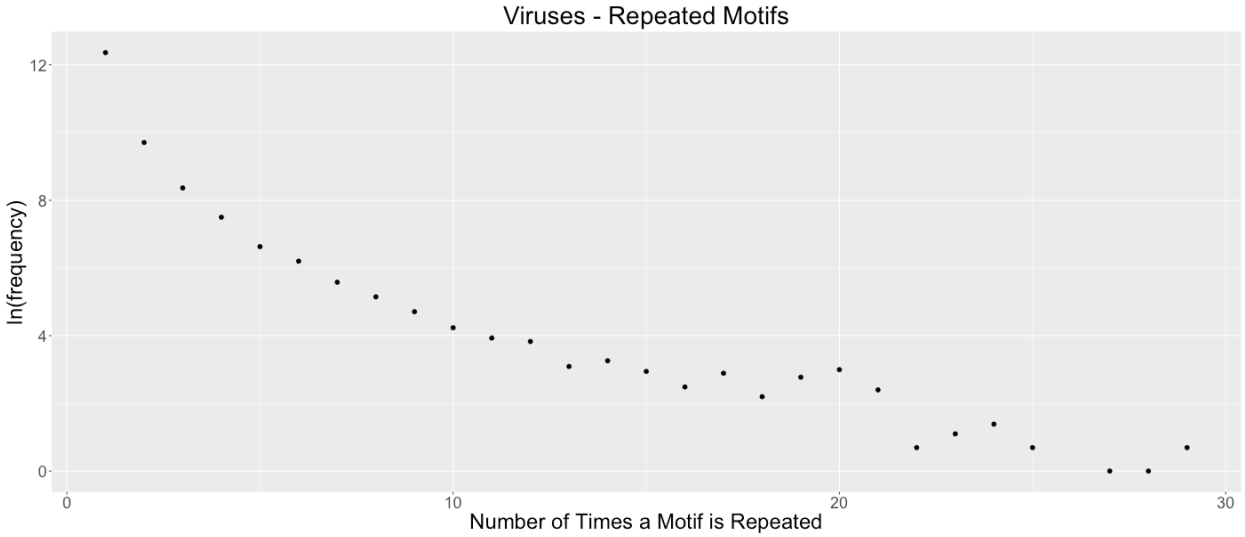
Supplementary Figure 38, Chapter 5:



Supplementary Figure 39, Chapter 5:



Supplementary Figure 40, Chapter 5:



Appendix 5: Supplementary Figures and Tables for Chapter 6

Table S1, Chapter 6. A comprehensive list of the 113 viruses with their highest correlating protein, accompanied by the Pearson's r correlation and the respective p-value. Bolded rows were found to be insignificant. Unnamed viral proteins are designated by their position numbers in the following format— Pos: start position-stop position.

Virus Accession Number	Virus Protein Name	Pearson's R Correlation Value	P-value	Highest Correlating Protein Accession Number	Protein Common Name
NC_000883	NS1	0.764596741	1.94E-13	NP_002763.2	TMPRSS15
NC_000898	U90	0.931483267	6.4E-29	NP_112561.2	TEX15
NC_001348	ICP4	0.798569441	2.68E-15	NP_787081.2	FAM181B
NC_001352	E1	0.725454272	1.2E-11	NP_037485.2	TMOD4
NC_001354	Pos: 951-2795	0.804857764	1.11E-15	NP_001273387.1	USP7
NC_001355	E1	0.798328333	2.77E-15	NP_940841.1	KBTBD3
NC_001356	E1	0.903438527	1.74E-24	NP_001138663.1	FAM200B
NC_001357	E1	0.805278655	1.05E-15	NP_940841.1	KBTBD3
NC_001405	L1	0.865302979	2.94E-20	NP_001073990.2	RASSF10
NC_001430	Pos: 727-7311	0.837550489	6.34E-18	NP_000123.1	F8
NC_001436	Pr55	0.752880597	7.22E-13	NP_001092872.1	CCNK
NC_001454	L3	0.792140958	6.41E-15	NP_612426.1	KTI12
NC_001457	Pos: 5345-6895	0.859158203	1.06E-19	NP_061854.1	DNAJC10
NC_001458	Pos: 822-2678	0.847795937	9.88E-19	NP_001273176.1	RALGPS2
NC_001460	E1B	0.806525776	8.74E-16	NP_001116801.1	ZBTB1
NC_001472	Pos: 742-7290	0.800822126	1.96E-15	NP_005224.2	EPHA3
NC_001488	Pos: 807-2108	0.748225962	1.19E-12	NP_001073882.3	NOBOX
NC_001490	Pos: 629-7168	0.891321462	5.65E-23	NP_002175.2	IL6ST
NC_001526	L1	0.807134439	8E-16	NP_942089.1	MAP4K5
NC_001531	Pos: 961-2781	0.852165343	4.29E-19	NP_079114.3	THNSL1
NC_001576	Pos: 791-2836	0.785723092	1.48E-14	NP_899059.1	RAB27A
NC_001583	Pos: 878-2794	0.787008282	1.26E-14	NP_940841.1	KBTBD3
NC_001586	Pos: 850-2778	0.799660538	2.31E-15	NP_940841.1	KBTBD3
NC_001587	Pos: 5430-7016	0.749586507	1.03E-12	NP_057654.2	ERGIC2
NC_001591	E1	0.845045382	1.65E-18	NP_078787.2	HAUS3
NC_001593	L1	0.744112558	1.84E-12	NP_001167579.1	ZBED6
NC_001595	Pos: 5798-7315	0.770647823	9.56E-14	NP_001273644.1	AGTPBP1
NC_001596	E1	0.844374112	1.86E-18	NP_940841.1	KBTBD3
NC_001612	Pos: 751-7332	0.842341207	2.7E-18	NP_001116105.1	CPS1
NC_001617	Pos: 619-7113	0.86771873	1.74E-20	NP_002175.2	IL6ST
NC_001664	IE1	0.893813269	2.86E-23	NP_653091.3	CASC5
NC_001676	Pos:828-2729	0.787967453	1.11E-14	NP_940841.1	KBTBD3
NC_001690	E1	0.855417316	2.26E-19	NP_001092688.1	RAD51AP2
NC_001691	E1	0.876751214	2.23E-21	NP_940841.1	KBTBD3

NC_001693	E1	0.894934035	2.1E-23	NP_940841.1	KBTBD3
NC_001716	IE1	0.927833476	3.03E-28	NP_001073973.2	RBM44
NC_001722	Pos: 1103-2668	0.737893765	3.5E-12	NP_002408.3	MKI67
NC_001781	L	0.876166171	2.56E-21	NP_065982.1	KIAA1586
NC_001796	Pos: 8646-15347	0.903986563	1.47E-24	NP_065982.1	KIAA1586
NC_001798	UL39	0.904920752	1.1E-24	NP_036567.2	SHC2
NC_001802	Pr55	0.78047161	2.89E-14	NP_001093866.1	C2orf73
NC_001806	UL30	0.90801467	4.15E-25	NP_055778.2	SBNO2
NC_001897	Pos: 703-7242	0.890389641	7.26E-23	NP_001017975.3	HFM1
NC_001943	Pos: 86-4380	0.830734096	2.04E-17	NP_114161.3	SPATA16
NC_002645	Pos: 293-12550	0.774229507	6.22E-14	NP_000099.2	DLD
NC_003266	L4	0.898683268	7.19E-24	NP_009115.2	NISCH
NC_003443	L	0.839684044	4.35E-18	NP_004645.2	USP9Y
NC_003461	L	0.866879002	2.09E-20	NP_065982.1	KIAA1586
NC_004104	E1	0.68207836	5.44E-10	NP_899059.1	RAB27A
NC_004148	L	0.867913209	1.67E-20	NP_065982.1	KIAA1586
NC_004295	VP1	0.773099678	7.13E-14	NP_114414.2	EIF2A
NC_004500	E1	0.880929983	8.18E-22	NP_004645.2	USP9Y
NC_005134	E1	0.851299523	5.07E-19	NP_001138663.1	FAM200B
NC_005147	Pos: 21507-22343	0.820880135	1.01E-16	NP_064506.3	UGGT2
NC_005831	Pos: 287-20475	0.750091303	9.77E-13	NP_037471.2	ALG6
NC_006273	IE1	0.87654333	2.35E-21	NP_055478.2	KDM4A
NC_006577	Pos: 22942-27012	0.756094354	5.07E-13	NP_852607.3	LRRC70
NC_007018	ORF2	0.774104535	6.31E-14	NP_005112.2	MED13
NC_007026	Pos: 828-2486	0.704735964	8.08E-11	NP_001024.1	RRM1
NC_007027	Pos: 94-1698	0.746908872	1.37E-12	NP_002717.3	PREP
NC_007455	VP1	0.768556356	1.22E-13	NP_803875.2	PKHD1L1
NC_007605	BALF5	0.934931283	1.36E-29	NP_620124.1	RHOT2
NC_008188	E1	0.85042555	6E-19	NP_940841.1	KBTBD3
NC_008189	E1	0.781785258	2.45E-14	NP_000305.3	PTEN
NC_009333	ORF75	0.91780911	1.47E-26	NP_002891.1	RBP3
NC_009334	BALF5	0.935906758	8.64E-30	NP_620124.1	RHOT2
NC_009996	Pos: 616-7050	0.834124398	1.15E-17	NP_004939.1	DSC1
NC_010329	E1	0.908048024	4.1E-25	NP_940841.1	KBTBD3
NC_010810	Pos: 956-7837	0.825974666	4.46E-17	NP_004939.1	DSC1
NC_010956	L4	0.884411516	3.44E-22	NP_009115.2	NISCH
NC_011202	L1	0.825443151	4.86E-17	NP_787072.2	EXOC8
NC_011203	L4	0.84556954	1.5E-18	NP_009115.2	NISCH
NC_011800	Pos: 1892-2533	0.744847797	1.71E-12	NP_056526.3	GLTSCR1
NC_012042	VP1	0.776501461	4.72E-14	NP_005424.1	YES1
NC_012213	E1	0.843291298	2.27E-18	NP_001138663.1	FAM200B
NC_012485	E1	0.883809966	4E-22	NP_940841.1	KBTBD3
NC_012486	E1	0.902494945	2.32E-24	NP_001138663.1	FAM200B

NC_012564	VP1	0.783191043	2.05E-14	NP_002899.1	REL
NC_012729	NS2	0.805392124	1.03E-15	NP_001073932.1	DYNC2H1
NC_012798	Pos: 139-6480	0.82589364	4.52E-17	NP_057190.2	SCFD1
NC_012801	Pos: 750-7124	0.824196863	5.94E-17	NP_001191195.1	GABRA4
NC_012802	Pos: 748-7128	0.834958942	9.94E-18	NP_001161829.1	PLA2G7
NC_012950	Pos: 21445-22281	0.818600204	1.44E-16	NP_064506.3	UGGT2
NC_012959	Pos: 22707-24845	0.842896843	2.44E-18	NP_009115.2	NISCH
NC_012986	Pos: 719-7831	0.755617438	5.35E-13	NP_004215.2	GPR50
NC_013035	E1	0.900330229	4.44E-24	NP_940841.1	KBTBD3
NC_014185	E1	0.928268261	2.53E-28	NP_940841.1	KBTBD3
NC_014952	E1	0.879231256	1.24E-21	NP_940841.1	KBTBD3
NC_014953	E1	0.904630266	1.21E-24	NP_940841.1	KBTBD3
NC_014954	E1	0.895597619	1.74E-23	NP_940841.1	KBTBD3
NC_014955	E1	0.905343727	9.67E-25	NP_940841.1	KBTBD3
NC_014956	E1	0.903382032	1.77E-24	NP_940841.1	KBTBD3
NC_015150	Pos: c5026-4790, c4437-2632	0.897789054	9.32E-24	NP_060862.3	C4orf21
NC_015630	Pos: 381-1076	0.54440122	0.00000332	NP_689786.2	RASEF
NC_016157	Pos: 817-2640	0.919910732	6.78E-27	NP_940841.1	KBTBD3
NC_017993	Pos: 805-2610	0.859261423	1.04E-19	NP_940841.1	KBTBD3
NC_017994	E1	0.868341489	1.52E-20	NP_940841.1	KBTBD3
NC_017995	Pos: 714-2546	0.883104334	4.77E-22	NP_001138663.1	FAM200B
NC_017996	Pos: 717-2534	0.881915256	6.42E-22	NP_940841.1	KBTBD3
NC_017997	Pos; 703-2502	0.825068761	5.16E-17	NP_112561.2	TEX15
NC_019023	E1	0.864842857	3.24E-20	NP_940841.1	KBTBD3
NC_019843	orf1ab	0.777846368	4E-14	NP_079265.2	PGAP1
NC_020890	large T antigen	0.894050364	2.68E-23	NP_001017975.3	HFM1
NC_021483	E1	0.858044662	1.34E-19	NP_001092688.1	RAD51AP2
NC_021568	Pos: 279-13433, 13433-21514	0.731131014	6.89E-12	NP_066012.1	METTL14
NC_021928	Pos: c5033-4821, c4421-2508	0.8986358	7.29E-24	NP_065982.1	KIAA1586
NC_022095	L1	0.818922205	1.37E-16	NP_001273644.1	AGTPBP1
NC_022518	Pos: 6451-8550	0.801092218	1.89E-15	NP_001121143.1	LIFR
NC_022892	E1	0.856537918	1.81E-19	NP_065982.1	KIAA1586
NC_023874	Pos: 161-997	0.720321018	1.95E-11	NP_060146.2	GIN1
NC_023891	E1	0.88293482	4.98E-22	NP_940841.1	KBTBD3
NC_023984	Pos: 1362-7727	0.837884709	5.98E-18	NP_036434.1	LPHN2
NC_024694	Pos: 1 - 1113	0.676628221	8.4E-10	NP_054860.1	CNTNAP2

Table S2, Chapter 6. A comprehensive list of the 113 viruses with the number of genes in the virus, the number of highly correlating human genes, and the number of highly correlating human proteins per viral protein. Viruses are ordered in accending order based on the number of highly correlating human genes per viral gene.

Virus Accession Number	Number of Genes in Virus	Number of Highly Correlating Genes in Humans	Number of Highly Correlating Human Proteins per Viral Protein
NC_015630	3	0	0
NC_004104	7	4	0.57
NC_012986	1	1	1
NC_001436	6	7	1.17
NC_024694	4	13	3.25
NC_001488	6	27	4.5
NC_011800	6	28	4.67
NC_007026	2	15	7.5
NC_005831	6	47	7.83
NC_001722	9	91	10.11
NC_001352	7	91	13
NC_023874	2	32	16
NC_001595	6	104	17.33
NC_001357	8	152	19
NC_001454	34	655	19.26
NC_006577	8	165	20.63
NC_021568	2	50	25
NC_001576	7	221	31.57
NC_001587	6	219	36.5
NC_001348	73	2843	38.95
NC_001593	7	331	47.29
NC_000883	6	317	52.83
NC_019843	11	582	52.91
NC_001355	9	478	53.11
NC_001460	36	1950	54.17
NC_001583	6	328	54.67
NC_001676	7	391	55.86
NC_001526	8	456	57
NC_008189	6	353	58.83
NC_001802	10	629	62.9
NC_002645	8	613	76.63
NC_001586	6	517	86.17
NC_015150	5	435	87
NC_007027	1	93	93
NC_011202	38	3637	95.71
NC_007455	4	392	98

NC_001781	11	1079	98.09
NC_017997	7	691	98.71
NC_001354	11	1096	99.64
NC_012950	12	1268	105.67
NC_005147	9	970	107.78
NC_012042	4	438	109.5
NC_004500	7	787	112.43
NC_013035	7	837	119.57
NC_008188	6	720	120
NC_004295	6	747	124.5
NC_022095	6	750	125
NC_012564	4	555	138.75
NC_004148	9	1314	146
NC_001405	38	5628	148.11
NC_000898	104	15694	150.9
NC_012485	7	1083	154.71
NC_006273	169	26217	155.13
NC_001664	88	13960	158.64
NC_012213	5	801	160.2
NC_003461	10	1706	170.6
NC_003266	38	7275	191.45
NC_001798	77	14790	192.08
NC_022892	6	1160	193.33
NC_010956	38	7500	197.37
NC_017993	7	1382	197.43
NC_001690	7	1464	209.14
NC_021483	7	1467	209.57
NC_001596	7	1470	210
NC_014953	7	1498	214
NC_012959	36	7762	215.61
NC_001591	6	1327	221.17
NC_014952	7	1601	228.71
NC_011203	39	9069	232.54
NC_001531	8	1903	237.88
NC_012729	5	1212	242.4
NC_003443	7	1720	245.71
NC_020890	5	1235	247
NC_010329	7	1744	249.14
NC_012486	7	1768	252.57
NC_001691	7	1771	253
NC_023891	7	1843	263.29
NC_001356	7	1844	263.43
NC_021928	7	1879	268.43

NC_005134	7	1893	270.43
NC_014956	7	1894	270.57
NC_001796	8	2167	270.88
NC_016157	7	1969	281.29
NC_001457	7	1980	282.86
NC_014954	7	1981	283
NC_014955	7	2051	293
NC_017994	7	2061	294.43
NC_014185	7	2076	296.57
NC_009333	86	26437	307.41
NC_001458	7	2182	311.71
NC_001693	7	2316	330.86
NC_001806	77	26054	338.36
NC_019023	6	2070	345
NC_017996	7	2500	357.14
NC_007018	2	769	384.5
NC_001716	86	33651	391.29
NC_017995	7	2784	397.71
NC_001943	2	1088	544
NC_022518	1	592	592
NC_001472	1	753	753
NC_007605	95	85227	897.13
NC_009334	80	82905	1036.31
NC_001612	1	1133	1133
NC_009996	1	1157	1157
NC_001617	1	1193	1193
NC_010810	1	1223	1223
NC_012802	1	1408	1408
NC_001490	1	1423	1423
NC_012798	1	1437	1437
NC_023984	1	1453	1453
NC_012801	1	1482	1482
NC_001430	1	1720	1720
NC_001897	1	1918	1918
Average	15.74	4161.41	303.36
Total	1779	470239	34279.52

Table S3, Chapter 6. A comprehensive list of where the highest correlating human protein with respect to a human-infecting virus is most highly expressed.

Virus Accession Number	Highest Correlating Human Protein Accession Number	Region(s) Where Human Protein is Most Highly Expressed
NC_000883	NP_002763.2	Stomach glandular cells
NC_000898	NP_112561.2	Testis, urinary tract, and brain
NC_001348	NP_787081.2	Myocytes in heart muscle, lateral ventricle, cerebral cortex, hippocampus
NC_001352	NP_037485.2	Myocytes in skeletal muscle, and glandular cells in the stomach.
NC_001354	NP_001273387.1	Liver, pancreas, digestive tract, male reproductive system, endocrine
NC_001355	NP_940841.1	Skeletal muscle, smooth muscle, epidermal cells, hepatocytes in liver
NC_001356	NP_001138663.1	GI-tract, gallbladder, and the blood and immune system
NC_001357	NP_940841.1	Smooth muscle cells
NC_001405	NP_001073990.2	Stomach, kidney, fallopian tube,
NC_001430	NP_000123.1	Adipocytes of soft tissue, placenta, tubule cells in the kidney
NC_001436	NP_001092872.1	Hematopoietic cells in bone marrow, glandular cells in the stomach
NC_001454	NP_612426.1	Glandular cells of the GI tract, urinary tract cells, adrenal glands
NC_001457	NP_061854.1	Glandular cells of the epididymis and the endometrium
NC_001458	NP_001273176.1	Testis.
NC_001460	NP_001116801.1	Kidney, testis, stomach, esophagus, vagina, skin, lung, and heart
NC_001472	NP_005224.2	Low expression everywhere
NC_001488	NP_001073882.3	No information found
NC_001490	NP_002175.2	Stomach cells, prostate, kidney, liver, pancreas, heart muscle
NC_001526	NP_942089.1	Female reproductive system
NC_001531	NP_079114.3	Stomach
NC_001576	NP_899059.1	Stomach and rectum
NC_001583	NP_940841.1	Smooth muscle cells
NC_001586	NP_940841.1	Smooth muscle cells
NC_001587	NP_057654.2	Heart muscle cells, and some GI-tract cells.
NC_001591	NP_078787.2	Stomach
NC_001593	NP_001167579.1	GI-tract and female reproductive system
NC_001595	NP_001273644.1	Testis
NC_001596	NP_940841.1	Smooth muscle cells
NC_001612	NP_001116105.1	Stomach and liver
NC_001617	NP_002175.2	Stomach cells, prostate, kidney, liver, pancreas, heart muscle
NC_001664	NP_653091.3	Testis
NC_001676	NP_940841.1	Smooth muscle cells

