ABSTRACT

WISOTSKY, SARAH ROSE. Explorations of Site-to-Site Synonymous Substitution Rate Variability. (Under the direction of Spencer Muse).

Models of sequence substitution are used to accurately reflect the process of sequence evolution and the underlying biological mechanisms that drive it. Because of advancements in both computational power and our biological understanding of evolutionary processes, new models that better reflect the truth have been proposed. My work focuses on modeling synonymous substitution rate variation across sites, or SRV. There is a growing body of literature that shows that synonymous substitutions can impact downstream products and experience selective pressures. However, most current models of sequence substitution assume a constant synonymous substitution rate across sites. Here I present my findings that show site to site synonymous rate variation is of an appreciable magnitude, is widespread throughout many genes and orders and that by not accounting for this phenomena in our molecular inferences mis-estimation of parameters is likely. First I introduce the basic concepts surrounding models of sequence evolution and rate variation. In my second chapter I analyze Metazoan mitochondrial DNA alignments to show that SRV is both widespread and of a comparable magnitude to that of the nonsynonymous rate variation. As an offshoot of this first project I also revived an unpublished result from Dr. Frank Mannino's thesis that shows the number of rate categories used to estimate a discrete distribution results in an upper bound on the estimated variance as is discussed in Chapter 3. Finally, in collaboration with Dr. Sergei Kosakovsky Pond's group at Temple University, we introduce a new method of gene-wide episodic selection detection that incorporates site to site variability of the synonymous substitution rate, BUSTED+SRV. We compare BUSTED+SRV to the previous method BUSTED using both empirical data from the Selectome database and data simulated using the BUSTED+SRV framework and find that BUSTED has a high false positive rate when SRV is present. The combination evidence of how widespread a phenomena SRV is as well as its magnitude and the high false positive rate of BUSTED suggests that other statistical inference assuming a constant synonymous rate from site to site could

be similarly impacted.

Explorations of Site-to-Site Synonymous Substitution Rate Variability

by
Sarah Rose Wisotsky

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2018

APPROVED BY:

David Reif	Jung-Ying Tzeng
Ignazio Carbone	Spencer Muse Chair of Advisory Committee

DEDICATION

To my parents, for always encouraging and supporting me.

BIOGRAPHY

Sarah (Sadie) Rose Wisotsky was born in Honolulu, HI and raised in Birmingham, AL. She received her Bachelors of Science from Auburn University in 2013 in Cellular, Molecular, and Microbial Biology. There she worked in a Plant Pathology Lab under the guidance of Dr. Leonardo De La Fuente and lab manager, Jennifer Parker, studying Citrus Greening Disease. It was there that she learned about Bioinformatics, a field that combines her love of biology with her passion for data analysis. After starting the Bioinformatics Program at NC State University in Raleigh, NC, she found an interest in Molecular Evolution and started her dissertation work with Dr. Spencer Muse. Sadie plans on continuing her work as a post-doc in the Kosakovsky Pond group at Temple University in Philadelphia, PA.

ACKNOWLEDGEMENTS

To start I'd like to thank my committee, Dr. Spencer Muse, Dr. David Reif, Dr. Jung-Ying Tzeng, and Dr. Ignazio Carbone and my graduate school representative Lisa McGraw for their assistance throughout this process. In particular, I would like to thank my advisor, Dr. Spencer Muse for his guidance throughout my time here at NC State University. I feel incredibly lucky to have found a mentor and a friend that I enjoy working with and has taught me so much and that I also enjoy talking to about everything from board games to musicals.

I would also like to thank Dr. Sergei Kosakovsky Pond for answering countless questions about HyPhy and code, being an excellent collaborator and offering me an opportunity to pursue a postdoc.

To everyone in the BRC. Thank you all for being understanding when I was first learning to use the cluster and repeatedly took it over entirely. Thank you all for the amazing food and conversations at every potluck.