NC_001690	NP_001092688.1	Male reproductive system
NC_001691	NP_940841.1	Smooth muscle cells
NC_001693	NP_940841.1	Smooth muscle cells
NC_001716	NP_001073973.2	Testis
NC_001722	NP_002408.3	Blood, immune system
NC_001781	NP_065982.1	Seminal vesicle in men, and the breast in women
NC_001796	NP_065982.1	Seminal vesicle in men, and the breast in women
NC_001798	NP_036567.2	Varied expression everywhere
NC_001802	NP_001093866.1	Male reproductive system and GI-tract
NC_001806	NP_055778.2	Liver cells, skeletal muscle, cerebral cortex, endocrine glands, lung
NC_001897	NP_001017975.3	Lung cells and skeletal muscles
NC_001943	NP_114161.3	Testis and cerebellum
NC_002645 NC_003266	NP_000099.2 NP_009115.2	Nearly everywhere, except skin, gallbladder, cerebellum, heart muscle, adrenal gland, bronchus
NC_003443	NP_004645.2	Prostate
NC_003461	NP_065982.1	Seminal vesicle in men, and the breast in women
NC_004104	NP_899059.1	Stomach and rectum
NC_004148	NP_065982.1	Seminal vesicle in men, and the breast in women
NC_004295	NP_114414.2	Skin
NC_004500	NP_004645.2	Prostate
NC_005134	NP_001138663.1	GI-tract, gallbladder, and the blood and immune system
NC_005147	NP_064506.3	Testis and the brain
NC_005831	NP_037471.2	Both male and female reproductive systems
NC_006273	NP_055478.2	Stomach, testis, and brain
NC_006577	NP_852607.3	Hippocampus, heart muscle, parathyroid gland
NC_007018	NP_005112.2	Bone marrow, and testis
NC_007026	NP_001024.1	Testis, lymph nodes, and lateral ventricles
NC_007027	NP_002717.3	GI-tract, and endometrium in women
NC_007455	NP_803875.2	Spleen and bone marrow
NC_007605	NP_620124.1	Stomach, placenta, skeletal muscle, and cerebral cortex
NC_008188	NP_940841.1	Smooth muscle cells
NC_008189	NP_000305.3	Cerebral cortex
NC_009333	NP_002891.1	No information found
NC_009334	NP_620124.1	Stomach, placenta, skeletal muscle, and cerebral cortex
NC_009996	NP_004939.1	highest expression in the skin keratinocytes
NC_010329	NP_940841.1	Smooth muscle cells
NC_010810	NP_004939.1	Skin keratinocytes
NC_010956	NP_009115.2	Skin, gallbladder, cerebellum, heart muscle, adrenal gland, bronchus
NC_011202	NP_787072.2	Adrenal gland, cerebellum, stomach, and placenta
NC_011203	NP_009115.2	Skin, gallbladder, cerebellum, heart muscle, adrenal gland, bronchus

NC_011800	NP_056526.3	Medium/high expression everywhere
NC_012042	NP_005424.1	Testis, stomach, and placenta
NC_012213	NP_001138663.1	GI-tract, gallbladder, blood and immune system
NC_012485	NP_940841.1	Smooth muscle cells
NC_012486	NP_001138663.1	GI-tract, gallbladder, blood and immune system
NC_012564	NP_002899.1	Blood, immune system, women reproductive system, and GI-tract
NC_012729	NP_001073932.1	GI-tract
NC_012798	NP_057190.2	Pancreas, testis, kidney, and placenta
NC_012801	NP_001191195.1	Cerebral cortex
NC_012802	NP_001161829.1	Appendix, prostate, placenta, lymph node, and spleen
NC_012950	NP_064506.3	Testis and the brain
NC_012959	NP_009115.2	Skin, gallbladder, cerebellum, heart muscle, adrenal gland, bronchus
NC_012986	NP_004215.2	Kidney and smooth muscle tissue
NC_013035	NP_940841.1	Smooth muscle cells
NC_014185	NP_940841.1	Smooth muscle cells
NC_014952	NP_940841.1	Smooth muscle cells
NC_014953	NP_940841.1	Smooth muscle cells
NC_014954	NP_940841.1	Smooth muscle cells
NC_014955	NP_940841.1	Smooth muscle cells
NC_014956	NP_940841.1	Smooth muscle cells
NC_015150	NP_060862.3	No information available
NC_015630	NP_689786.2	GI-tract and urinary tract
NC_016157	NP_940841.1	Smooth muscle cells
NC_017993	NP_940841.1	Smooth muscle cells
NC_017994	NP_940841.1	Smooth muscle cells
NC_017995	NP_001138663.1	GI-tract, gallbladder, blood and immune system
NC_017996	NP_940841.1	Smooth muscle cells
NC_017997	NP_112561.2	Low expression everywhere
NC_019023	NP_940841.1	Smooth muscle cells
NC_019843	NP_079265.2	Testis, placenta and parathyroid gland
NC_020890	NP_001017975.3	Lung cells and skeletal muscles
NC_021483	NP_001092688.1	Stomach, male reproductive system, and skin
NC_021568 NC_021928	NP_066012.1 NP_065982.1	Testis and stomach, Seminal vesicle in men, and the breast in women
NC_022095	NP_001273644.1	Testis
NC_022518	NP_001121143.1	Male reproductive tissue and in the heart
NC_022892	NP_065982.1	Seminal vesicle in men, and the breast in women
NC_023874	NP_060146.2	Tonsil, stomach, and pancreas
NC_023891	NP_940841.1	Smooth muscle cells
NC_023984	NP_036434.1	Skeletal and smooth muscle, tonsils, small intestine, colon
NC_024694	NP_054860.1	Cerebral cortex

Appendix 6: Supplementary Figures and Tables for Chapter 7

Algorithms

Supplementary Algorithm 1, Chapter 7.

Algorithm 1: JustOrthologs

```
function compareSequenceSets(f1, f2):
Input: 2 sorted sets of sequences
Output: set of orthologous pairs
orthologous_pairs = {}
for s1 in f1:
    best_total = 2 * (getCdsCount(s1) - 1)
    overall_best_seq = ''
    for s2 in f2:
        total = 0
        bssf = 0.05
        best_seq = ''
        if |getCdsCount(s1) - getCdsCount(s2)| <= 3:
             for c1 in s1:
                 for c2 in s2:
                     if |c1| == |c2| && |c2| > 15:
                          d = sumDifferencesOfAllDinucleotideCounts(c1, c2)
                          if d < bssf:
                              bssf = d
                              best_seq = s2
                 if bssf < 0.05: total += bssf
                 else: total += 2
             if type(total) == "float" && total <= best_total:
                 best_total = total
                 overall_best_seq = best_seq
    if |overall_best_seq| > 0:
        orthologous_pairs U= {(s1, overall_best_seq)}
return orthologous_pairs

Input: 2 sets of sequences as f1, f2
Output: orthologs
orthologs = compareSequenceSets(sort(f1), sort(f2)) U= compareSequenceSets(sort(f2), sort(f1))
for o in orthologs:
    if !( oi-1 -> oi-2 && oi-2 -> oi-j):
      print o
```

Supplementary Algorithm 2, Chapter 7.

Algorithm 2: JustOrthologs -d

```
function compareSequenceSets(f1, f2):
Input: 2 sorted sets of sequences
Output: set of orthologous pairs
orthologous_pairs = {}
for s1 in f1:
    overall_best_seq = ''
    for s2 in f2:
        total = 0
        bssf = 0.1
        best_seq = ''
        if |getCdsCount(s1) - getCdsCount(s2)| <= 3:
            for c1 in s1:
                for c2 in s2:
                    if |c1| == |c2| && |c2| > 15:
                        d = sumDifferencesOfAllDinucleotideCounts(c1, c2) - 2.2
                        h = highestDinucleotideDifference(c1, c2)
                        if h < 0.03 && d < bssf:
                            bssf = d
                            best_seq = s2
                if bssf < 0.1: total += bssf
                else: total += 2
            if type(total) == "float" && ( (getCdsCount(s1) < 6 && total < -2.15) || (getCdsCount(s1) >= 6 && total < 0) ):
                best_total = total
                overall_best_seq = best_seq
    if |overall_best_seq| > 0:
        orthologous_pairs ∪= {(s1, overall_best_seq)}
return orthologous_pairs

Input: 2 sets of sequences as f1, f2
Output: orthologs
orthologs = compareSequenceSets(sort(f1), sort(f2)) ∪= compareSequenceSets(sort(f2), sort(f1))
for o in orthologs:
    if !( oi-1 -> oi-2 && oi-2 -> oi-j):
      print o
```

Supplementary Algorithm 3, Chapter 7.

Algorithm 3: JustOrthologs -c

```
Input: 2 sets of sequences as f1, f2
Output: orthologs
orthologs = JustOrthologs(f1, f2) ∪= JustOrthologs -d (f1, f2)
for o in orthologs:
   if !( oi-1 -> oi-2 && oi-2 -> oi-j):
      print o
```

Supplementary Notes

Supplementary Note 1, Chapter 7. Tuning the threshold parameter is a simple, yet time-consuming process. It requires orthologs to be annotated in the species used to tune the parameter. After choosing which species to tune the threshold parameter, run JustOrthologs for values between 0.01 and 1.00, incremented by 0.01 (i.e., 100 times), saving the output in different files. Then count the number of correctly classified orthologs, false positive orthologs, and calculate the precision and accuracy of each run. Based on precision and accuracy scores, choose the threshold value that best suites the needs of the research. The commands are presented for executing this process, where [s1] represents the first species, [s2] represents the second species, [o] represents the output file, and [t] represents the threshold.

JustOrthologs for closely related species:
```
python justOrthologs.py -s [s1] -q [s2] -o [o] -r [t]
```
JustOrthologs for distantly related species:
```
python justOrthologs.py -d -s [s1] -q [s2] -o [o] -r [t]
```
JustOrthologs combined approach:
```
python justOrthologs.py -c -s [s1] -q [s2] -o [o] -r [t]
```

Supplementary Note 2, Chapter 7. Since JustOrthologs relies heavily on identifying CDS regions that are the same length, it is possible for two gene to have the same CDS region lengths and similar dinucleotide percentages without having similar sequences. For this reason, precision and accuracy are lowest when genes have fewer CDS regions. For instance, if a simulated gene has one CDS region with 21 nucleotides, it could randomly have very similar same dinucleotide percentage as a different sequence if portions of the gene are rearranged:

Sequence 1: `ATGAAATTTCCCGGGATCTAA`
Sequence 2: `ATGCCCATCTTTAAAGGGTAA`

In this case, the start codon and the stop codon are identical, but the sequence alignment would be very poor because of the nucleotide rearrangements. However, the sequences have the same nucleotide composition, and the same codons in a different order. So, two-thirds of the dinucleotide percentages will be identical even without chance collisions with subsequent codons (e.g ATCCAG and ATCCCC share a CC dinucleotide between codons).

Supplementary Tables

Supplementary Table 1, Chapter 7. Comparing the strengths and weaknesses of the algorithms used. All three algorithms also use a time-intensive all-versus-all BLAST to recover orthologous groups.

Algorithm	Advantages	Disadvantages
OrthoMCL	Widely used	Complicated 13-step process
	High recall	
OrthoFinder	Single-step process	Slow
	High precision	Several software dependencies
OMA	Comprehensive ortholog database	Strict directory structure
		Not easily scriptable

Supplementary Table 2, Chapter 7. Species included in the combined analysis of 1 197 species.

Eukaryota	Archaea
Ooceraea biroi	*Thermococcus sp. 5-4*
Pseudocercospora fijiensis	*Methanopyrus sp. KOL6*
Carlito syrichta	*Halorientalis sp. IM1011*
Exaiptasia pallida	*Methanonatronarchaeum thermophilum*
Baudoinia panamericana	*Candidatus Micrarchaeota archaeon MIA14*
Wallemia mellicola	*Natronomonas sp. CBA1134*
Blastomyces gilchristii	*Natrialbaceae archaeon JW/NM-HA 15*
Phialophora attae	*Candidatus Nitrosomarinus catalina*
Metarhizium majus	*Haladaptatus sp. W1*
Sinocyclocheilus anshuiensis	*Sulfolobus sp. A20*
Mitosporidium daphniae	*Methanolobus sp. YSF-3*
Paraphaeosphaeria sporulosa	*Halodesulfurarchaeum formicicum*
Talaromyces atroroseus	*Natrialba sp. SSL1*
Kalmanozyma brasiliensis	*Methanosphaera sp. A6*
Fonsecaea erecta	*Methanobrevibacter sp. A54*
Auricularia subglabra	*Methanobacterium sp. A39*
Xylona heveae	*Methanobrevibacter sp. A27*
Cladophialophora psammophila	*Methanosarcina sp. A14*
Penicillium rubens	*Halorubrum sp. SD683*
Saccharomyces eubayanus	*Halorubrum sp. SD612*
Tetrapisispora blattae	*Halolamina sp. CBA1230*
Verticillium alfalfae	*Nitrosopumilus sp. Nsub*
Paracoccidioides lutzii	*Candidatus Methanomethylophilus sp. 1R26*
Zymoseptoria tritici	*Halorubrum tropicale*
Aureobasidium subglaciale	*Halobacterium sp. CBA1132*
Wickerhamomyces ciferrii	*Haloarcula sp. CBA1128*
Drosophila rhopaloa	*Haloarcula sp. CBA1127*
Exophiala aquamarina	*Halorubrum aethiopicum*
Nannospalax galili	*Methanosarcina flavescens*
Ogataea parapolymorpha	*Thermococcus piezophilus*
Heterobasidion irregulare	*Haloarcula rubripromontorii*
Morus notabilis	*Haladaptatus sp. R4*
Fonsecaea multimorphosa	*Haloparvum sedimenti*
Salpingoeca rosetta	*Thermococcus sp. 2319x1*
Blastocystis sp. subtype 4	*Cuniculiplasma divulgatum*
Balearica regulorum	*Vulcanisaeta sp. EB80*
Dendrobium catenatum	*Thermoproteus sp. CP80*
Fukomys damarensis	*Caldivirga sp. MU80*
Cutaneotrichosporon oleaginosum	*Haloarcula sp. K1*
Schizosaccharomyces cryophilus	*Methanobrevibacter sp. YE315*
Sugiyamaella lignohabitans	*Pyrococcus kukulkanii*
Aspergillus fumigatus	*Halanaeroarchaeum sulfurireducens*
Polistes dominula	*Candidatus Nitrosotenuis cloacae*
Hyphopichia burtonii	*Haloarcula sp. CBA1115*

Colletotrichum fioriniae	*Thermococcus sp. EP1*
Marssonina brunnea	*Candidatus Nitrosopumilus adriaticus*
Fonticula alba	*Candidatus Methanoplasma termitum*
Amyelois transitella	*Methanosphaera sp. WGK6*
Pseudogymnoascus destructans	*Methanoculleus sediminis*
Plasmodium gaboni	*Thermofilum uzonense*
Harpegnathos saltator	*Haloferax sp. SB3*
Dinoponera quadriceps	*Haloferax sp. SB29*
Fibroporia radiculosa	*Haloferax sp. Q22*
Habropoda laboriosa	*Haloprofundus marisrubri*
Nematocida parisii	*Haloferax sp. ATB1*
Encephalitozoon romaleae	*Thermococcus eurythermalis*
Cladophialophora immunda	*Halorubrum sp. BV1*
Metarhizium robertsii	*Methanobacterium sp. SMA-27*
Scedosporium apiospermum	*Methanoculleus sp. MH98A*
Aureobasidium namibiae	*methanogenic archaeon ISO4-H5*
Chlorella variabilis	*Halostagnicola sp. A56*
Thecamonas trahens	*Methanosarcina sp. 1.H.T.1A.1*
Atta colombica	*Methanosarcina sp. 1.H.A.2.2*
Eufriesea mexicana	*Methanosarcina sp. 2.H.A.1B.4*
Metarhizium brunneum	*Methanosarcina sp. 2.H.T.1A.6*
[Candida] auris	*Methanosarcina sp. 2.H.T.1A.8*
Galeopterus variegatus	*Methanosarcina sp. 2.H.T.1A.3*
Colletotrichum gloeosporioides	*Methanosarcina sp. 2.H.T.1A.15*
Trachymyrmex cornetzi	*Thermofilum carboxyditrophus*
Cladophialophora yegresii	*Halorubrum sp. HD13*
Kwoniella mangrovensis	*Halorubrum sp. EA1*
Ogataea polymorpha	*Halolamina sediminis*
Komagataella phaffii	*Haloferax massiliensis*
Cyphomyrmex costatus	*Halapricum salinum*
Diachasma alloeum	*Halobellus rufus*
Kwoniella pini	*Methanosarcina sp. WWM596*
Neodiprion lecontei	*Methanosarcina sp. WH1*
Kazachstania africana	*Methanosarcina sp. MTP4*
Dactylellina haptotyla	*Methanosarcina sp. Kolksee*
Camelus ferus	*Candidatus Halobonum tyrrellensis*
Entamoeba nuttalli	*Halorubrum sp. SD626R*
Vollenhovia emeryi	*Haloterrigena mahii*
Amphimedon queenslandica	*Candidatus Nitrosopelagicus brevis*
Trichoderma gamsii	*Halobacterium hubeiense*
Pestalotiopsis fici	*Thaumarchaeota archaeon N4*
Lachancea thermotolerans	*Candidatus Methanomassiliicoccus intestinalis*
Propithecus coquereli	*Candidatus Methanoperedens nitroreducens*
Endocarpon pusillum	*Halorubrum halodurans*
Acytostelium subglobosum	*Halorubrum sp. IB24*

Exophiala xenobiotica	*Halorubrum sp. EB13*
Falco cherrug	*Halorubrum sp. EA8*
Pseudogymnoascus verrucosus	*Halolamina rubra*
Spathaspora passalidarum	*Methanobacterium sp. MB1*
Manacus vitellinus	*Halorientalis persicus*
Pseudozyma hubeiensis	*Thermofilum adornatus*
Ziziphus jujuba	*Candidatus Nitrosotenuis chungbukensis*
Kwoniella dejecticola	*Methanococcoides vulcani*
Kwoniella bestiolae	*Methanobrevibacter boviskoreani*
Lepidothrix coronata	*Salinarchaeum sp. Harcht-BSK1*
Magnaporthe oryzae	*Archaeoglobus sulfaticallidus*
Neofusicoccum parvum	*Methanocaldococcus bathoardescens*
Sinocyclocheilus rhinocerous	*Sulfolobus sp. JCM 16833*
Monomorium pharaonis	*Acidiplasma sp. MBA-1*
Pundamilia nyererei	*Halopiger goleimassiliensis*
Micromonas commoda	*Halopiger djelfimassiliensis*
Hyalella azteca	*Candidatus Methanomethylophilus alvus*
Cyphellophora europaea	*Pyrodictium delaneyi*
Miniopterus natalensis	*Halopenitus malekzadehii*
Halyomorpha halys	*Halococcus sediminicola*
Pochonia chlamydosporia	*Haloferax sp. BAB2207*
Antrostomus carolinensis	*Halopelagius longus*
Candida orthopsilosis	*Halococcus agarilyticus*
Pneumocystis murina	*Candidatus Nitrosopumilus sediminis*
Cavenderia fasciculata	*Halorubrum hochstenium*
Rhincodon typus	*Haloferax sp. ATCC BAA-646*
Verruconis gallopava	*Haloferax sp. ATCC BAA-645*
Coccomyxa subellipsoidea	*Haloferax sp. ATCC BAA-644*
Tupaia chinensis	*Halopiger salifodinae*
Wallemia ichthyophaga	*Thermogladius calderae*
Cynoglossus semilaevis	*Halorubrum sp. T3*
Ostreococcus 'lucimarinus'	*Natrinema salaciae*
Pterocles gutturalis	*Pyrococcus sp. ST4*
Cryptosporidium hominis	*Methanocella arvoryzae*
Diplodia corticola	*Methanocella conradii*
Phanerochaete carnosa	*Halorubrum sp. AJ67*
Bigelowiella natans	*Natrinema sp. J7-1*
Grosmannia clavigera	*Candidatus Nitrosopumilus salaria*
Myotis davidii	*Candidatus Acidianus copahuensis*
Lottia gigantea	*Metallosphaera yellowstonensis*
Pyrus x bretschneideri	*Natronobacterium texcoconense*
Agrilus planipennis	*Natronorubrum texcoconense*
Phaeoacremonium minimum	*Candidatus Nitrosoarchaeum koreensis*
Orussus abietinus	*Methanomassiliicoccus luminyensis*
Moniliophthora roreri	*Halohasta litchfieldiae*
Pseudomyrmex gracilis	*Thermoplasmatales archaeon BRNA1*

Anoplophora glabripennis	*Halopenitus persicus*
Larimichthys crocea	*Thermococcus sp. 4557*
Exophiala oligosperma	*Methanosarcina soligelidi*
Exophiala mesophila	*Nitrososphaera viennensis*
Cephus cinctus	*Nitrosopumilus sp. SJ*
Chaetomium thermophilum	*Nitrosopumilus sp. AR*
Fomitiporia mediterranea	*Pyrococcus yayanosii*
Plasmodium coatneyi	*Halalkalicoccus paucihalophilus*
Melampsora larici-populina	*Candidatus Nitrosoarchaeum limnia*
Punctularia strigosozonata	*Metallosphaera cuprina*
Coccidioides posadasii	*Haloarchaeobius iranensis*
Calidris pugnax	*Vulcanisaeta moutnovskia*
Capsaspora owczarzaki	*Palaeococcus pacificus*
Egretta garzetta	*Thermococcus sp. PK*
Leptosomus discolor	*Halovenus aranensis*
Chlamydotis macqueenii	*Methanoregula formicica*
Hipposideros armiger	*Methanobacterium lacus*
Chrysochloris asiatica	*Methanobacterium paludis*
Pseudopodoces humilis	*Vulcanisaeta thermophila*
Corvus cornix	*Methanothermobacter sp. CAT2*
Jatropha curcas	*Halomicrobium zhouii*
Dufourea novaeangliae	*halophilic archaeon DL31*
Python bivittatus	*Halobacterium sp. DL1*
Beauveria bassiana	*Halolamina pelagica*
Nestor notabilis	*Halogranum salarium*

Buceros rhinoceros	*Aciduliprofundum sp. MAR8-339*
Ordospora colligata	*Methanocaldococcus villosus*
Bactrocera latifrons	*Halogranum gelatinilyticum*
Aspergillus nidulans	*Halogranum amylolyticum*
Beta vulgaris	*Halorientalis regularis*
Folsomia Candida	*Halobellus clavatus*
Vigna radiata	*Methermicoccus shengliensis*
Leptomonas pyrrhocoris	*Methanocaldococcus sp. FS406-22*
Aphanomyces invadans	*Natronorubrum sediminis*
Ceratina calcarata	*Halostagnicola kamekurae*
Boleophthalmus pectinirostris	*Haloterrigena daqingensis*
Phialocephala scopiformis	*Thermococcus paralvinellae*
Gekko japonicus	*Geoglobus acetivorans*
Monoraphidium neglectum	*Acidianus hospitalis*
Stegastes partitus	*Halogeometricum limi*
Pogonomyrmex barbatus	*Haloplanus vescus*
Megachile rotundata	*Halogeometricum rufum*
Diuraphis noxia	*Haladaptatus litoreus*
Condylura cristata	*Halopelagius inordinatus*
Ceratosolen solmsi	*Halogranum rubrum*
Nectria haematococca	*Methanoculleus horonobensis*
Zootermopsis nevadensis	*Acidiplasma aeolicum*
Acanthaster planci	*Haloterrigena salina*
Pteropus vampyrus	*Candidatus Korarchaeum cryptofilum*
Bombus impatiens	*Halarchaeum acidiphilum*

Arachis ipaensis	*Methanofollis ethanolicus*
Arachis duranensis	*Methanolobus profundi*
Galdieria sulphuraria	*Halorubrum kocurii*
Nipponia nippon	*Methanosphaerula palustris*
Drosophila biarmipes	*Haladaptatus cibarius*
Nanorana parkeri	*Halomicrobium katesii*
Diaphorina citri	*Halorhabdus tiamatea*
Paracoccidioides brasiliensis	*Methanolobus psychrophilus*
Tauraco erythrolophus	*Natrinema gari*
Pediculus humanus	*Methanosarcina horonobensis*
Lipotes vexillifer	*Halorubrum californiense*
Picoides pubescens	*Natronomonas moolapensis*
Fusarium verticillioides	*Halorubrum halophilum*
Aethina tumida	*Natronococcus jeotgali*
Isaria fumosorosea	*Haloterrigena jeotgali*
Parasteatoda tepidariorum	*Halalkalicoccus jeotgali*
Dichomitus squalens	*Halogeometricum pallidum*
Tetrapisispora phaffii	*Halorubrum litoreum*
Scleropages formosus	*Methanobacterium veterum*
Saprolegnia diclina	*Natrinema sp. J7-2*
Aphanomyces astaci	*Haloferax elongans*
Gregarina niphandrodes	*Haloferax mucosum*
Nicrophorus vespilloides	*Haloarcula amylolytica*
Batrachochytrium dendrobatidis	*Haloterrigena hispanica*
Spizellomyces punctatus	*Natronorubrum sulfidifaciens*

Myotis brandtii	*Methanobrevibacter sp. 87.7*
Hippocampus comes	*Halopiger xanaduensis*
Nilaparvata lugens	*Methanobacterium arcticum*
Herrania umbratica	*Haloferax prahovense*
Maylandia zebra	*Aciduliprofundum boonei*
Nothobranchius furzeri	*Haloplanus natans*
Bactrocera oleae	*Haloterrigena limicola*
Cryptococcus amylolentus	*Halorubrum lipolyticum*
Camponotus floridanus	*Halorubrum aidingense*
Gloeophyllum trabeum	*Halorubrum arcis*
Postia placenta	*Haladaptatus paucihalophilus*
Protobothrops mucrosquamatus	*Methanoregula boonei*
Pogona vitticeps	*Halobacterium jilantaiense*
Vavraia culicis	*Halostagnicola larsenii*
Acromyrmex echinatior	*Pyrococcus sp. NA2*
Prunus mume	*Thermococcus onnurineus*
Glarea lozoyensis	*Halorubrum ezzemoulense*
Saprolegnia parasitica	*Halococcus thailandensis*
Bipolaris oryzae	*Halorubrum chaoviator*
Fusarium pseudograminearum	*Halovivax asiaticus*
Hammondia hammondi	*Halococcus hamelinensis*
Phaethon lepturus	*Acidiplasma cupricumulans*
Eutypa lata	*Thermococcus kodakarensis*
Gavialis gangeticus	*Natronorubrum thiooxidans*

Tinamus guttatus	*Haloferax larsenii*
Coturnix japonica	*Haloterrigena saccharevitans*
Setosphaeria turcica	*Methanosaeta harundinacea*
Metarhizium acridum	*Methanobrevibacter olleyae*
Exophiala spinifera	*Haloquadratum walsbyi*
Polistes canadensis	*Thermococcus thioreducens*
Camelina sativa	*Halorubrum terrestre*
Cladophialophora bantiana	*Methanolinea tarda*
Bubalus bubalis	*Haloferax sulfurifontis*
Rhinolophus sinicus	*Natronolimnobius baerhuensis*
Selaginella moellendorffii	*Natronolimnobius innermongolicus*
Rhinocladiella mackenziei	*Thermococcus sp. AM4*
Cladophialophora carrionii	*Haloarcula californiae*
Serpula lacrymans	*Acidilobus saccharovorans*
Sphaerulina musiva	*Methanobrevibacter millerae*
Citrus clementina	*Methanolacinia paynteri*
Corvus brachyrhynchos	*Aeropyrum camini*
Moesziomyces antarcticus	*Halobiforma lacisalsi*
Anthracocystis flocculosa	*Thermococcus celericrescens*
Linepithema humile	*Methanobrevibacter sp. ABM4*
Trichosporon asahii	*Natrinema altunense*
Capsella rubella	*Methanotorris formicicus*
Monosiga brevicollis	*Caldisphaera lagunensis*
Acanthochromis polyacanthus	*Thermococcus nautili*
Colletotrichum higginsianum	*Methanobrevibacter wolinii*

Coniophora puteana	*Methanobrevibacter gottschalkii*
Cimex lectularius	*Thermococcus radiotolerans*
Microtus ochrogaster	*Thermococcus gammatolerans*
Phalaenopsis equestris	*Thermoproteus uzoniensis*
Thermothelomyces thermophila	*Thermococcus sibiricus*
Dendroctonus ponderosae	*Halosimplex carlsbadense*
Malassezia pachydermatis	*Methanosarcina lacustris*
Malassezia sympodialis	*Vulcanisaeta distributa*
Malassezia globosa	*Thermococcus cleftensis*
Papilio polytes	*Natrialba taiwanensis*
Papilio machaon	*Ignicoccus hospitalis*
Populus euphratica	*Methanothermococcus okinawensis*
Sinocyclocheilus grahami	*Halobiforma haloterrestris*
Cordyceps militaris	*Halorhabdus utahensis*
Eutrema salsugineum	*Methanothermobacter marburgensis*
Nannochloropsis gaditana	*Methanothermobacter thermautotrophicus*
Sphaeroforma arctica	*Methanothermobacter wolfeii*
Bos mutus	*Halobiforma nitratireducens*
Trypanosoma grayi	*Natrialba aegyptia*
Eucalyptus grandis	*Natrialba hulunbeirensis*
Acropora digitifera	*Thermococcus sp. P6*
Ostreococcus tauri	*Haloterrigena thermotolerans*
Microplitis demolitor	*Halorubrum tebenquichense*
Theileria orientalis	*Methanoculleus chikugoensis*

Rasamsonia emersonii	*Methanobacterium congolense*
Phytophthora sojae	*Haloferax alexandrinus*
Papilio xuthus	*Geoglobus ahangari*
Fopius arisanus	*Sulfolobus tokodaii*
Wasmannia auropunctata	*Thermococcus guaymasensis*
Trachymyrmex zeteki	*Methanomethylovorans hollandica*
Lobosporangium transversale	*Ferroplasma acidarmanus*
Pieris rapae	*Desulfurococcus amylolyticus*
Trichoderma atroviride	*Natrinema versiforme*
Trichophyton verrucosum	*Methanoculleus sp. MAB1*
Microsporum canis	*Staphylothermus hellenicus*
Nannizzia gypsea	*Archaeoglobus profundus*
Trichophyton benhamiae	*Methanoculleus bourgensis*
Nomascus leucogenys	*Methanocorpusculum labreanum*
Rhinopithecus roxellana	*Palaeococcus ferrophilus*
Rhinopithecus bieti	*Methanobrevibacter ruminantium*
Coniosporium apollinis	*Methanocaldococcus fervens*
Chlorocebus sabaeus	*Picrophilus torridus*
Ficedula albicollis	*Caldivirga maquilingensis*
Taeniopygia guttata	*Ferroplasma acidiphilum*
Arabidopsis lyrata	*Methanocaldococcus vulcanius*
Pantholops hodgsonii	*Thermococcus siculi*
Myotis lucifugus	*Thermococcus pacificus*
Encephalitozoon intestinalis	*Thermococcus gorgonarius*
Debaryomyces fabryi	*Methanocorpusculum bavaricum*

Fragaria vesca	*Natrinema pallidum*
Merops nubicus	*Natrinema pellirubrum*
Colius striatus	*Natrialba chahannaoensis*
Apaloderma vittatum	*Methanocaldococcus infernus*
Acanthisitta chloris	*Methanobrevibacter oralis*
Labrus bergylta	*Halorubrum coriense*
Tyto alba	*Natrialba asiatica*
Cuculus canorus	*Halorubrum vacuolatum*
Guillardia theta	*Natronorubrum tibetense*
Eurypyga helias	*Haloterrigena turkmenica*
Cariama cristata	*Halococcus saccharolyticus*
Mesitornis unicolor	*Natronorubrum bangense*
Ascoidea rubescens	*Halogeometricum borinquense*
Colobus angolensis	*Archaeoglobus veneficus*
Pyrenophora teres	*Halomicrobium mukohataei*
Austrofundulus limnaeus	*Thermococcus barophilus*
Haliaeetus leucocephalus	*Methanobrevibacter filiformis*
Plasmodium inui	*Thermococcus chitonophagus*
Elaeis guineensis	*Ferroglobus placidus*
Plutella xylostella	*Ignicoccus islandicus*
Trichoderma reesei	*Thermosphaera aggregans*
Castor canadensis	*Pyrolobus fumarii*
Jaculus Jaculus	*Hyperthermus butylicus*
Eimeria necatrix	*Methanolacinia petrolearia*
Juglans regia	*Thermococcus barossii*
Necator americanus	*Thermococcus zilligii*
Charadrius vociferus	*Pyrococcus horikoshii*
Nicotiana attenuata	*Thermococcus peptonophilus*

Geospiza fortis	*Haloarcula hispanica*
Poecilia mexicana	*Thermoplasma volcanium*
Poecilia latipinna	*Thermococcus profundus*
Poecilia formosa	*Methanobrevibacter curvatus*
Orbicella faveolata	*Methanobrevibacter cuticularis*
Entamoeba dispar	*Sulfolobus metallicus*
[Candida] tanzawaensis	*Picrophilus oshimae*
Babjeviella inositovora	*Natronobacterium gregoryi*
[Candida] tenuis	*Natronococcus amylolyticus*
Nematostella vectensis	*Halobaculum gomorrense*
Eremothecium sinecaudum	*Haloarcula argentinensis*
Eremothecium cymbalariae	*Metallosphaera sedula*
Cyanidioschyzon merolae	*Sulfolobus islandicus*
Pyrenophora tritici-repentis	*Methanococcus aeolicus*
Bipolaris sorokiniana	*Methanosalsum zhilinae*
Dictyostelium discoideum	*Methanohalophilus portucalensis*
Eimeria mitis	*Methanobrevibacter arboriphilus*
Zonotrichia albicollis	*Methanococcus maripaludis*
Aureococcus anophagefferens	*Methanosarcina siciliae*
Monopterus albus	*Methanolobus vulcani*
Capronia coronata	*Halococcus salifodinae*
Capronia epimyces	*Haloferax gibbonsii*
Ictidomys tridecemlineatus	*Haloferax denitrificans*
Pygocentrus nattereri	*Halorubrum sodomense*
Vittaforma corneae	*Haloarcula sinaiiensis*

Candida dubliniensis	*Pyrococcus abyssi*
Phoenix dactylifera	*Methanococcoides burtonii*
Sorex araneus	*Natronococcus occultus*
Prunus avium	*Halorubrum distributum*
Pneumocystis jirovecii	*Haloarcula japonica*
Drosophila bipectinata	*Haloarcula vallismortis*
Bathycoccus prasinos	*Natrialba magadii*
Aspergillus nomius	*Methanohalobium evestigatum*
Botrytis cinerea	*Methanosphaera stadtmanae*
Stereum hirsutum	*Methanoplanus limicola*
Fonsecaea pedrosoi	*Pyrodictium occultum*
Nosema ceranae	*Thermoplasma acidophilum*
Lonchura striata	*Sulfolobus solfataricus*
Neurospora tetrasperma	*Sulfolobus acidocaldarius*
Bipolaris victoriae	*Staphylothermus marinus*
Tuber melanosporum	*Pyrobaculum islandicum*
Micromonas pusilla	*Desulfurococcus mucosus*
Alligator sinensis	*Thermoproteus tenax*
Phascolarctos cinereus	*Thermofilum pendens*
Chaetomium globosum	*Thermococcus litoralis*
Cryptococcus gattii VGI	*Thermococcus celer*
Talaromyces marneffei	*Pyrococcus furiosus*
Aegilops tauschii	*Natronomonas pharaonis*
Octopus bimaculoides	*Haloferax lucentense*
Priapulus caudatus	*Haloferax mediterranei*
Athalia rosae	*Halococcus morrhuae*

Aotus nancymaae	*Halorubrum saccharovorum*
Gavia stellata	*Halorubrum lacusprofundi*
Kryptolebias marmoratus	*Haloferax volcanii*
Lodderomyces elongisporus	*Halobacterium salinarum*
Clavispora lusitaniae	*Haloarcula marismortui*
Gaeumannomyces tritici	*Archaeoglobus fulgidus*
Penicillium digitatum	*Methanococcoides methylutens*
Aspergillus fischeri	*Methanothrix soehngenii*
Pelecanus crispus	*Methanolobus tindarius*
Vanderwaltozyma polyspora	*Methanosarcina vacuolata*
Ipomoea nil	*Methanosarcina acetivorans*
Thielavia terrestris	*Methanosarcina thermophila*
Stomoxys calcitrans	*Methanosarcina mazei*
Thalassiosira pseudonana	*Methanosarcina barkeri*
Thamnophis sirtalis	*Methanomicrobium mobile*
Chinchilla lanigera	*Methanospirillum hungatei*
Trachymyrmex septentrionalis	*Methanofollis liminatans*
Galendromus occidentalis	*Methanoculleus thermophilus*
Mixia osmundae	*Methanoculleus marisnigri*
Purpureocillium lilacinum	*Methanogenium cariaci*
Uncinocarpus reesii	*Methanocaldococcus jannaschii*
Aspergillus terreus	*Methanotorris igneus*
Eremothecium gossypii	*Methanococcus voltae*
Entamoeba invadens	*Methanococcus vannielii*

Acinonyx jubatus	*Methanothermococcus thermolithotrophicus*
Neolamprologus brichardi	*Methanothermus fervidus*
Tetranychus urticae	*Methanohalophilus halophilus*
Colletotrichum graminicola	*Methanohalophilus mahii*
Enterocytozoon bieneusi	*Methanobrevibacter smithii*
Perkinsus marinus	*Methanobacterium formicicum*
Caenorhabditis remanei	
Takifugu rubripes	
Otolemur garnettii	
Microcebus murinus	
Vicugna pacos	
Fulmarus glacialis	
Opisthocomus hoazin	
Bombus terrestris	
Drosophila kikkawai	
Drosophila ficusphila	
Drosophila elegans	
Drosophila busckii	
Sporothrix schenckii	
Rhodotorula graminis	
Laccaria bicolor	
Trichoderma virens	
Vitis vinifera	
Gossypium raimondii	
Gossypium arboreum	
Neospora caninum	
Crassostrea gigas	

Neomonachus schauinslandi
Eptesicus fuscus
Ursus maritimus
Helicoverpa armigera
Copidosoma floridanum
Drosophila takahashii
Drosophila eugracilis
Kluyveromyces lactis
Cyprinodon variegatus
Elephantulus edwardii
Rhagoletis zephyria
Zeugodacus cucurbitae
Drosophila suzukii
Talaromyces stipitatus
Tarenaya hassleriana
Solanum pennellii
Anolis carolinensis
Encephalitozoon hellem
Saimiri boliviensis
Bactrocera dorsalis
Verticillium dahliae
Penicillium expansum
Metschnikowia bicuspidata
Naumovozyma dairenensis
Naumovozyma castellii
Brachypodium distachyon

Pelodiscus sinensis
Solenopsis invicta
Parastagonospora nodorum
Heterostelium pallidum
Monodelphis domestica
Arthrobotrys oligospora
Amborella trichopoda
Myzus persicae
Melopsittacus undulatus
Blastocystis hominis
Atta cephalotes
Trichoplax adhaerens
Saccoglossus kowalevskii
Heterocephalus glaber
Octodon degus
Cavia porcellus
Rattus norvegicus
Mus pahari
Mus musculus
Mus caroli
Meriones unguiculatus
Peromyscus maniculatus
Mesocricetus auratus
Cricetulus griseus
Dipodomys ordii
Marmota Marmota
Oryctolagus cuniculus
Ochotona princeps
Manis javanica

Ovis aries
Capra hircus
Bos indicus
Bos taurus
Bison Bison
Odocoileus virginianus
Camelus dromedarius
Camelus bactrianus
Sus scrofa
Orycteropus afer
Ceratotherium simum
Equus przewalskii
Equus caballus
Equus asinus
Loxodonta africana
Trichechus manatus
Balaenoptera acutorostrata
Physeter catodon
Tursiops truncatus
Orcinus orca
Leptonychotes weddellii
Odobenus rosmarus
Panthera tigris
Panthera pardus
Felis catus
Mustela putorius
Ailuropoda melanoleuca
canis lupus
Homo sapiens
Pongo abelii
Pan troglodytes
Pan paniscus
Gorilla Gorilla
Mandrillus leucophaeus

Papio anubis
Macaca nemestrina
Macaca mulatta
Macaca fascicularis
Cercocebus atys
Cebus capucinus
Callithrix jacchus
Rousettus aegyptiacus
Pteropus alecto
Echinops telfairi
Erinaceus europaeus
Dasypus novemcinctus
Sarcophilus harrisii
Ornithorhynchus anatinus
Calypte anna
Pygoscelis adeliae
Aptenodytes forsteri
Phalacrocorax carbo
Sturnus vulgaris
Parus major
Serinus canaria
Meleagris gallopavo
Gallus Gallus
Numida Meleagris
Haliaeetus albicilla
Aquila chrysaetos
Falco peregrinus
Columba livia
Chaetura pelagica
Anser cygnoides
Anas platyrhynchos
Apteryx australis
Struthio Camelus

Crocodylus porosus
Alligator mississippiensis
Chrysemys picta
Chelonia mydas
Xenopus tropicalis
Xenopus laevis
Paralichthys olivaceus
Notothenia coriiceps
Lates calcarifer
Haplochromis burtoni
Oreochromis niloticus
Oryzias latipes
Xiphophorus maculatus
Poecilia reticulata
Fundulus heteroclitus
Salmo salar
Oncorhynchus mykiss
Oncorhynchus kisutch
Esox lucius
Ictalurus punctatus
Astyanax mexicanus
Cyprinus carpio
Danio rerio
Clupea harengus
Lepisosteus oculatus
Latimeria chalumnae
Callorhinchus milii
Branchiostoma belcheri
Branchiostoma floridae

Ciona intestinalis
Strongylocentrotus purpuratus
Lingula anatina
Trichogramma pretiosum
Apis florea
Apis dorsata
Apis cerana
Apis mellifera
Nasonia vitripennis
Musca domestica
Drosophila obscura
Drosophila serrata
Drosophila arizonae
Drosophila willistoni
Drosophila yakuba
Drosophila virilis
Drosophila simulans
Drosophila sechellia
Drosophila pseudoobscura
Drosophila persimilis
Drosophila navojoa
Drosophila mojavensis
Drosophila miranda
Drosophila melanogaster
Drosophila grimshawi
Drosophila erecta
Drosophila ananassae
Ceratitis capitata

Loa Loa
Culex quinquefasciatus
Anopheles gambiae
Aedes albopictus
Aedes aegypti
Bombyx mori
Tribolium castaneum
Bemisia tabaci
Acyrthosiphon pisum
Ixodes scapularis
Limulus polyphemus
Mizuhopecten yessoensis
Biomphalaria glabrata
Aplysia californica
Helobdella robusta
Trichinella spiralis
Brugia malayi
Caenorhabditis elegans
Caenorhabditis briggsae
Opisthorchis viverrini
Schistosoma haematobium
Schistosoma mansoni
Hydra vulgaris
Encephalitozoon cuniculi
Exophiala dermatitidis
Ichthyophthirius multifiliis
Tetrahymena thermophila
Paramecium tetraurelia

Theileria parva
Theileria annulata
Theileria equi
Babesia microti
Babesia bigemina
Babesia bovis
Plasmodium yoelii
Plasmodium vinckei
Plasmodium fragile
Plasmodium vivax
Plasmodium reichenowi
Plasmodium knowlesi
Plasmodium falciparum
Plasmodium cynomolgi
Plasmodium chabaudi
Plasmodium berghei
Toxoplasma gondii
Cryptosporidium muris
Cryptosporidium parvum
Eimeria maxima
Eimeria tenella
Eimeria acervulina
Dictyostelium purpureum
Naegleria gruberi
Entamoeba histolytica
Acanthamoeba castellanii
Giardia intestinalis
Trichomonas vaginalis
Trypanosoma cruzi

Trypanosoma brucei
Leishmania panamensis
Leishmania infantum
Leishmania mexicana
Leishmania major
Leishmania donovani
Leishmania braziliensis
Tsuchiyaea wingfieldii
Saitoella complicata
Alternaria alternata
Trichophyton rubrum
Fusarium graminearum
Fusarium oxysporum
Coccidioides immitis
Candida tropicalis
[Candida] glabrata
Candida albicans
Coprinopsis cinerea
Agaricus bisporus
Schizophyllum commune
Trametes versicolor
Puccinia graminis
Rhodotorula toruloides
Tilletiaria anomala
Ustilago maydis
Tremella mesenterica

Cryptococcus neoformans
Sclerotinia sclerotiorum
Sordaria macrospora
Podospora anserina
Neurospora crassa
Aspergillus oryzae
Aspergillus niger
Aspergillus flavus
Aspergillus clavatus
Aspergillus aculeatus
Histoplasma capsulatum
Leptosphaeria maculans
Bipolaris zeicola
Bipolaris maydis
Kockovaella imperatae
Debaryomyces hansenii
Zygosaccharomyces rouxii
Yarrowia lipolytica
Torulaspora delbrueckii
Saccharomyces cerevisiae
Meyerozyma guilliermondii
Wickerhamomyces anomalus
Pichia membranifaciens
Scheffersomyces stipitis
Pichia kudriavzevii
Cyberlindnera jadinii

Schizosaccharomyces octosporus
Schizosaccharomyces japonicus
Schizosaccharomyces pombe
Phycomyces blakesleeanus
Phytophthora parasitica
Phytophthora infestans
Pneumocystis carinii
Asparagus officinalis
Musa acuminata
Ananas comosus
Zea mays
Sorghum bicolor
Setaria italica
Oryza brachyantha
Oryza sativa
Nelumbo nucifera
Sesamum indicum
Erythranthe guttata
Solanum tuberosum
Nicotiana tomentosiformis
Nicotiana tabacum
Nicotiana sylvestris
Solanum lycopersicum
Capsicum annuum
Daucus carota

Ricinus communis
Vigna angularis
Phaseolus vulgaris
Medicago truncatula
Lupinus angustifolius
Glycine max
Cicer arietinum
Cajanus cajan
Prunus persica
Malus domestica
Raphanus sativus
Brassica oleracea
Brassica rapa
Brassica napus
Arabidopsis thaliana
Populus trichocarpa
Momordica charantia
Cucumis sativus
Cucumis melo
Theobroma cacao
Gossypium hirsutum
Physcomitrella patens
Auxenochlorella protothecoides
Volvox carteri
Chlamydomonas reinhardtii
Emiliania huxleyi
Phaeodactylum tricornutum
Chondrus crispus
Citrus sinensis

Supplementary Table 3, Chapter 7. A complete list of the ortholog groups identified by combineOrthoGroups. This table contains 207 MB of text data. There are 622,843 lines of data, and the first 100 lines fill over 7,000 pages when pasted into this dissertation. As such, I am omitting this supplementary file from my dissertation, although it is available at *Bioinformatics* online. The first column is the gene name for the group with the most named genes. Each subsequent column includes a species name followed by a ":" followed by the gene accession number. If there is a gene annotation for that gene, it follows surrounded by parentheses. For example: "Parus_major:XP_015490149.1(KLHL23)" is a valid entry, where *Parus major* is the species name, XP_015490149.1 is the gene accession, and KLHL23 is the gene annotation.