To my partner and girlfriend, Bri Oleson, I would not have made it through this with as much of my sanity intact if not for you. You've been amazing and helped me with everything from reminding me to eat something to sitting through countless practice presentations and reading through countless drafts.

To my friends. To those of you in my program that have made goofing around in the several offices we've had in our time here memorable: Kim, Will, Kyle, Michele, and Maria. To those of you I met in classes who became my support there and outside: Jamie, Natalie, Mary, Marcella, Emma, and Richard. To those of you on our occasional trivia team, 'Zelda is the Boy': William, Racheal, Katie, and Molly. To Frankie, my Wynonna Earp friend from afar. Thank you all for all of your support and good times. I'm sure I'm missing some folks but just know if you've been my friend here, you've had a hand in making this happen.

To my parents, thank you so much for everything you do and have done. Even when you couldn't physically be here you've always supported and I couldn't be luckier. To my mom, Rebecca Wisotsky, you've always inspired me with your intelligence and your desire to always keep learning about

yourself and about the world around you. Sometimes it's about Judaic texts and other times it's a new conspiracy theory but you've always got something new to talk about when you call. To my dad, Joel Wisotsky, you've taught me not to settle and speak up if I think something is wrong. You've always encouraged me to follow my passion and that's led me here.

TABLE OF CONTENTS

LIST OF TABLES . . . . . **ix**

LIST OF FIGURES . . . . . **x**

Chapter 1 INTRODUCTION **1**
1.1 Components of Substitution Models 1
1.2 Examples of Substitution Models 4
1.2.1 Nucleotide Models 4
1.2.2 Amino Acid Models 9
1.2.3 Codon Models 10
1.3 Likelihood 14
1.3.1 Maximum Likelihood Estimation of D 16
1.4 Counting Methods 17
1.4.1 Codon Models and Counting Methods 18
1.5 Rate Variation 20
1.5.1 Nonsynonymous Rate Variation 22
1.5.2 Synonymous Rate Variation 23
1.6 Tests of Selection 25
1.6.1 Site Models 26
1.6.2 Branch Models 28
1.6.3 Branch-Site Models 29
1.6.4 Effects of Synonymous Rate Variation 33
1.7 Description of Data Sets Used in this Thesis 36
1.7.1 General Composition 36
1.7.2 Comparison of Analysis on the Data Sets 37
1.8 Overview 38
1.9 Figures 38

Chapter 2 Widespread Site-to-Site Synonymous Substitution Rate Variation **53**
2.1 Authors 53
2.2 Contribution 54
2.3 Abstract 54
2.4 Introduction 54
2.5 Results and Discussion 56
2.5.1 Prevalence of Synonymous Rate Variation 57
2.5.2 Magnitude of Synonymous Rate Variation 57
2.5.3 Synonymous Rate Variation is Widespread 57
2.6 Implications 58
2.7 Materials and Methods 59
2.7.1 Sequences 59
2.7.2 Alignments 60
2.7.3 Trees 60

2.7.4 Statistical Methods . . . 60
2.8 Figures . . . 61

Chapter 3 Unexpected Consequences Stemming from the Selection of the Number of Discrete Rate Categories . . . 66
3.1 Authors . . . 66
3.2 Contribution . . . 67
3.3 Abstract . . . 67
3.4 Introduction . . . 67
3.5 Materials and Methods . . . 68
3.5.1 Data . . . 68
3.5.2 Analysis . . . 68
3.6 Results and Discussion . . . 69
3.6.1 Upper Bound Derivation . . . 69
3.6.2 Bias of Estimated Discrete CV . . . 69
3.6.3 Bias of Shape Parameter . . . 70
3.7 Implications . . . 70
3.8 Acknowledgments . . . 71
3.9 Supplementary Materials . . . 72
3.9.1 Shifting Distribution of Estimates . . . 72
3.10 Brief Communication Text . . . 72
3.11 Figures . . . 77

Chapter 4 Accounting for Site-to-Site Synonymous Rate Variability Reveals High False Positive Rate in Test of Selection . . . 84
4.1 Authors . . . 84
4.2 Contribution . . . 85
4.3 Abstract . . . 85
4.4 Introduction . . . 86
4.5 Materials and Methods . . . 87
4.5.1 Empirical Data . . . 87
4.5.2 Model . . . 88
4.5.3 Simulation . . . 89
4.5.4 Implementation . . . 89
4.6 Results . . . 89
4.6.1 Performance of BUSTED+SRV . . . 89
4.6.2 Accounting for SRV Improves Model Fit . . . 90
4.6.3 Selectome Analysis Reveals High False Positive Rate . . . 90
4.6.4 Simulation Finds High False Positives . . . 91
4.7 Discussion . . . 93
4.8 Acknowledgments . . . 95
4.9 Figures and Tables . . . 95

BIBLIOGRAPHY . . . 104

APPENDICES **106**
Appendix A NCBI Accession Numbers 107
A.1 Description 107
A.2 Table 107
Appendix B Simulation Parameters 139
B.1 Description 139
B.2 Table 139

LIST OF TABLES

Table 1.1 **Branch-Site Rate Categories.** A table describing the combination of foreground and background branch-site categories and their proportions. 41

Table 4.1 **Two-way Table of Selectome Positive Selection.** (UNCORRECTED $P \leq 0.05$) Fraction of data sets under selection. The fraction of total data sets categorized by if there is evidence of selection according to BUSTED and according to BUSTED+SRV. 95

Table A.1 **NCBI Accession numbers.** Table listing the NCBI accession number for the mitochondrial data sets as well as the order and species used. 107

Table B.1 **Simulation Parameters.** Table of simulation parameters for the simulations discussed in 4. Not every combination of these parameters were run but they are listed here to give a general idea of the breadth of the simulation. 140

LIST OF FIGURES

Figure 1.1 **General Time Reversible Model.** Four state, nucleotide, continuous time reversible Markov Chain. Arrows represent changes from one state to the next. The variable r_i on each arrow represents the rate of substitution from one state to the next. 39

Figure 1.2 **Simple, Rooted Tree.** A simple rooted tree with the root at node 0 of four sequences. 39

Figure 1.3 **Unrooted Tree.** Unrooted tree with 4 descendant sequences. 39

Figure 1.4 **Example of Discrete Distribution.** Gamma distribution with $\lambda_\alpha = 2$ and $k = 4$ represented by the red line. The histogram represents the 4 rate categories. Dashed lines represent the mean of each category, z_k. 40

Figure 1.5 **Varying Shape Parameters of the Gamma Distribution.** Lines represent the continuous gamma distribution with $\lambda_\alpha = 1$, $\lambda_\alpha = 2$, $\lambda_\alpha = 10$ where the scale parameter equals the shape parameter. 40

Figure 1.6 **Comparison of Synonymous and Nonsynonymous Coefficients of Variation (CV).** For each of the 721 datasets, we plot its estimated synonymous and nonsynonymous CV. Points below the blue line are datasets where the synonymous CV exceeds that of nonsynonymous CV; points below the red line had a synonymous CV of at least half the nonsynonymous CV. The vertical line of points on the left represents a numerical artifact for datasets with synonymous CV effectively zero. 41

Figure 1.7 **Tree of Metazoan Order for the Mitochondrial Data Set.** This tree of the 56 orders comprised into the mitochondrial data set was generated using NCBI's taxonomy common tree tool. 42

Figure 1.8 **Range of Sites and Sequences for Data Sets.** Histograms describing the number of sequences (a) and sites (b) per alignment. Histograms are split for the Mitochodrial (red) and Selectome (blue) data sets. 42

Figure 1.9 **Range of Estimated ω_3 of Data Sets.** Histograms describing the estimated ω_3 for the Mitochodrial (red) and Selectome (blue) data sets according to BUSTED (a) and BUSTED+SRV (b). 43