Supplementary Table 4, Chapter 7. Statistics for each ortholog group. Same as Table 2 from the main text, except all ortholog groups are included and the last column is not included. There are 622,844 lines on data in this file. We include the first 500 lines in this dissertation, although the whole table is available upon request and at *Bioinformatics* online.

Same Annotation	Other Annotation	Unknown Annotation	Total Genes
127	0	63	190
178	0	7	185
172	1	7	180
155	2	21	178
169	0	9	178
169	1	5	175
166	0	5	171
165	1	5	171
163	1	6	170
165	0	4	169
161	0	7	168
162	0	5	167
161	1	4	166
163	0	3	166
152	1	13	166
161	0	4	165
156	0	9	165
159	0	6	165
160	0	5	165
160	0	4	164
159	0	5	164
158	0	5	163
156	1	5	162
156	0	5	161
158	0	3	161
153	0	7	160
149	0	9	158

154	0	3	157
146	0	11	157
153	0	4	157
139	1	15	155
151	0	3	154
148	0	5	153
149	0	4	153
145	0	7	152
150	0	2	152
144	0	7	151
146	0	4	150
146	1	3	150
141	0	8	149
142	0	7	149
138	0	10	148
127	1	20	148
145	0	3	148
143	0	4	147
141	0	6	147
142	0	4	146
141	0	4	145
131	5	9	145
135	6	3	144
139	0	5	144
26	33	84	143
138	0	5	143
136	1	5	142
139	0	3	142
135	1	6	142
118	3	19	140
127	0	13	140
136	0	3	139
130	0	9	139
135	0	4	139
25	35	79	139
125	2	11	138
136	0	2	138
120	2	15	137
133	0	4	137
135	0	2	137
132	0	5	137
132	0	5	137
132	0	5	137

133	0	4	137
130	1	5	136
131	0	5	136
129	0	7	136
127	0	8	135
132	0	3	135
98	31	5	134
122	1	11	134
128	0	5	133
128	2	2	132
127	0	5	132
130	0	2	132
127	0	4	131
125	1	5	131
100	26	4	130
126	0	4	130
74	32	24	130
118	0	12	130
126	0	3	129
125	0	4	129
126	0	3	129
127	0	2	129
16	25	88	129
120	0	8	128
120	0	7	127
99	25	3	127
121	0	6	127
122	0	5	127
117	0	10	127
110	0	16	126
92	12	21	125
18	28	79	125
122	0	3	125
121	1	3	125
97	0	27	124
20	31	73	124
122	0	1	123
81	2	40	123
121	0	2	123
118	0	4	122
99	0	23	122
119	1	2	122
97	7	17	121

96	21	4	121
115	0	5	120
116	1	3	120
95	22	3	120
114	0	6	120
90	3	27	120
116	0	4	120
117	0	2	119
72	43	4	119
113	0	5	118
98	13	6	117
94	6	17	117
114	0	3	117
110	1	6	117
24	3	90	117
42	10	64	116
114	0	2	116
110	0	5	115
108	0	6	114
106	1	7	114
111	0	3	114
107	1	6	114
110	1	3	114
109	0	4	113
112	0	1	113
109	0	3	112
1	11	100	112
106	1	5	112
107	1	3	111
87	4	20	111
99	1	11	111
107	0	4	111
7	1	102	110
106	0	4	110
107	0	3	110
16	26	68	110
77	0	32	109
107	0	2	109
81	1	27	109
104	0	5	109
106	1	2	109
101	1	6	108
84	1	23	108

98	2	8	108
102	0	6	108
84	0	23	107
102	0	5	107
94	7	5	106
103	0	3	106
104	0	2	106
91	0	15	106
101	0	4	105
102	0	3	105
97	4	4	105
102	0	3	105
88	8	8	104
100	0	4	104
94	6	4	104
102	0	2	104
90	8	6	104
96	0	8	104
93	0	11	104
102	0	2	104
100	0	3	103
101	0	2	103
73	2	28	103
94	5	4	103
101	0	2	103
81	0	22	103
100	0	3	103
101	0	2	103
95	0	8	103
101	0	2	103
99	0	4	103
100	0	3	103
101	0	2	103
99	0	4	103
99	0	4	103
99	0	3	102
100	0	2	102
97	2	3	102
100	0	2	102
98	0	4	102
97	0	5	102
100	0	2	102
100	0	2	102

99	0	3	102
98	0	4	102
82	0	20	102
53	0	49	102
97	0	5	102
96	1	5	102
98	0	4	102
94	0	8	102
98	1	3	102
1	1	99	101
85	11	5	101
97	0	4	101
6	0	95	101
99	0	2	101
98	0	3	101
97	1	3	101
96	0	5	101
98	1	2	101
99	0	2	101
98	0	3	101
98	0	3	101
97	0	4	101
95	0	6	101
97	0	3	100
97	0	3	100
97	0	3	100
96	0	4	100
96	0	4	100
79	0	21	100
94	0	6	100
96	0	4	100
96	0	4	100
96	0	4	100
98	0	2	100
93	0	7	100
98	0	2	100
96	1	3	100
96	0	4	100
99	0	1	100
98	0	2	100
98	0	2	100
98	0	2	100
98	0	2	100

1	0	99	100
98	0	2	100
98	0	2	100
97	0	3	100
16	24	60	100
97	0	2	99
92	0	7	99
97	0	2	99
94	0	5	99
97	0	2	99
97	0	2	99
97	0	2	99
94	1	4	99
97	0	2	99
97	0	2	99
87	2	10	99
93	1	5	99
88	1	10	99
97	0	2	99
92	0	7	99
96	0	3	99
95	0	4	99
95	0	4	99
97	0	2	99
97	0	2	99
95	0	4	99
94	0	5	99
88	0	10	98
92	0	6	98
96	0	2	98
96	1	1	98
95	1	2	98
95	0	3	98
94	0	4	98
94	0	4	98
96	0	2	98
95	0	3	98
95	0	3	98
96	0	2	98
95	0	3	98
96	0	2	98
96	0	2	98
94	0	4	98

96	0	2	98
96	0	2	98
88	0	9	97
51	0	46	97
94	1	2	97
95	0	2	97
93	0	4	97
92	0	5	97
95	0	2	97
95	0	2	97
95	0	2	97
89	0	8	97
94	0	3	97
94	1	2	97
95	0	2	97
95	0	2	97
89	2	6	97
95	0	2	97
81	2	14	97
95	0	2	97
51	5	41	97
94	0	3	97
95	0	2	97
95	0	2	97
94	0	3	97
95	0	2	97
88	1	8	97
95	0	2	97
95	0	2	97
80	3	13	96
94	0	2	96
92	2	2	96
94	0	2	96
94	0	2	96
92	0	4	96
92	2	2	96
83	10	3	96
92	1	3	96
94	0	2	96
90	0	6	96
95	0	1	96
93	0	3	96
94	0	2	96

91	0	5	96
95	0	1	96
92	0	4	96
93	0	3	96
93	0	3	96
94	0	2	96
19	20	57	96
18	20	58	96
90	1	5	96
83	0	13	96
93	0	3	96
94	0	2	96
79	0	17	96
94	0	2	96
89	0	7	96
87	0	9	96
69	23	4	96
91	0	5	96
12	28	56	96
93	0	3	96
86	2	7	95
91	0	4	95
92	0	3	95
85	8	2	95
93	0	2	95
86	0	9	95
90	1	4	95
87	0	8	95
92	0	3	95
93	0	2	95
91	0	4	95
93	0	2	95
93	0	2	95
93	0	2	95
93	0	2	95
92	0	3	95
92	1	2	95
92	0	3	95
92	0	3	95
93	0	2	95
93	0	2	95
93	0	2	95
94	0	1	95

92	0	3	95
93	0	2	95
88	0	7	95
93	0	2	95
92	0	2	94
92	0	2	94
92	0	2	94
89	0	5	94
90	0	4	94
89	0	5	94
1	0	93	94
91	0	3	94
92	0	2	94
92	0	2	94
92	0	2	94
92	0	2	94
90	0	4	94
92	0	2	94
88	0	6	94
91	0	3	94
92	0	2	94
91	0	3	94
91	0	3	94
67	23	4	94
92	0	2	94
92	0	2	94
92	0	2	94
90	0	4	94
90	0	4	94
92	0	2	94
89	0	5	94
87	0	6	93
90	0	3	93
89	2	2	93
91	0	2	93
91	0	2	93
91	0	2	93
13	24	56	93
89	0	4	93
91	0	2	93
89	0	4	93
90	0	3	93
91	0	2	93

91	0	2	93
12	25	56	93
90	0	3	93
91	0	2	93
91	0	2	93
89	2	2	93
88	0	5	93
89	2	2	93
90	0	2	92
90	0	2	92
90	0	2	92
90	0	2	92
87	0	5	92
90	0	2	92
89	0	3	92
86	0	6	92
60	29	3	92
88	0	4	92
90	0	2	92
88	0	4	92
87	0	5	92
88	0	4	92
90	0	2	92
88	0	4	92
90	0	2	92
84	0	8	92
90	0	2	92
89	0	3	92
90	0	2	92
87	0	5	92
89	0	3	92
14	22	56	92
88	0	4	92
88	0	3	91
89	0	2	91
89	0	2	91
86	1	4	91
86	0	5	91
88	0	3	91
89	0	2	91
89	0	2	91
89	0	2	91
88	0	3	91

86	0	5	91
88	0	3	91
87	0	4	91
88	0	3	91
87	0	4	91
88	0	3	91
89	0	2	91
87	0	4	91
88	0	3	91
88	0	3	91
89	0	2	91
88	0	3	91
89	0	2	91
89	0	2	91
88	0	3	91
88	0	3	91
87	0	3	90
14	24	52	90
87	0	3	90
88	0	2	90
88	0	2	90
88	0	2	90
84	1	5	90
87	0	3	90
87	0	3	90
88	0	2	90
88	0	2	90
89	0	1	90
88	0	2	90
88	0	2	90
88	0	2	90
88	0	2	90
88	0	2	90
87	0	3	90
86	0	4	90
88	0	2	90
88	0	2	90
88	0	2	90
88	0	2	90
88	0	2	90
86	0	4	90
71	17	2	90
87	0	3	90

Supplementary Table 5, Chapter 7. A snapshot of how many genes are in each ortholog group. The first column is the number of genes in an ortholog group. The second column is the number of ortholog groups with that many genes in the group.

Number of Genes in the Ortholog Group	Number of Groups with that many genes
190	1
185	1
180	1
178	2
175	1
171	2
170	1
169	1
168	1
167	1
166	3
165	4
164	2
163	1
162	1
161	2
160	1
158	1
157	3
155	1
154	1
153	2
152	2
151	1
150	2
149	2
148	3
147	2
146	1
145	2
144	2
143	2
142	3
140	2
139	4
138	2
137	7
136	3
135	2
134	2
133	1
132	3
131	2
130	4
129	5
128	1
127	5
126	1
125	4
124	2
123	3
122	3
121	2
120	6
119	2
118	1
117	5
116	2
115	1
114	5
113	2
112	3
111	4
110	4
109	5
108	4
107	2
106	4
105	4
104	8
103	15
102	17
101	14
100	25
99	22

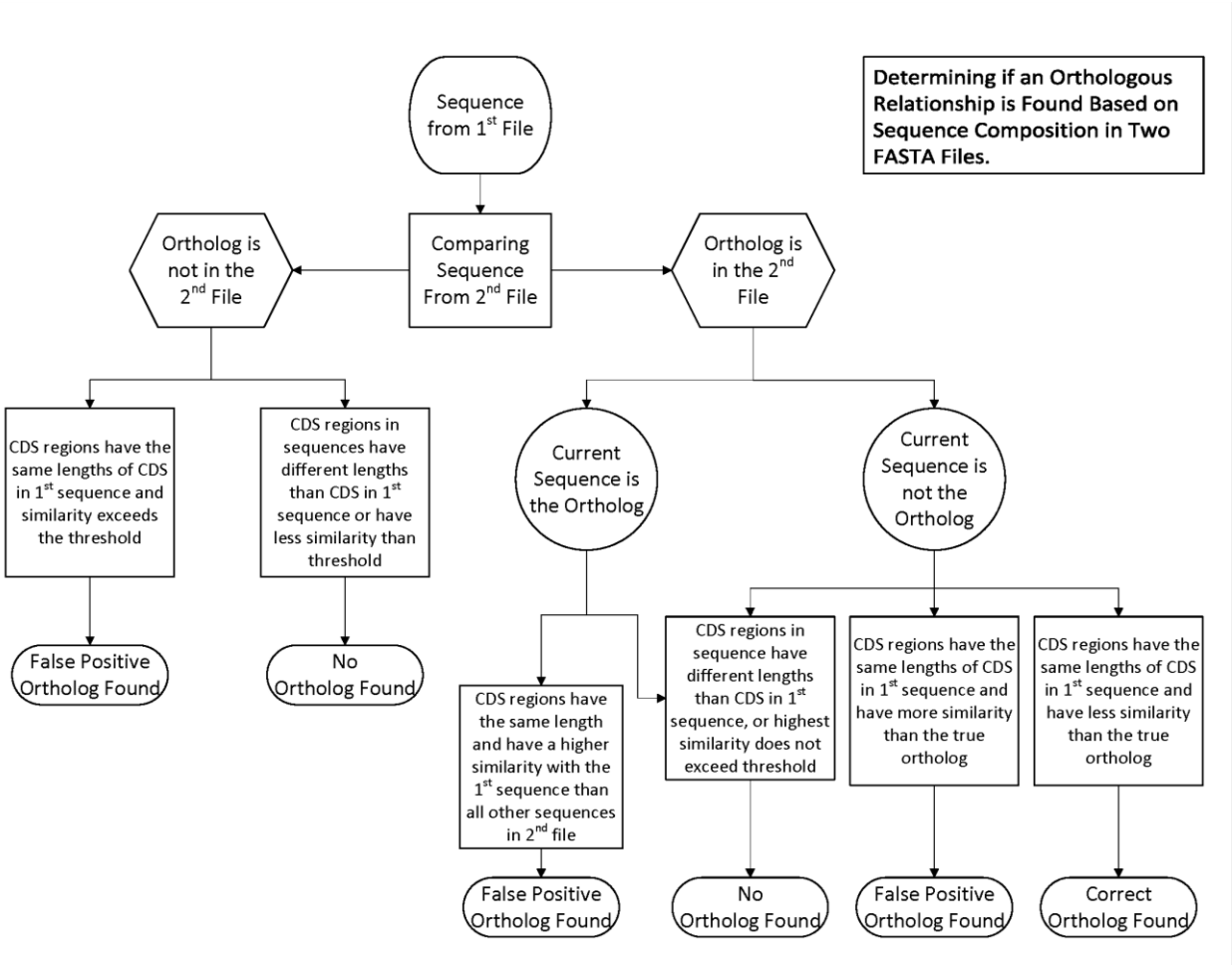
98	18
97	27
96	34
95	27
94	27
93	20
92	25
91	26
90	33
89	23
88	24
87	32
86	20
85	25
84	25
83	27
82	28
81	28
80	33
79	29
78	17
77	16
76	25
75	17
74	20
73	28
72	12
71	26
70	24
69	24
68	23
67	21
66	28
65	28
64	26
63	26
62	32
61	23
60	21
59	24
58	30
57	32
56	24

55	26
54	32
53	33
52	37
51	39
50	46
49	26
48	24
47	37
46	35
45	44
44	32
43	46
42	50
41	62
40	59
39	80
38	61
37	113
36	130
35	123
34	143
33	188
32	227
31	229
30	232
29	259
28	267
27	380
26	460
25	536
24	513
23	488
22	484
21	533
20	633
19	633
18	703
17	714
16	871
15	929
14	1148
13	1352

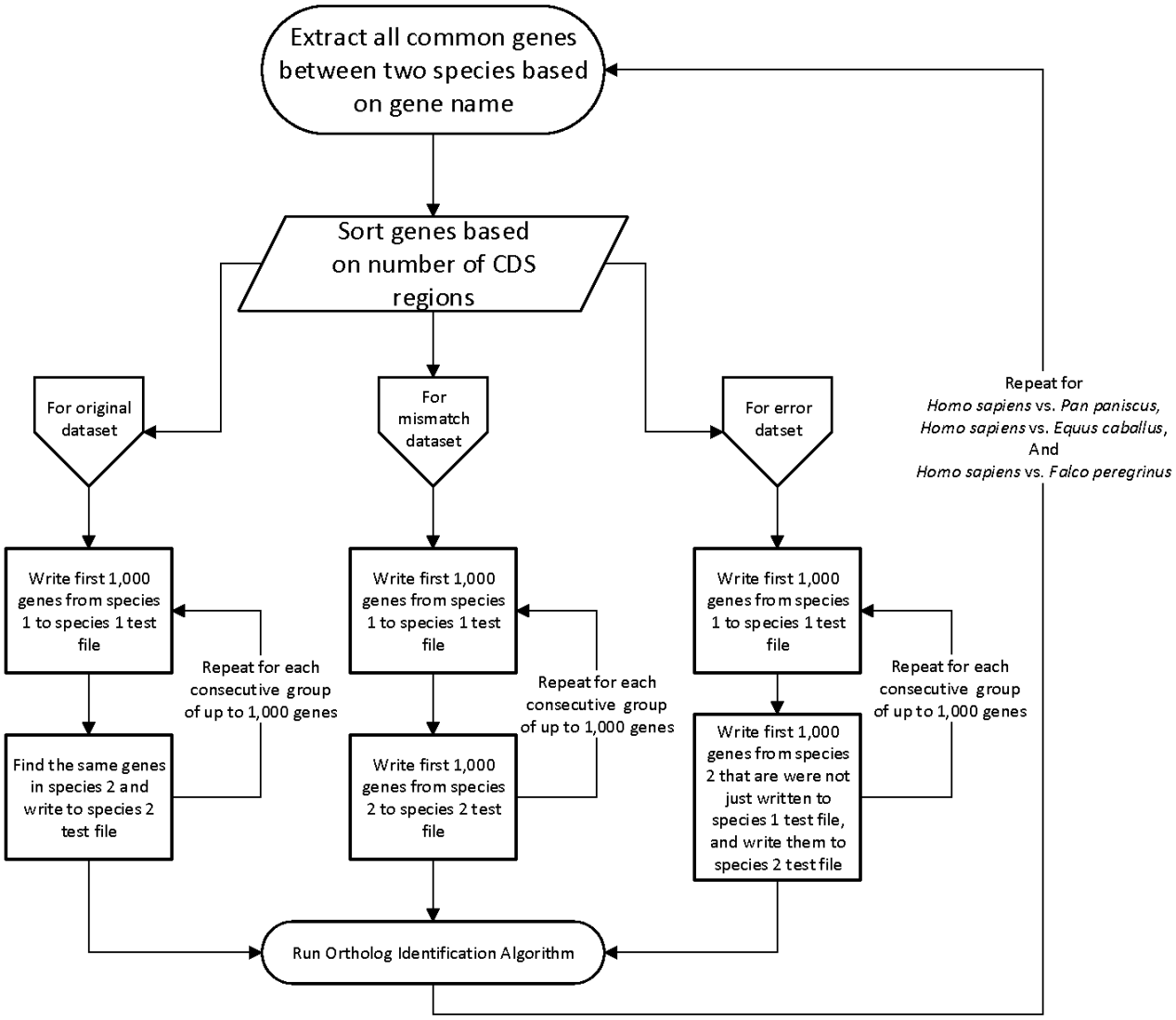
12	1790
11	2568
10	2776
9	3703
8	5559
7	7900

6	12117
5	24286
4	46429
3	114869
2	386442

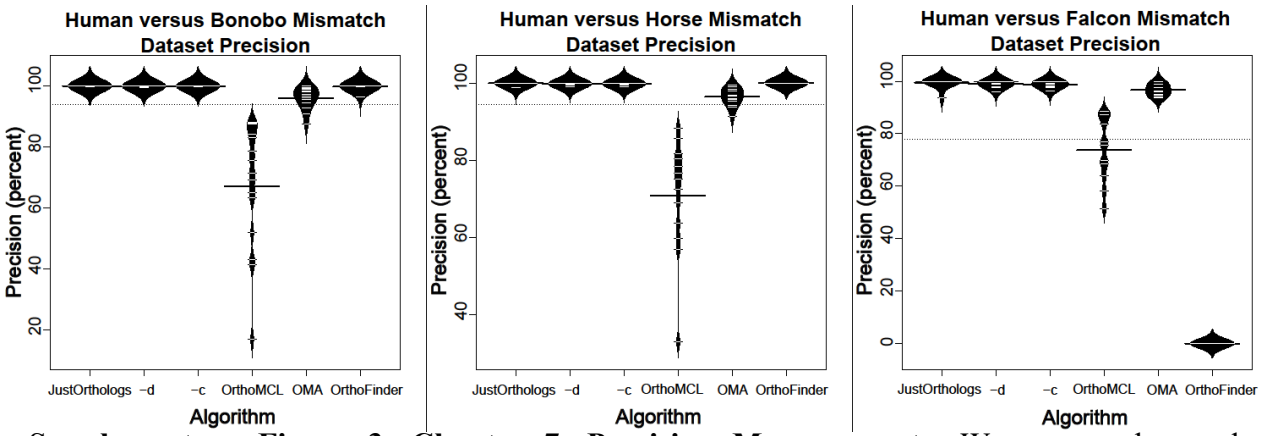
Supplementary Figures



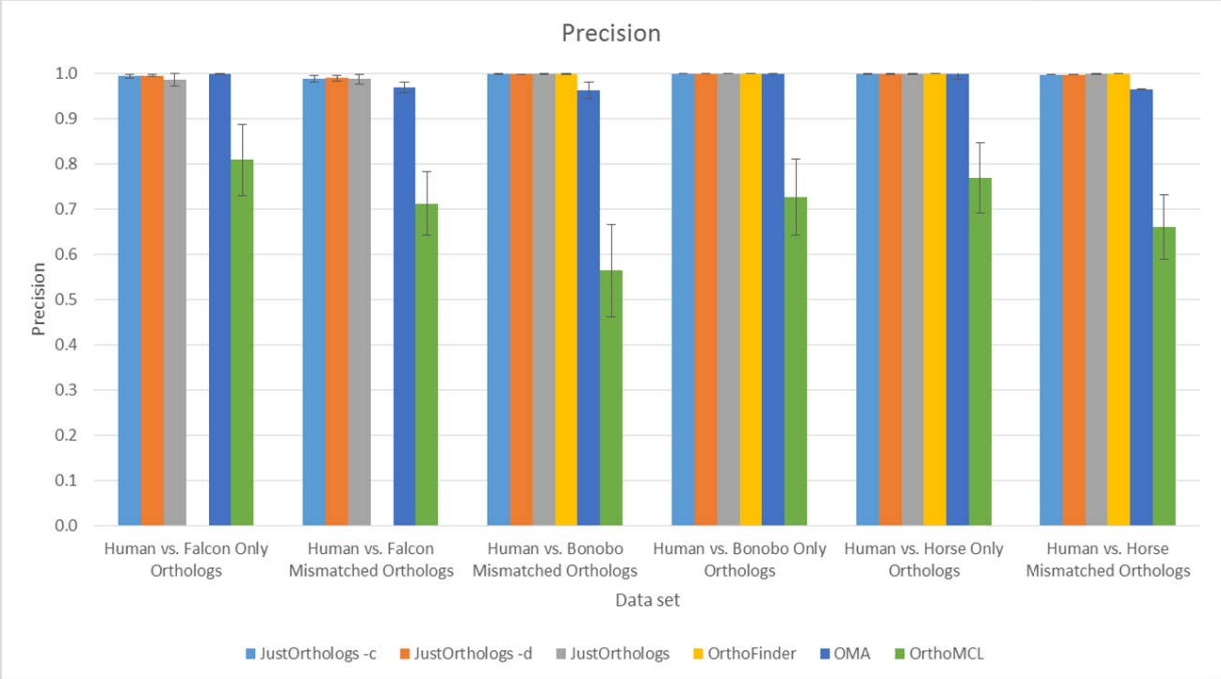
Supplementary Figure 1, Chapter 7. JustOrthologs decision process. A query sequence is selected from the first file and is compared to one sequence at a time in the second file (subject file), and the processes JustOrthologs follows are outlined. The first sequence is from the first file (containing gene sequences from the first species), and the second sequence is from the second file (containing gene sequences from the second species). Details describing CDS comparison and dinucleotide percentage thresholds are described in the text. Similarity refers to the comparison of dinucleotide compositions of each exon.



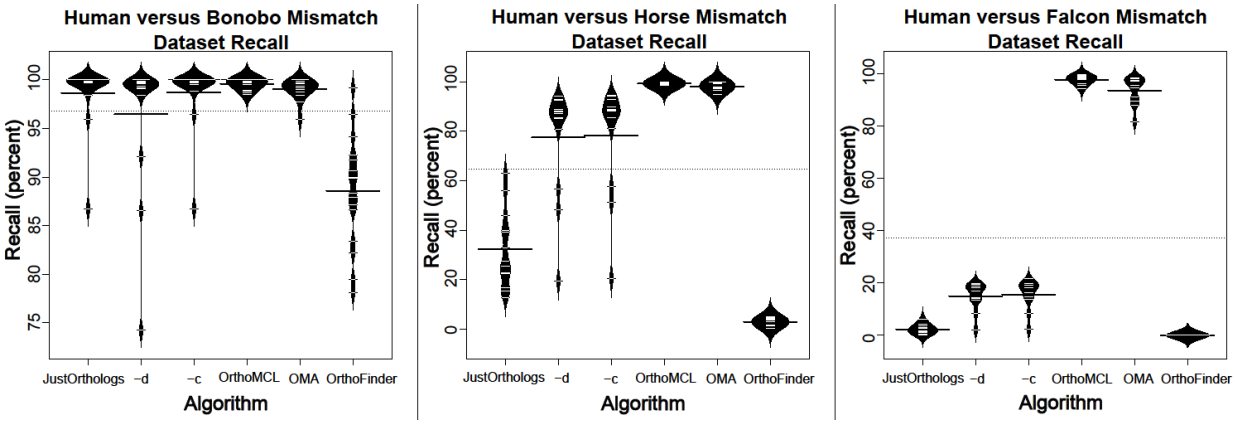
Supplementary Figure 2, Chapter 7. Test Set Creation. Three different types of test data sets are created. Original (left) data sets are all true positive orthologs, mismatch data sets (middle) are a random mix of true and false positive orthologs, and error data sets (right) contain no true orthologs. Each data set contains up to 1 000 sequences.



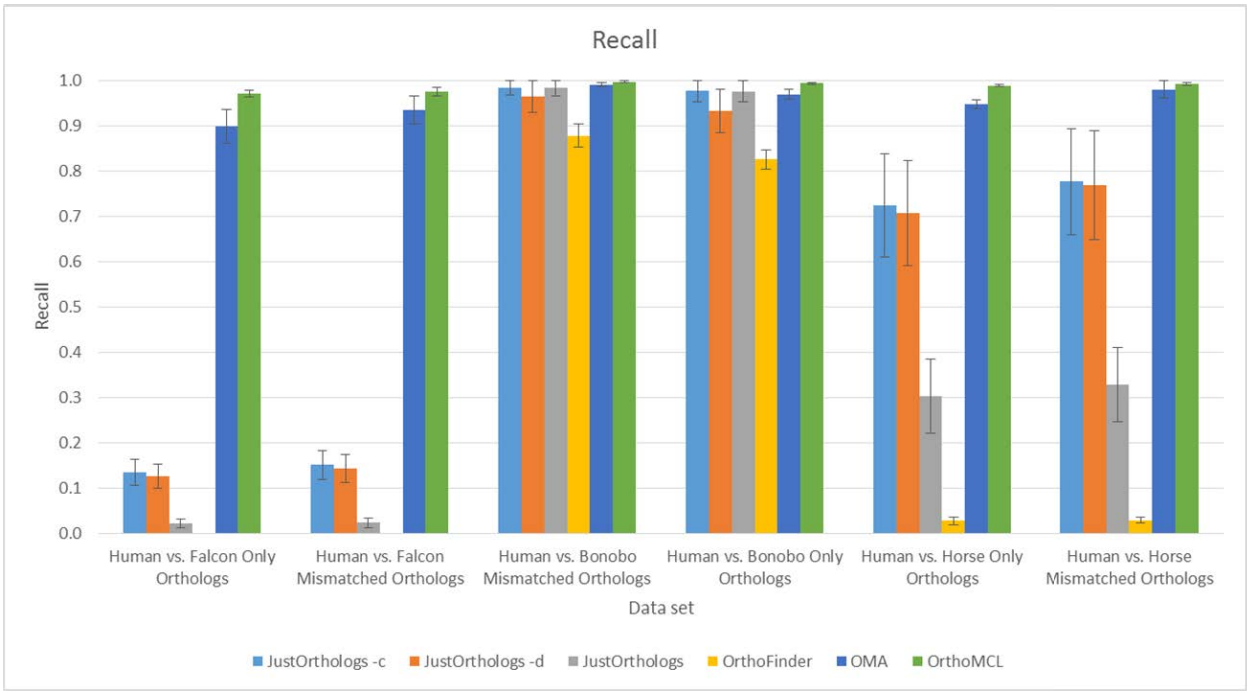
Supplementary Figure 3, Chapter 7. Precision Measurements. We present bean plots comparing the precision of the three different settings for JustOrthologs to OrthoMCL, OMA, and OrthoFinder for humans versus bonobos (left), humans versus horse (middle), and humans versus falcon (right). Bean plots display the individual tests as horizontal bars within a shaded density distribution, which allows for easy identification of outliers. The darker horizontal bar shown for each test set is the sample mean, and the dotted line that spans the entire chart is the overall mean from all samples and tests (Kampstra, 2008). Results in this figure are from mismatch test data sets (a mix of real and not real orthologs), which are the best approximation of a real data set. The x-axis labels refer to the algorithm used. From left to right, the algorithms used are: JustOrthologs for closely related species, JustOrthologs for distantly related species, JustOrthologs combined approach, OrthoMCL, OMA, and OrthoFinder.



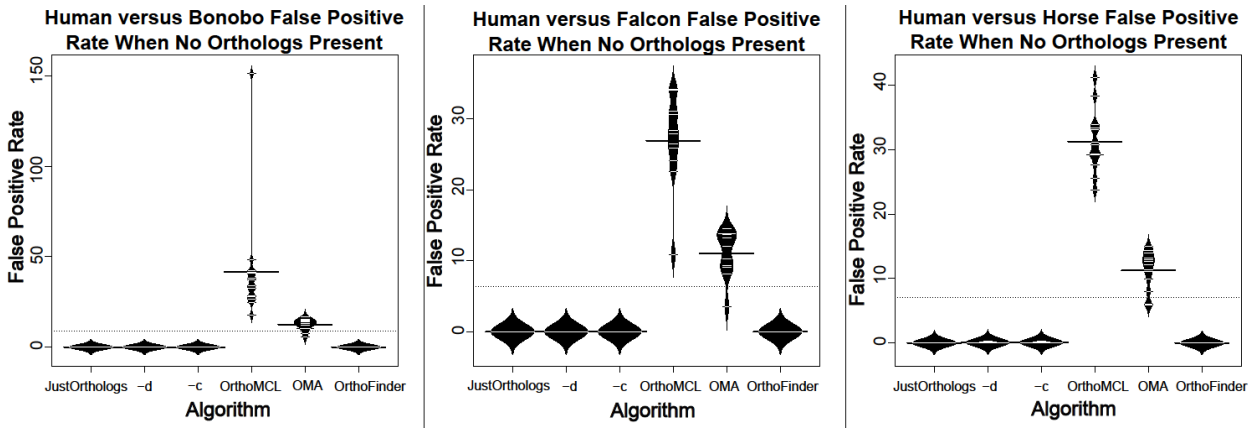
Supplementary Figure 4, Chapter 7. Precision for all datasets



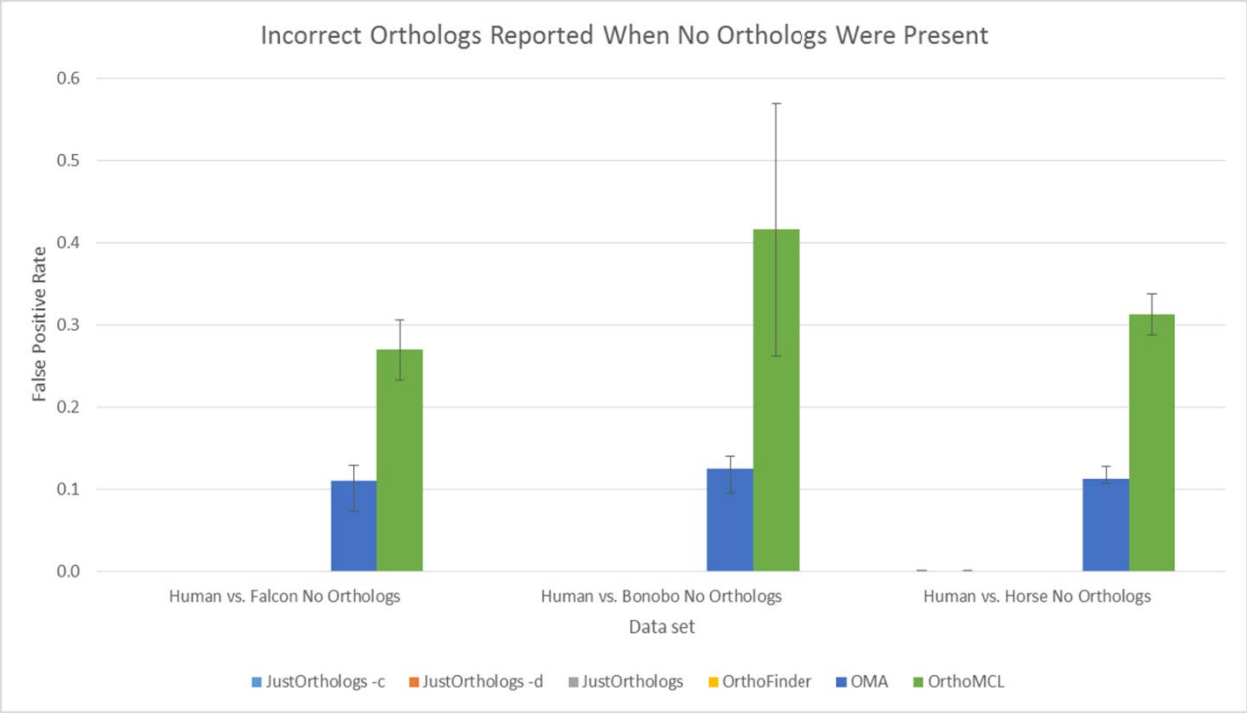
Supplementary Figure 5, Chapter 7. Recall Measurements. We present bean plots comparing the recall of the three different settings for JustOrthologs to OrthoMCL, OMA, and OrthoFinder for humans versus bonobos (left), humans versus horse (middle), and humans versus falcon (right). Results are from mismatch test data sets (a mix of real and not real orthologs), which are the best approximation of a real data set. Bean plots and the x-axis labels are as described in Fig. 3.



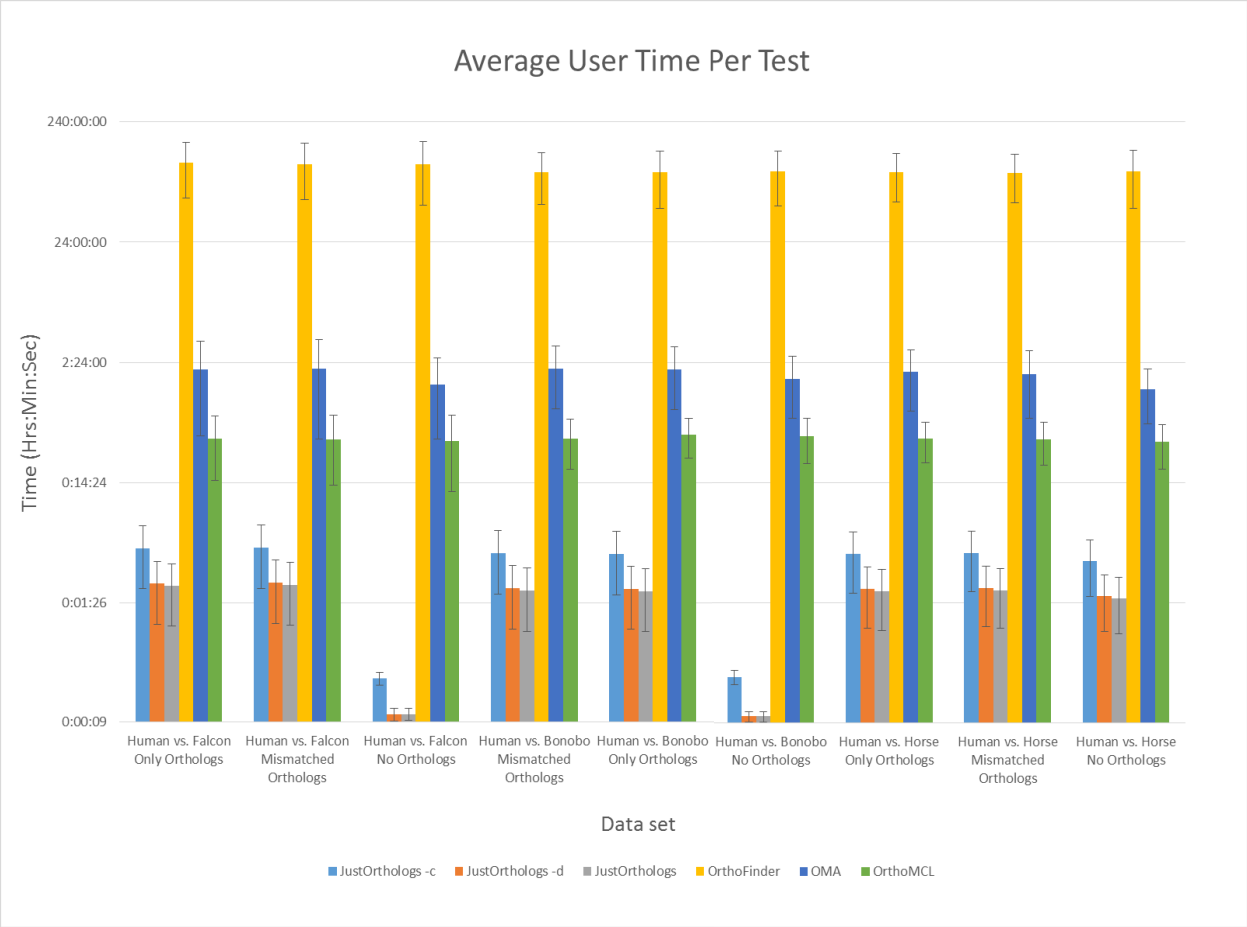
Supplementary Figure 6, Chapter 7. Recall for all datasets



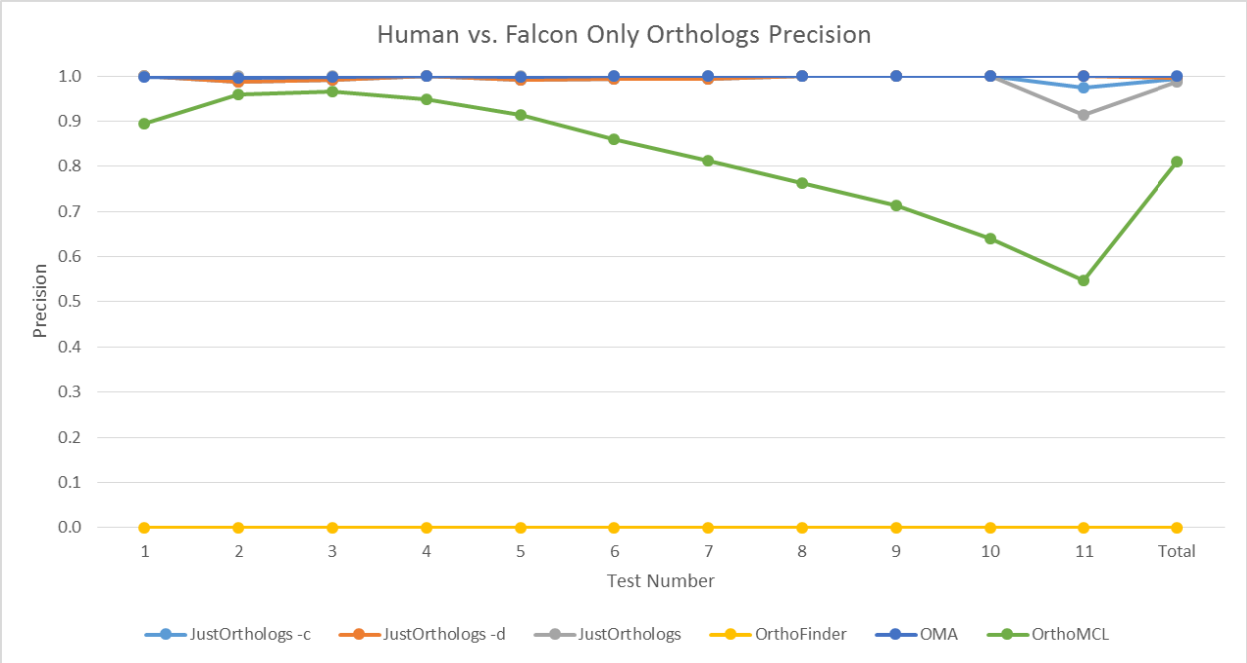
Supplementary Figure 7. False Positive Rate. We present bean plots comparing the false positive rate rate of the three different settings for JustOrthologs, OrthoMCL, OMA, and OrthoFinder for humans versus bonobos (left), humans versus horse (middle), and humans versus falcon (right). Results are from the error dataset, with no true orthologs. These graphs show how many orthologs are reported by each algorithm when no orthologs are present. Bean plots and the x-axis labels are as described in Fig. 3.



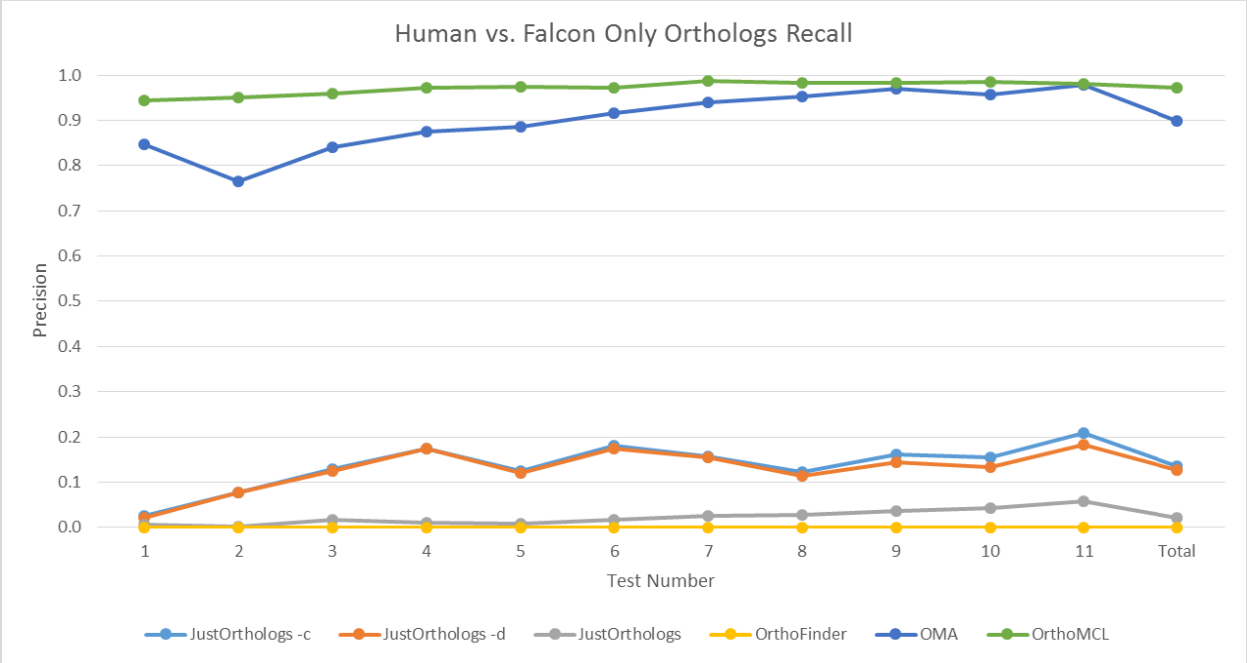
Supplementary Figure 8, Chapter 7. False Positive Rate for all datasets



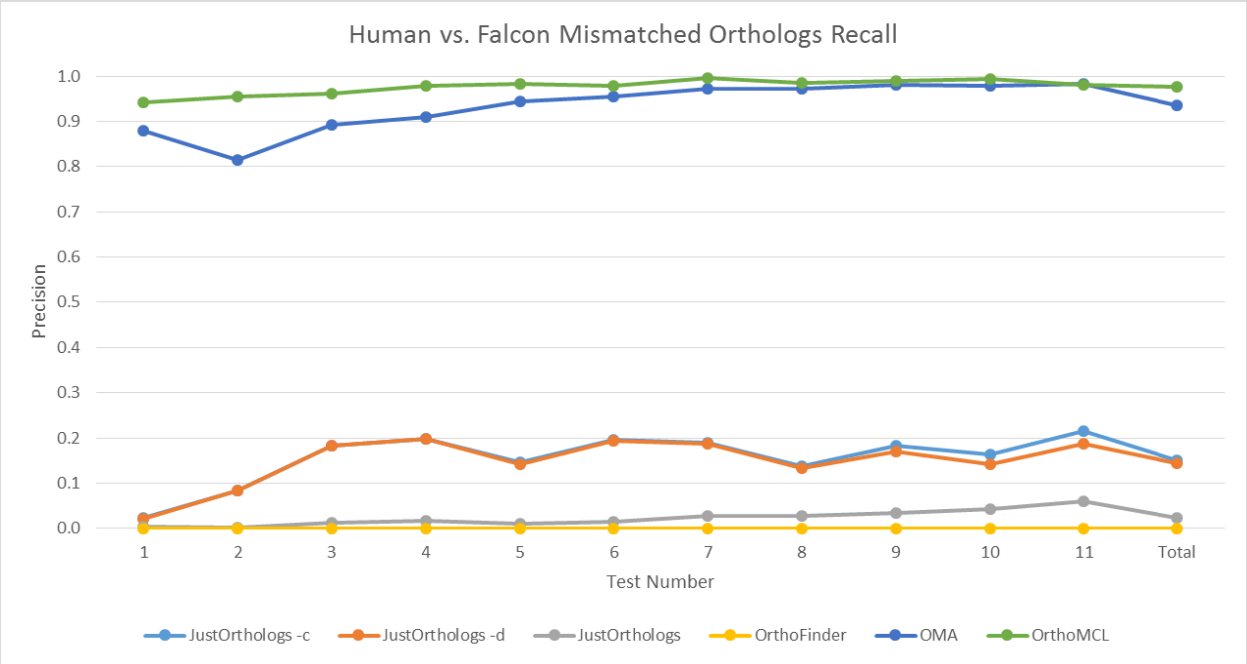
Supplementary Figure 9, Chapter 7. User Time for all datasets



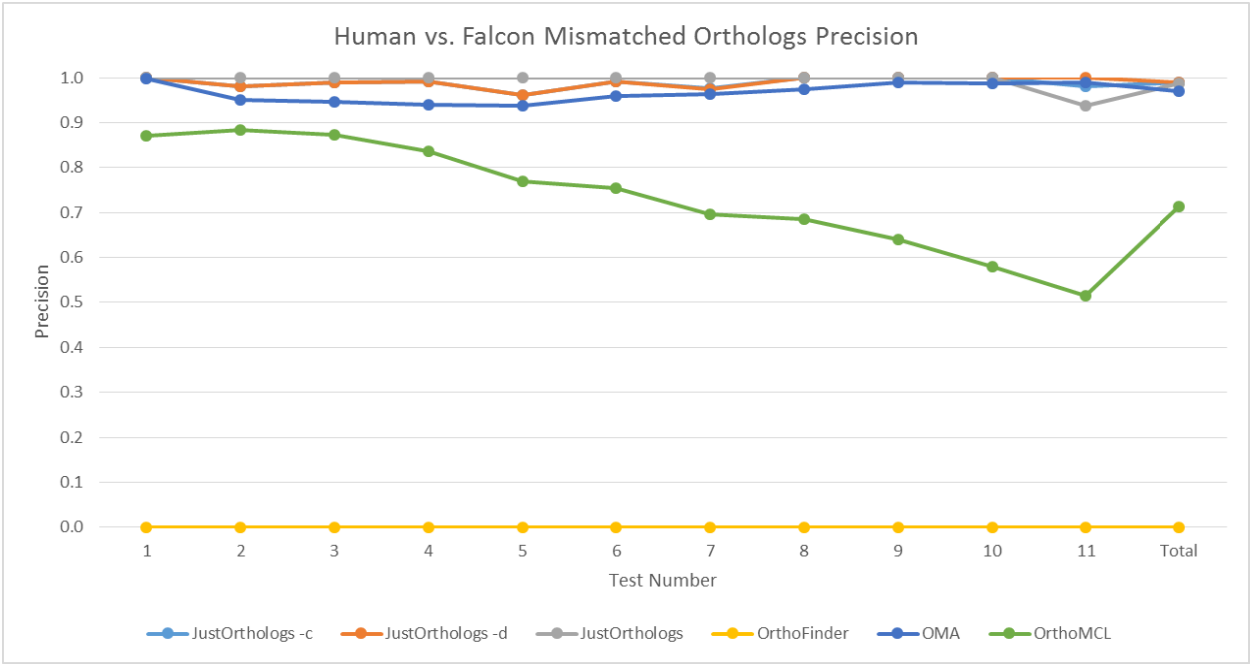
Supplementary Figure 10, Chapter 7. Precision for humans versus falcons for each test case where only orthologs are present



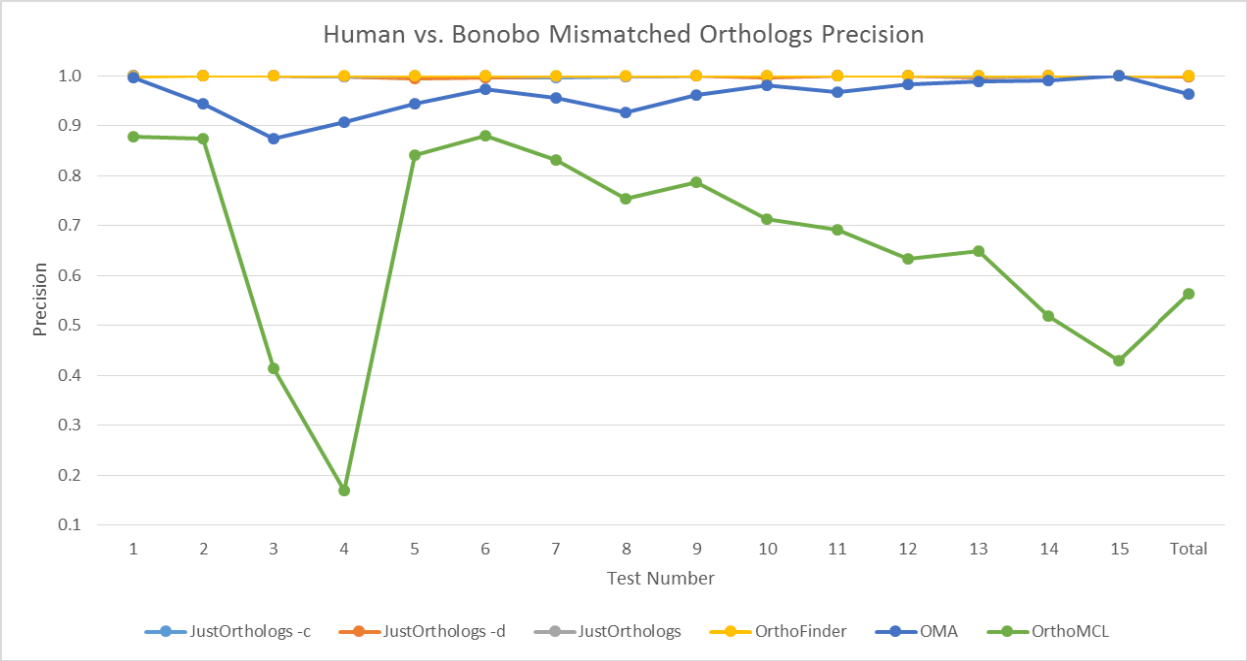
Supplementary Figure 11, Chapter 7. Recall for humans versus falcons for each test case where only orthologs are present



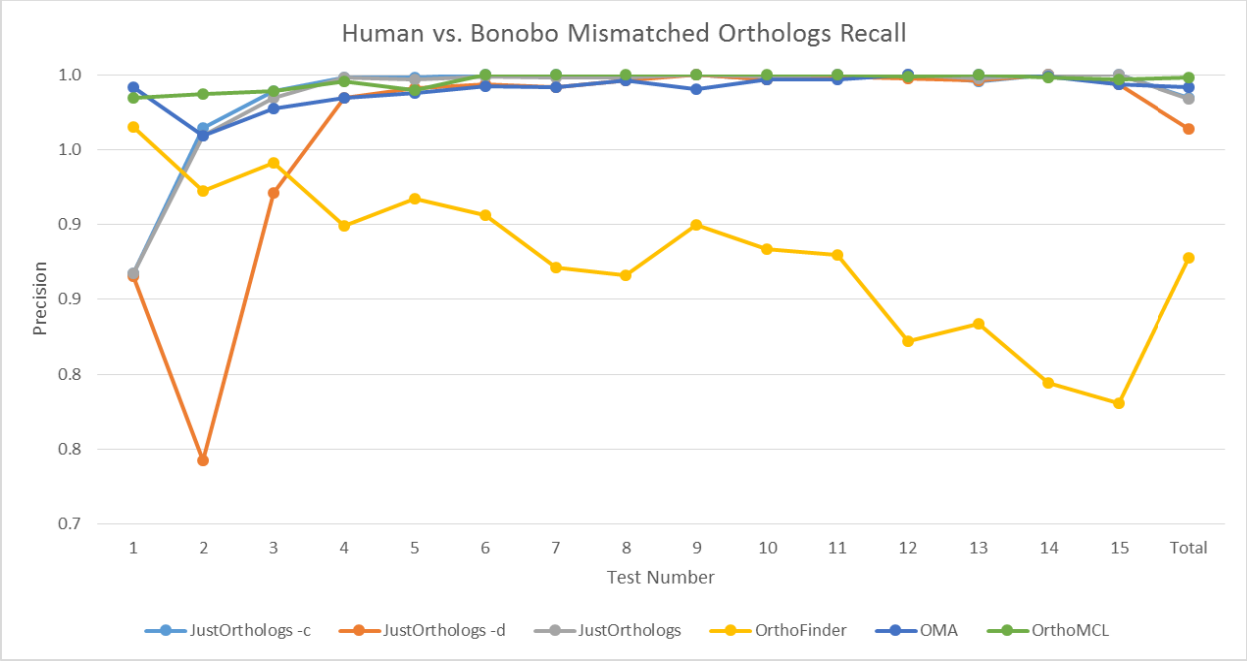
Supplementary Figure 12, Chapter 7. Recall for humans versus falcons for each test case where some orthologs are present



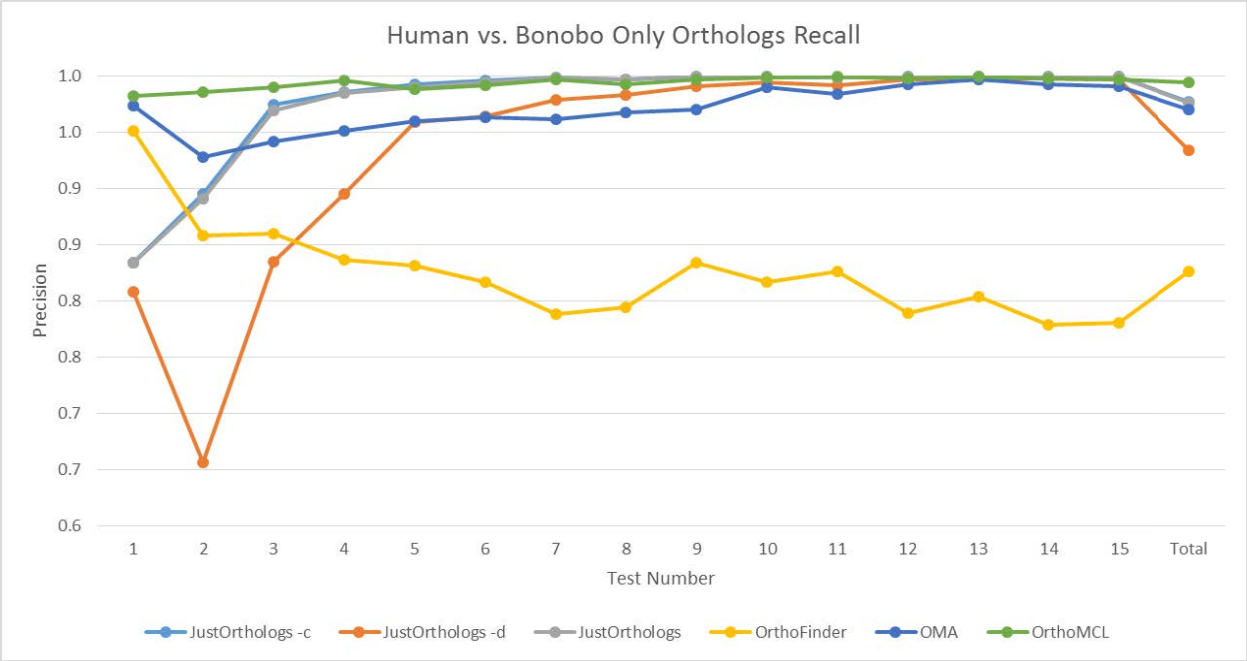
Supplementary Figure 13, Chapter 7. Precision for humans versus falcons for each test case where some orthologs are present



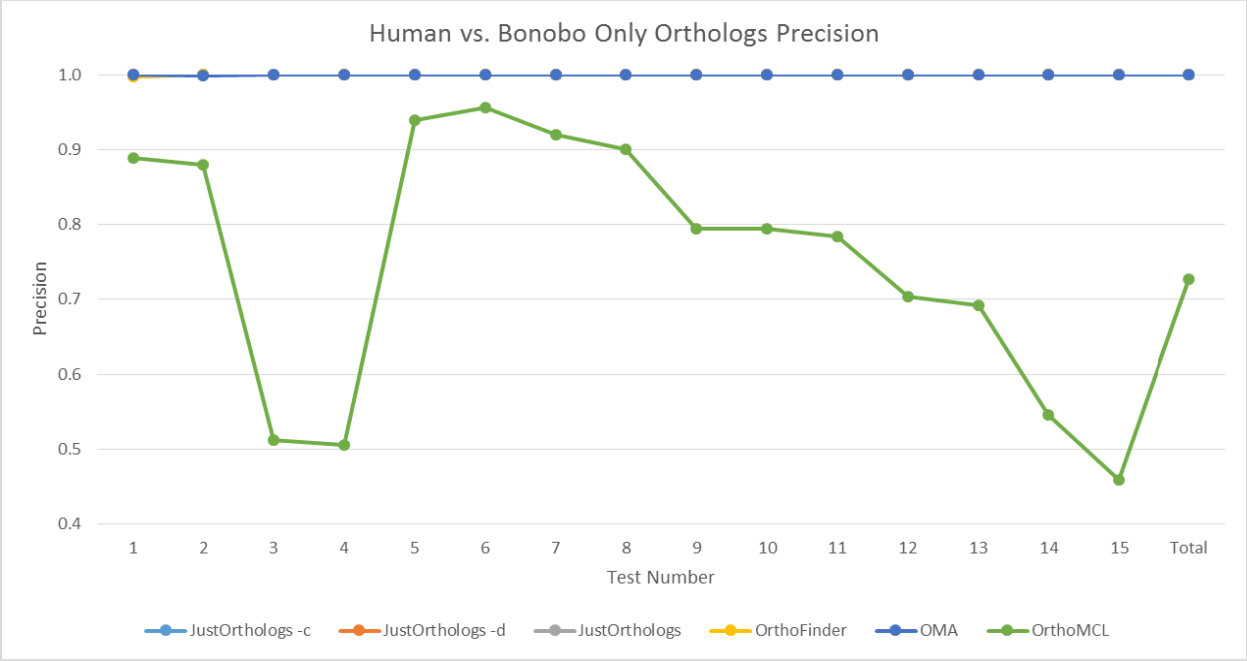
Supplementary Figure 14, Chapter 7. Precision for humans versus bonobo for each test case where some orthologs are present



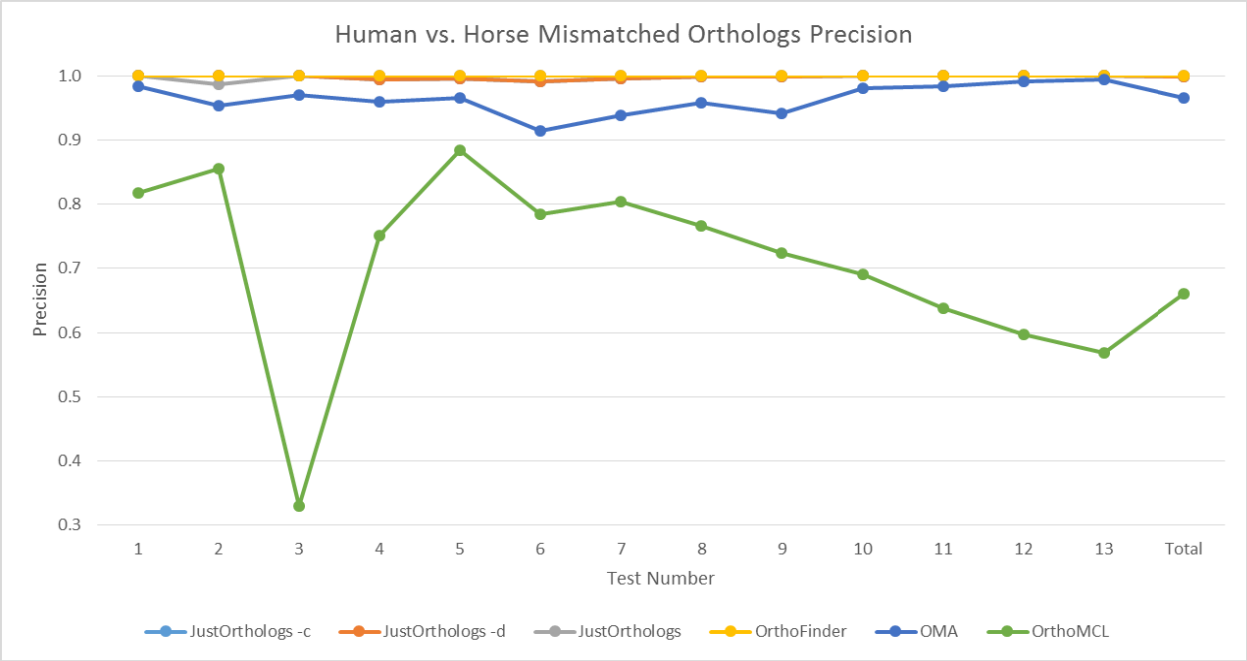
Supplementary Figure 15, Chapter 7. Recall for humans versus bonobo for each test case where some orthologs are present



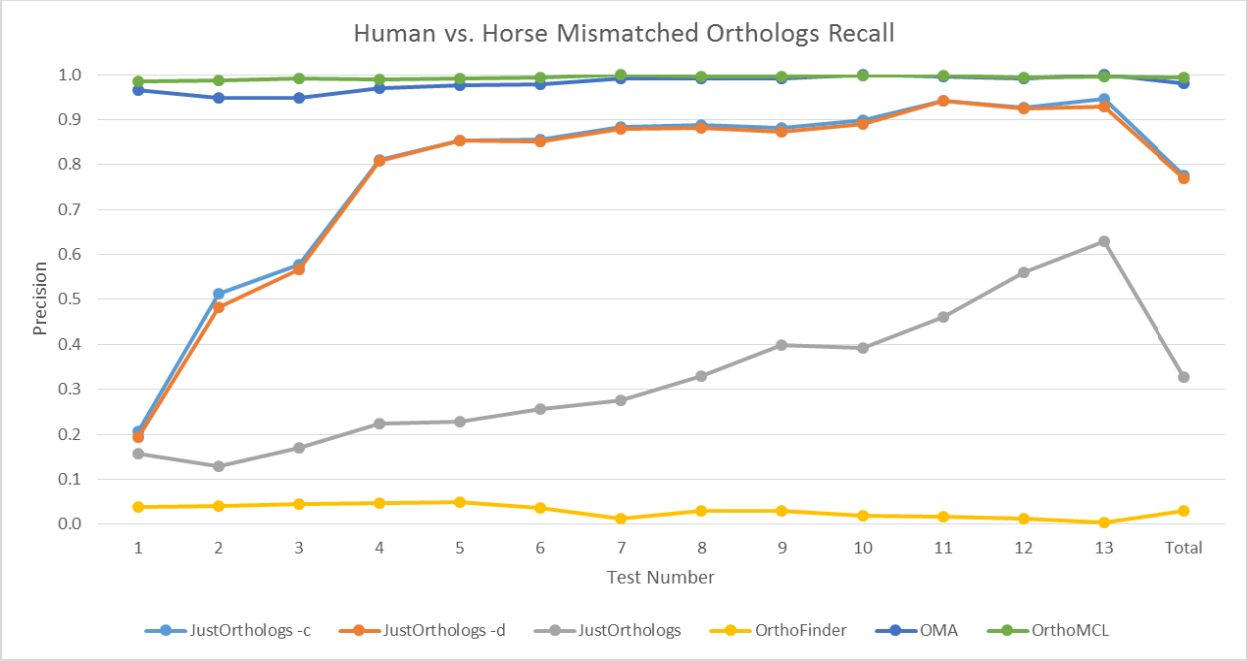
Supplementary Figure 16, Chapter 7. Recall for humans versus bonobo for each test case where only orthologs are present



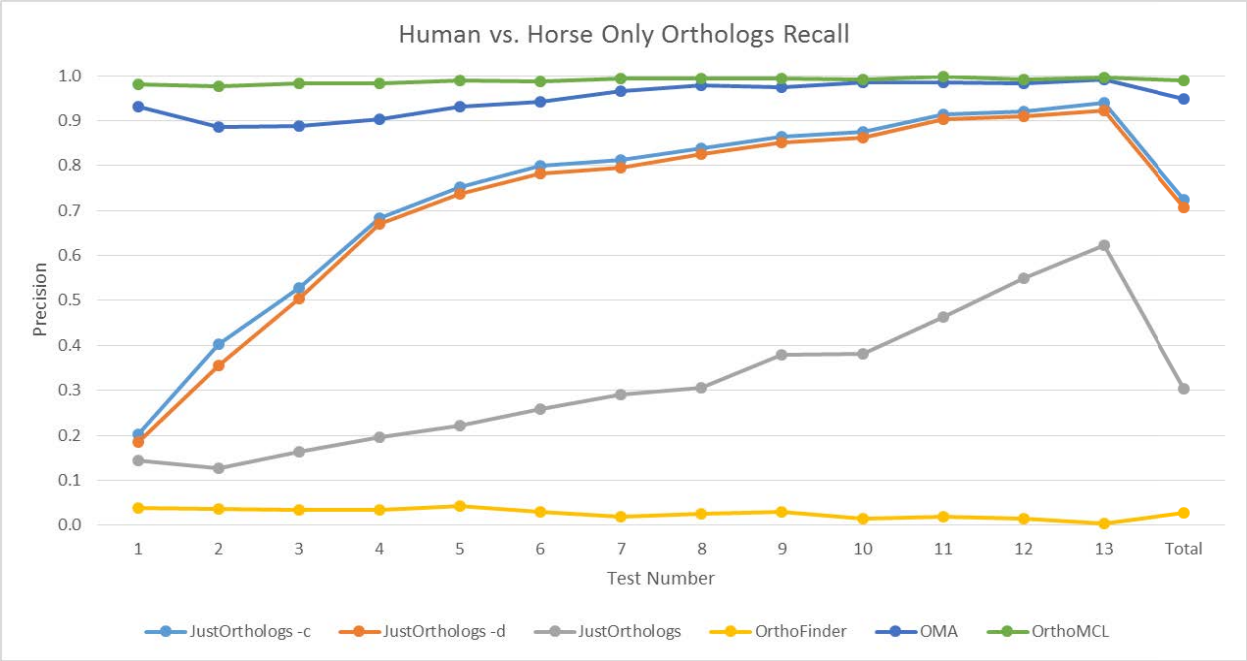
Supplementary Figure 17, Chapter 7. Precision for humans versus bonobo for each test case where only orthologs are present



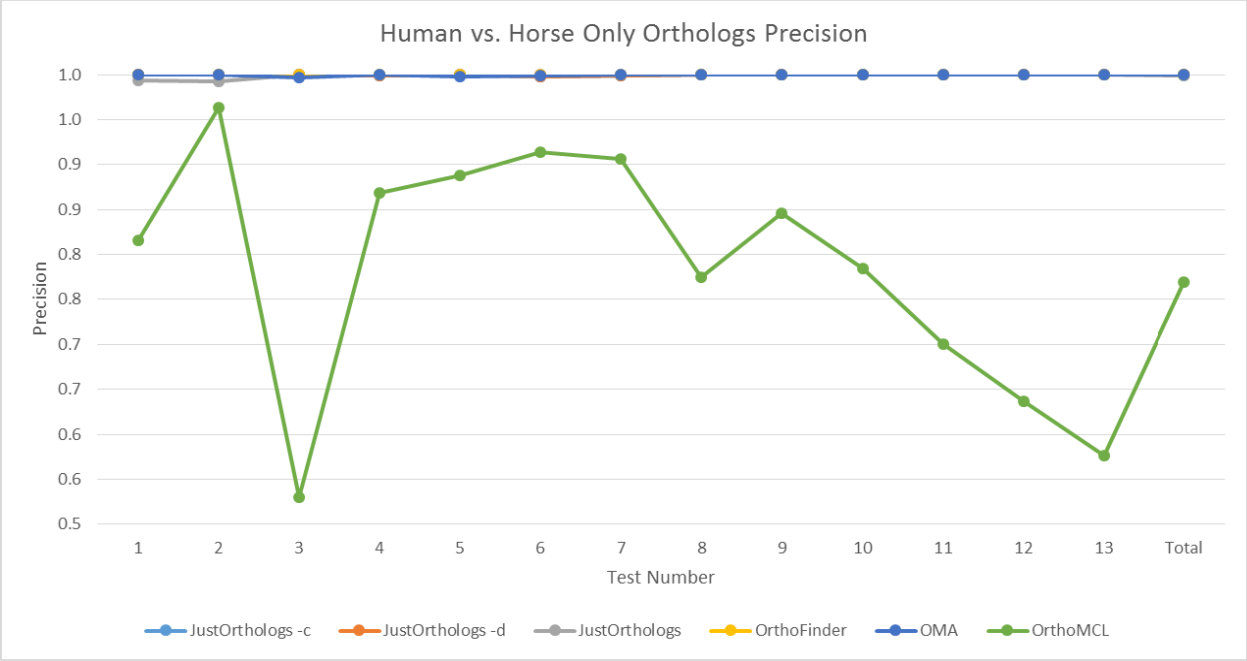
Supplementary Figure 18, Chapter 7. Precision for humans versus horse for each test case where only orthologs are present



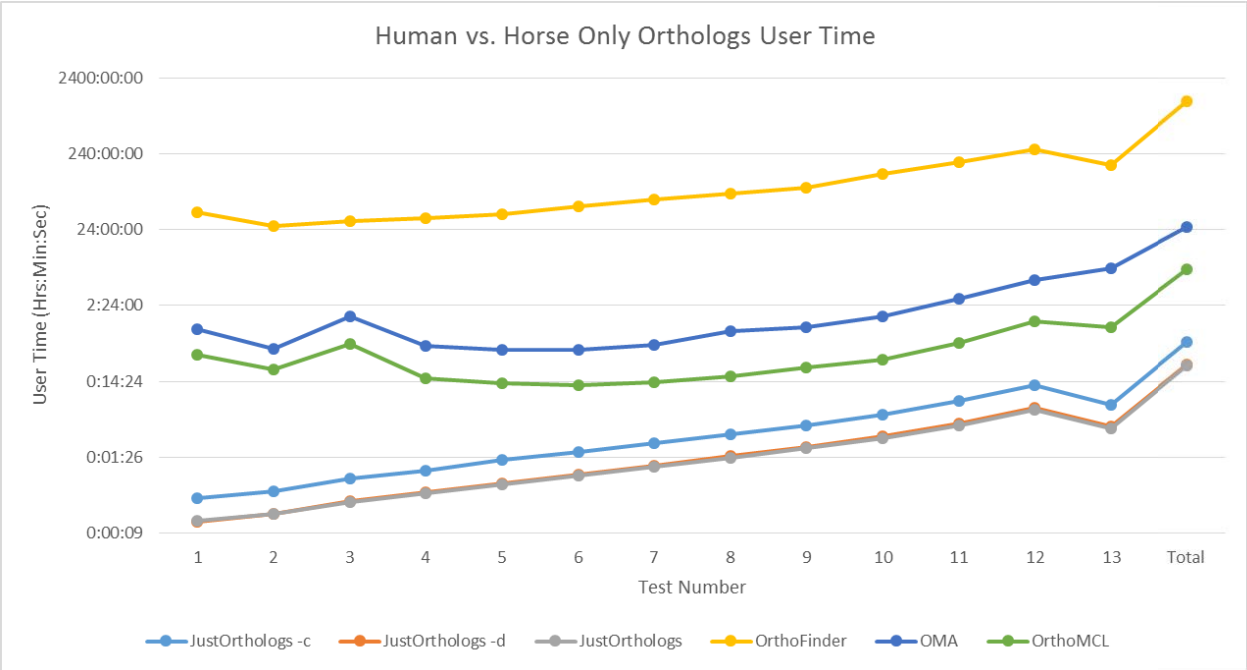
Supplementary Figure 19, Chapter 7. Recall for humans versus horse for each test case where only orthologs are present



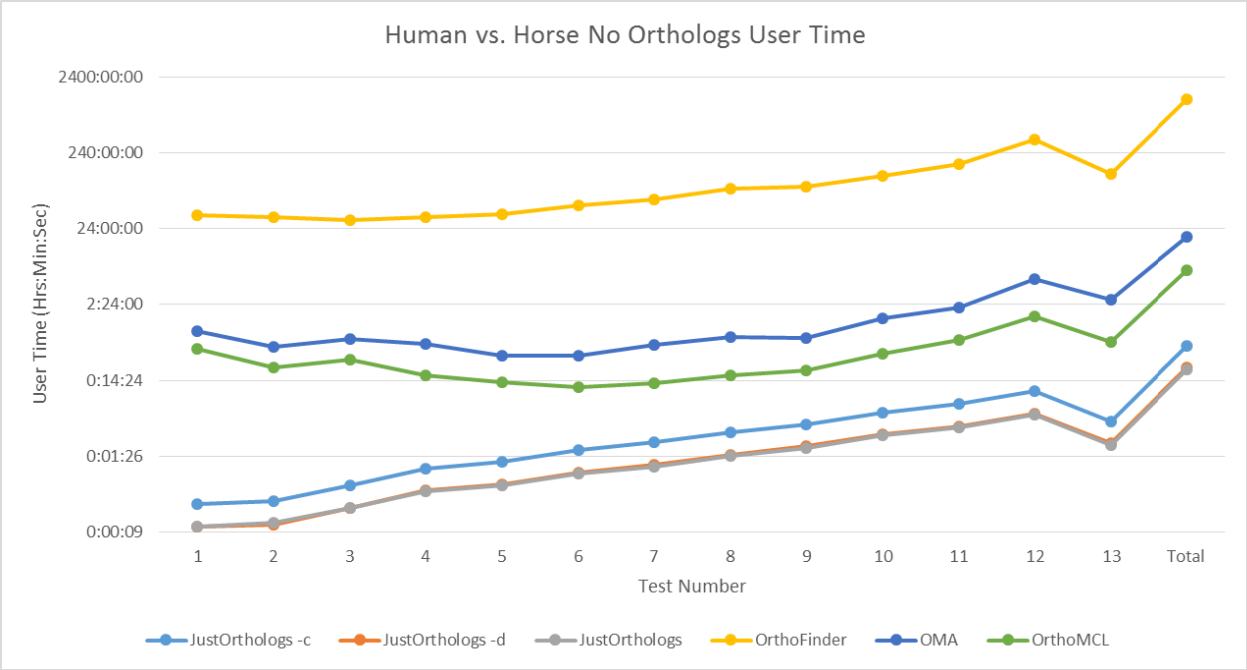
Supplementary Figure 20, Chapter 7. Recall for humans versus horse for each test case where only orthologs are present



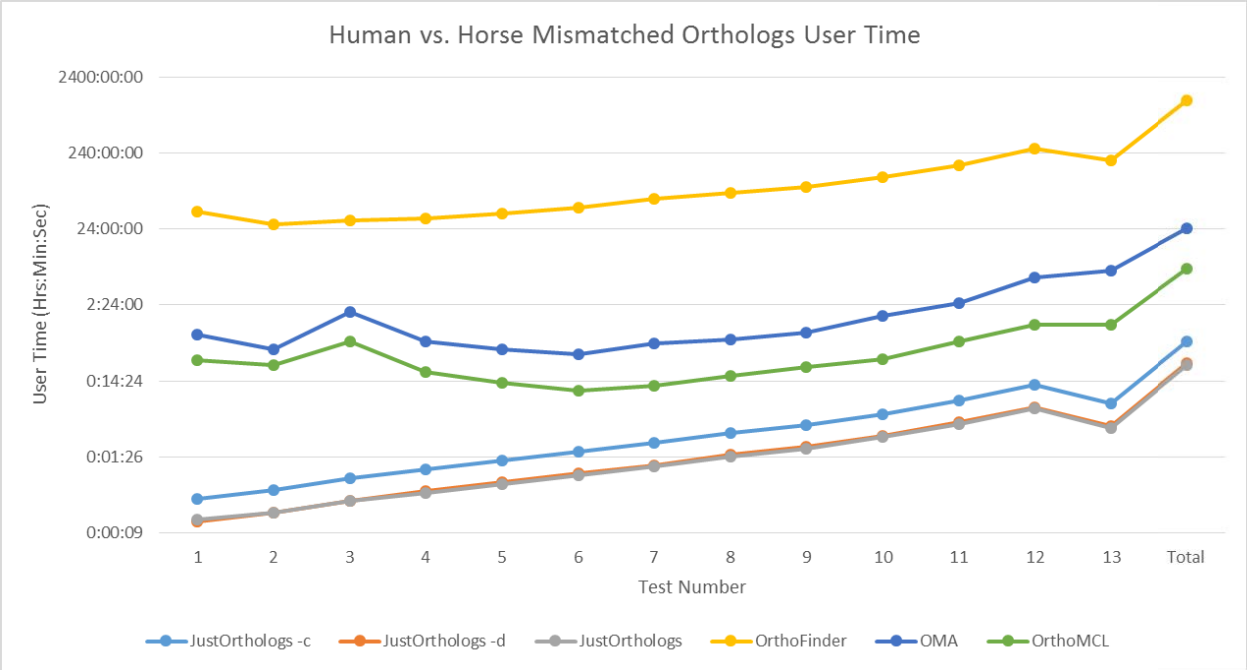
Supplementary Figure 21, Chapter 7. Precision for humans versus horse for each test case where only orthologs are present



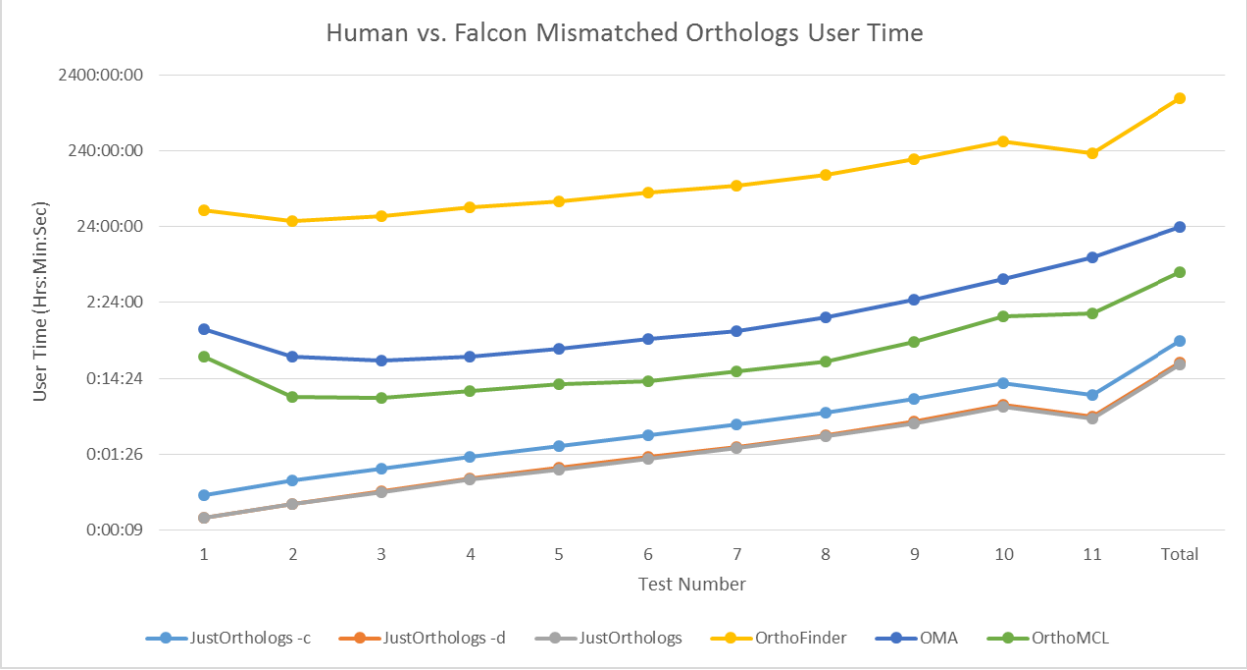
Supplementary Figure 22, Chapter 7. User time for humans versus horse for each test case where only orthologs are present



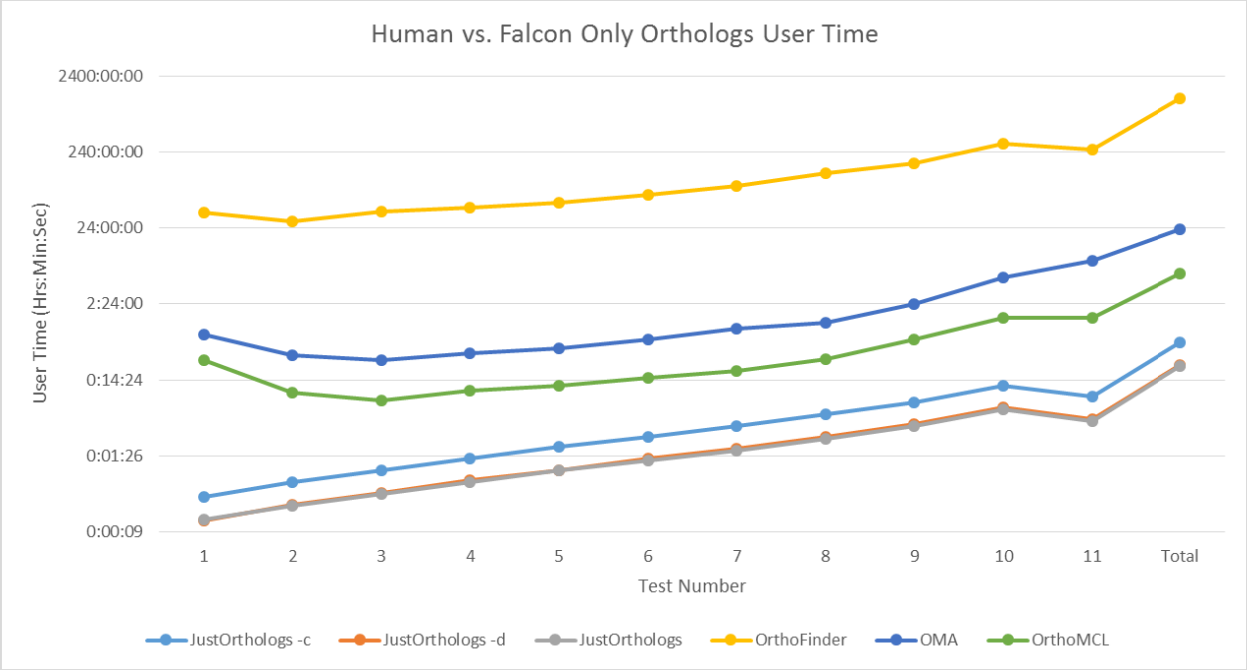
Supplementary Figure 23, Chapter 7. User time for humans versus horse for each test case where no orthologs are present



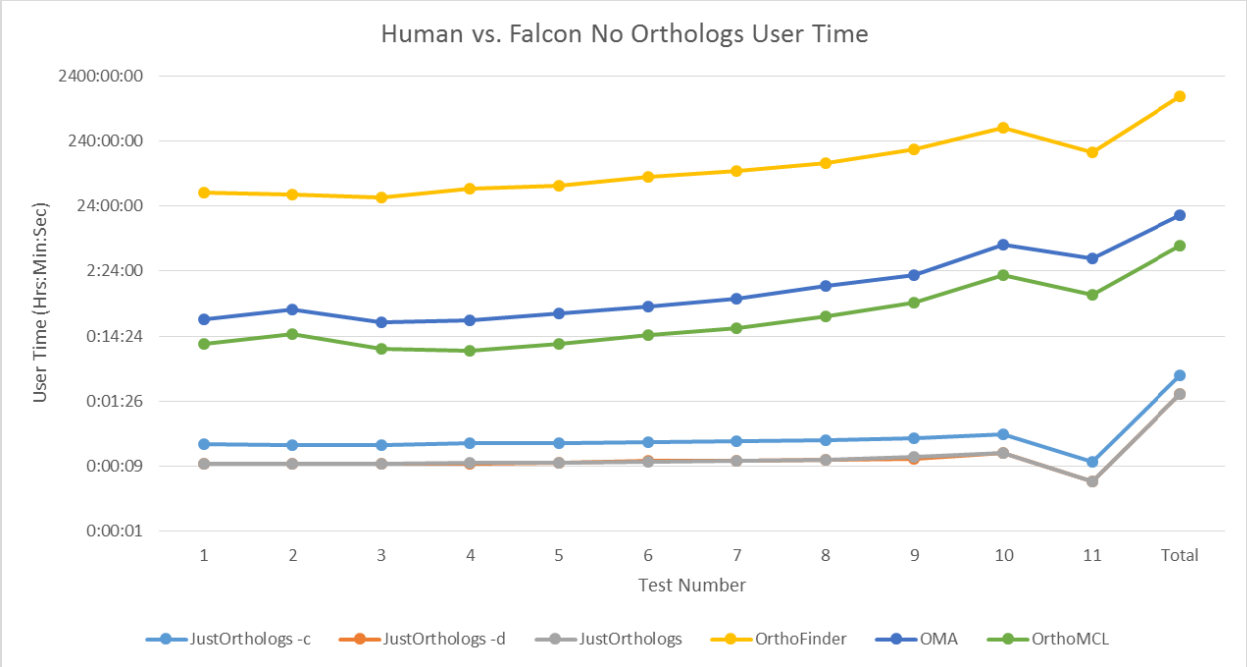
Supplementary Figure 24, Chapter 7. User time for humans versus horse for each test case where some orthologs are present



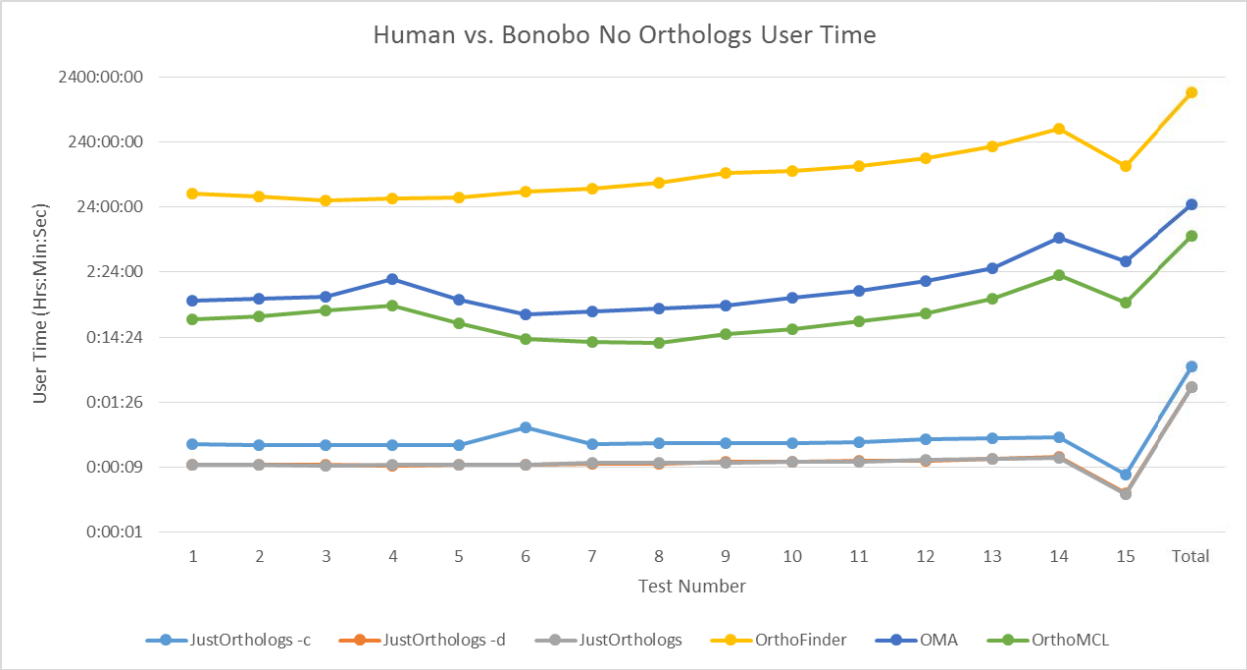
Supplementary Figure 25, Chapter 7. User time for humans versus falcon for each test case where some orthologs are present



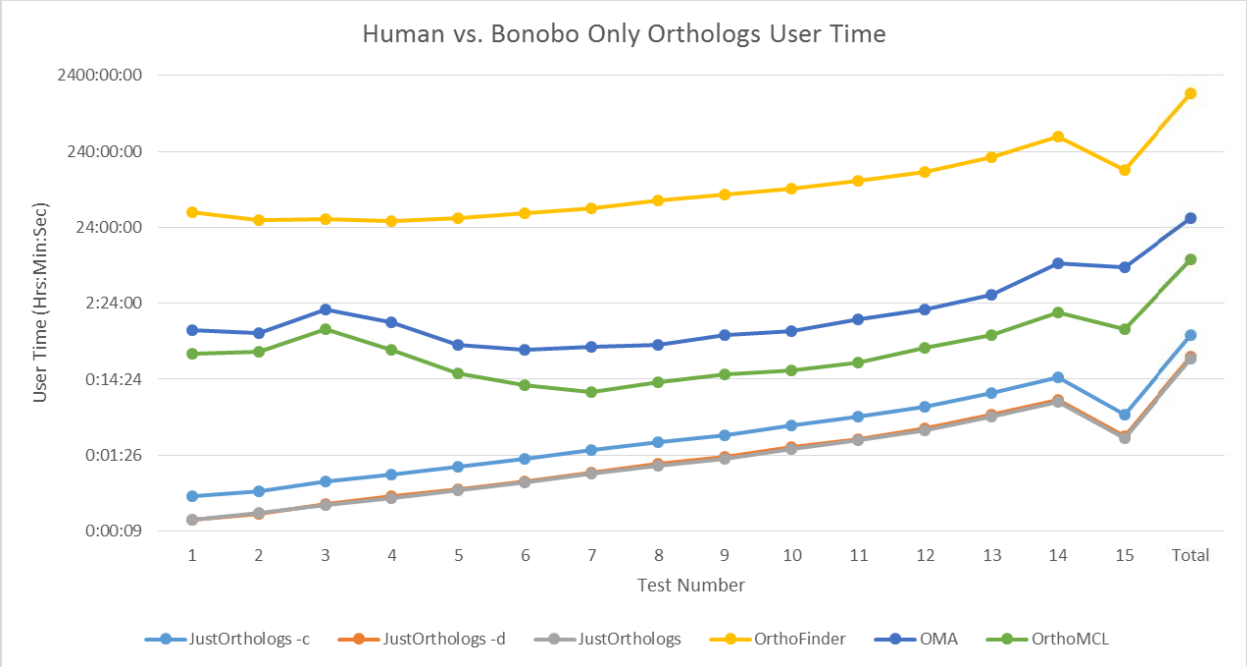
Supplementary Figure 26, Chapter 7. User time for humans versus falcon for each test case where only orthologs are present



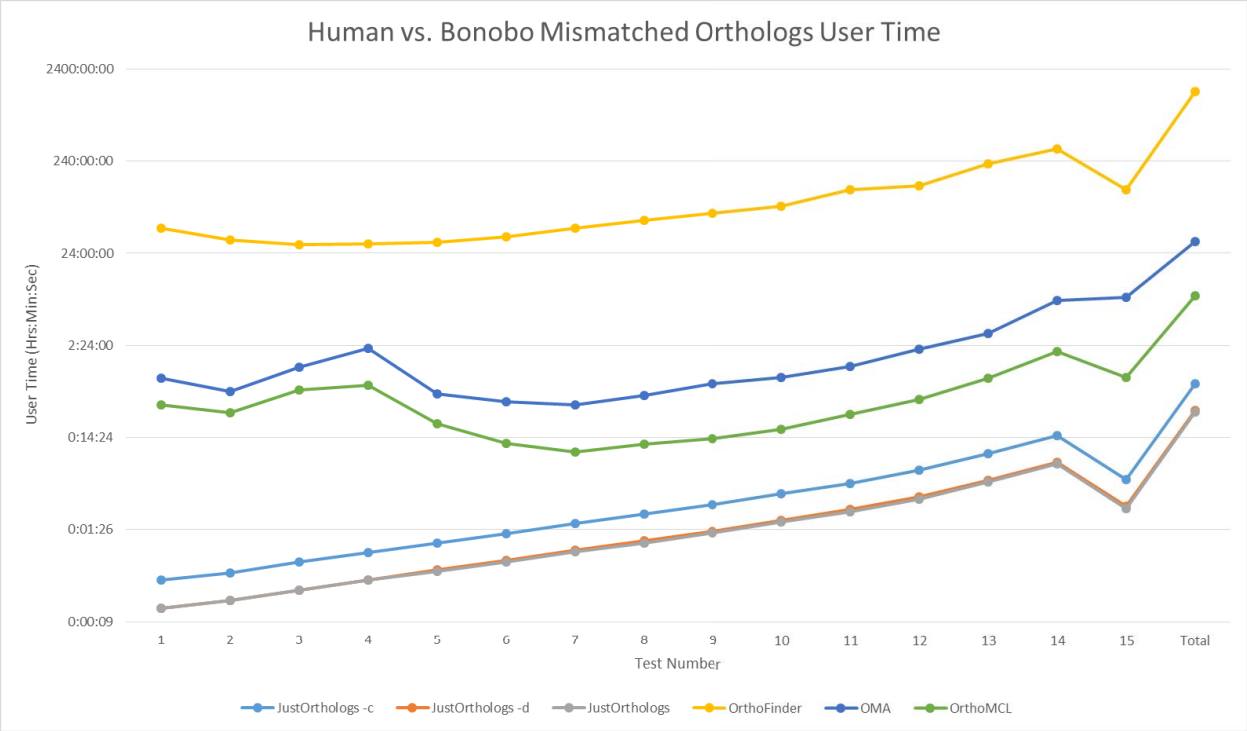
Supplementary Figure 27, Chapter 7. User time for humans versus falcon for each test case where no orthologs are present



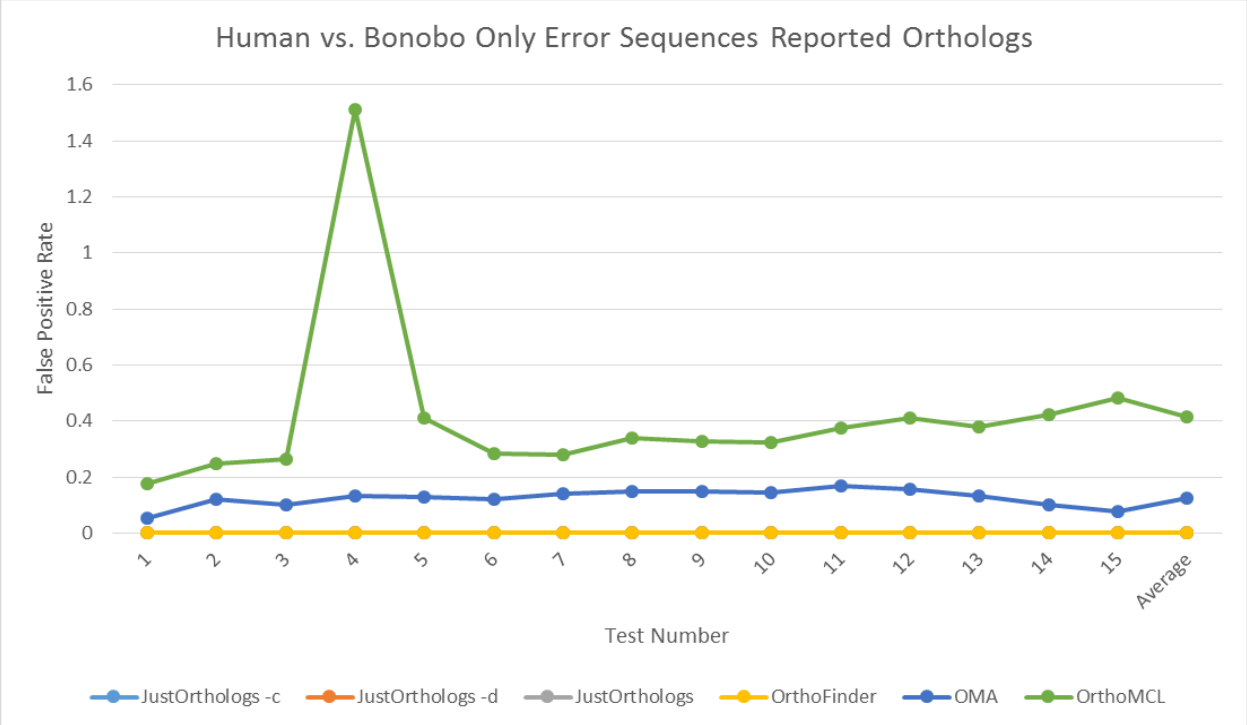
Supplementary Figure 28, Chapter 7. User time for humans versus bonobo for each test case where no orthologs are present



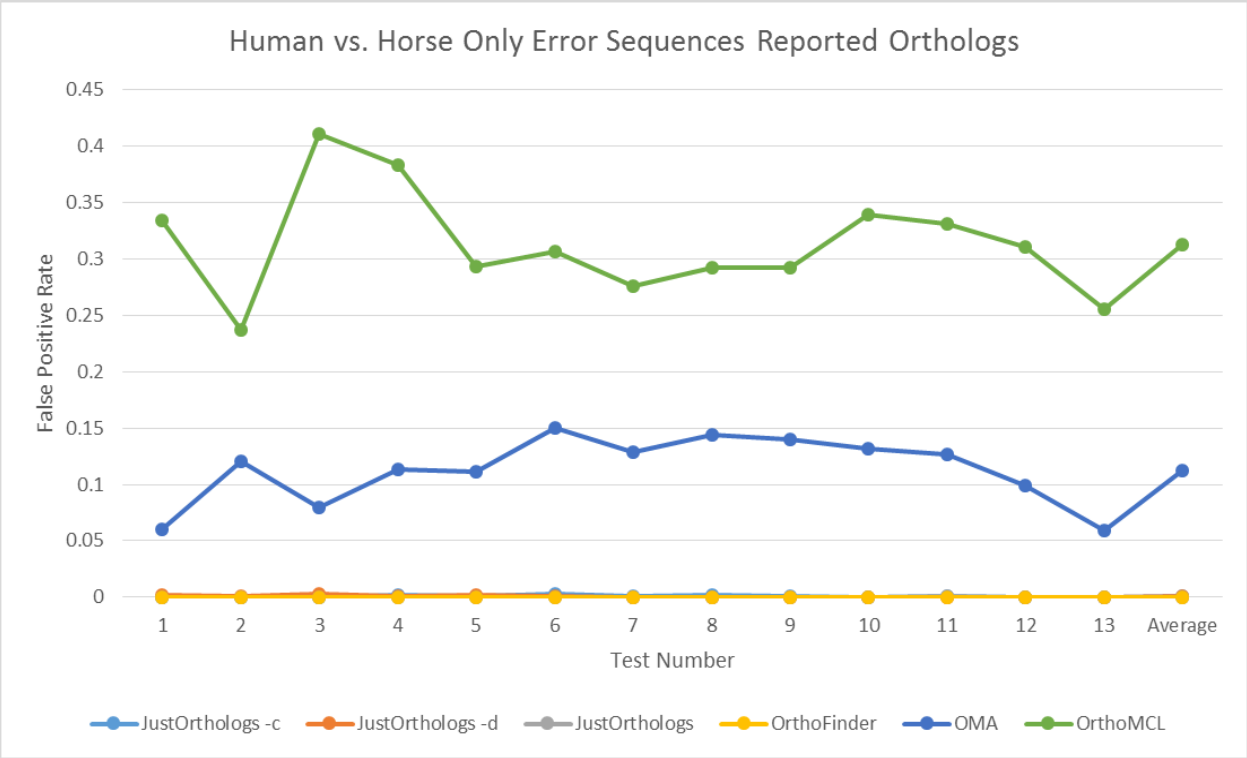
Supplementary Figure 29, Chapter 7. User time for humans versus bonobo for each test case where only orthologs are present



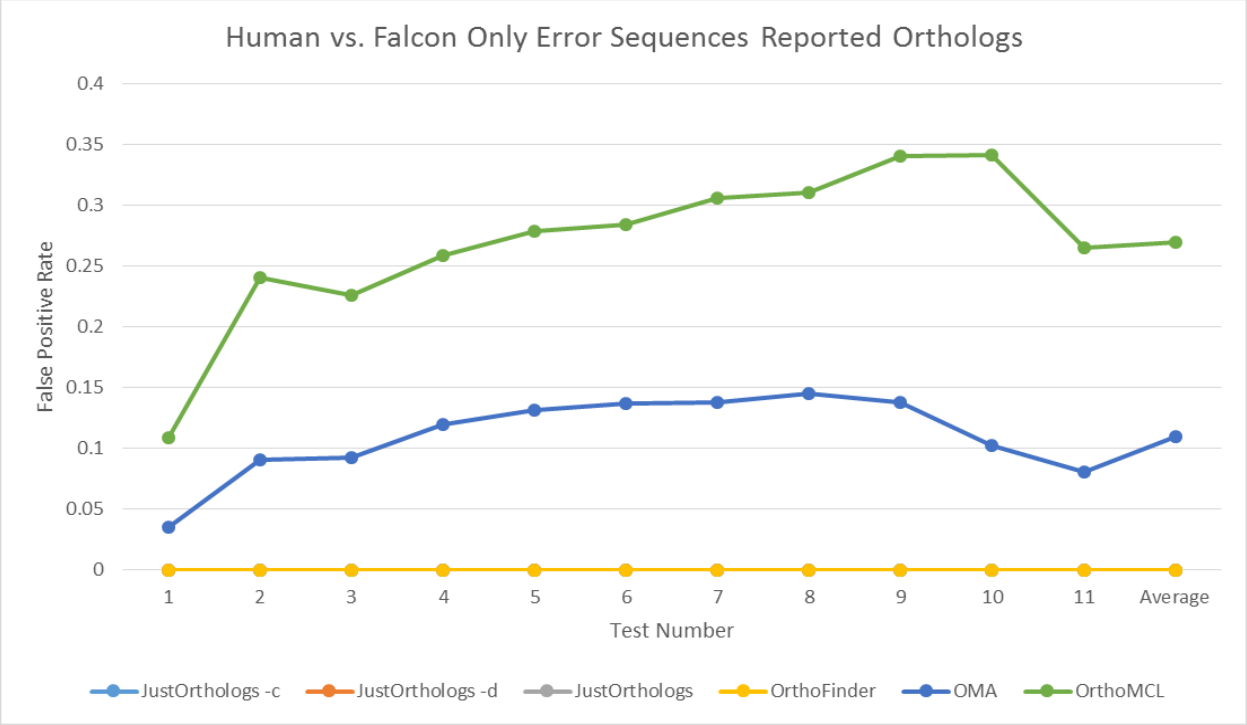
Supplementary Figure 30, Chapter 7. User time for humans versus bonobo for each test case where some orthologs are present



Supplementary Figure 31, Chapter 7. False positive orthologs reported for humans versus bonobo for each test case where no orthologs are present



Supplementary Figure 32, Chapter 7. False positive orthologs reported for humans versus horse for each test case where no orthologs are present



Supplementary Figure 33, Chapter 7. False positive orthologs reported for humans versus falcon for each test case where no orthologs are present