Figure 1.10 **Histogram of the Estimated Synonymous Coefficient of Variation (CV).** The distribution of estimated synonymous CVs according to BUSTED+SRV for the Selectome (blue) and mitochondrial (red) data sets. 43

Figure 2.1 **Results from the Likelihood Ratio Test for the Presence of Synonymous Rate Variability within Genes.** Datasets for each gene × order combination are found to be significant (at 0.01 or 0.05) or nonsignificant after a Bonferroni correction. The null hypothesis of no synonymous rate variation (SRV) from site to site is rejected for 57% of the combinations. Orders are arranged in phylogenetic relation to each other on the x-axis. NAs represent gene × order combinations that were unavailable for analysis. 62

Figure 2.2 **Comparison of Synonymous and Nonsynonymous Coefficients of Variation (CV).** For each of the 721 datasets, we plot its estimated synonymous and nonsynonymous CV. Points below the blue line are datasets where the synonymous CV exceeds that of nonsynonymous rates; points below the red line had a synonymous CV of at least half the nonsynonymous CV. The vertical line of points on the left represents a numerical artifact for datasets with synonymous CV effectively zero. 63

Figure 2.3 **Boxplots of the Ranges of Synonymous (A) and Nonsynonymous (B) Coefficients of Variation (CV) for each Gene Group.** Boxplots in A indicate the range of synonymous rate variation (SRV) across the metazoan orders for each gene. Boxplots in B represent the range of nonsynonymous rate variation across the metazoan orders for each gene. While the range of synonymous CVs is generally lower than that of the nonsynonymous CVs they are of the same order of magnitude. 64

Figure 2.4 **Boxplots of the Ranges of Synonymous (A) and Nonsynonymous (B) Coefficients of Variation (CV) for each Metazoan Order.** The boxplots in A and B represent the synonymous and nonsynonymous CV ranges respectively across the mitochondrial genes for each Metazoan order. The x-axis of orders is arranged in phylogenetic relation. 65

Figure 3.1 **Variation Bound Plots.** For each of 13 mtDNA genes for 56 metazoan orders, the estimated coefficient of variation (CV) for nonsynonymous rates is plotted against that for the synonymous rates using an increasing number of rate categories, K= 3, 4, 5, 7, 10. The horizontal line in each sub-plot is the maximum possible CV estimate according to the equation $\sqrt{K-1}$. Note that as the number of rate categories increase fewer points reach the theoretical upper bound. The upper bound does not appear to be consequential for the synonymous substitution rate estimates in this study. 77

Figure 3.2 **Estimated Nonsynonymous CV for Varying Rate Categories.** For all the mitochondrial data sets the estimated nonsynonymous coefficient of variation (CV) using 10 rate categories plotted against the estimated nonsynonymous CV using 3 categories (A), 5 rate categories (B), and 7 rate categories (D). Additionally, the nonsynonymous CV estimated with 3 rate categories was plotted against the nonsynonymous CV using 5 rate categories (C). The red line represents the x = y line. 78

Figure 3.3 **Estimated Synonymous CV for Varying Rate Categories.** For all the mitochondrial data sets the estimated synonymous coefficient of variation (CV) using 10 rate categories plotted against the estimated synonymous CV using 3 categories (A), 5 rate categories (B), and 7 rate categories (D). Additionally, the synonymous CV estimated with 3 rate categories was plotted against the synonymous CV using 5 rate categories (C). The red line represents the x = y line. 79

Figure 3.4 **Estimated Nonsynonymous Shape Parameter for Varying Rate Categories.** For all the mitochondrial data sets the estimated nonsynonymous shape parameter using 10 rate categories plotted against the estimated nonsynonymous shape parameter using 3 categories (A), 5 rate categories (B), and 7 rate categories (D). Additionally, the nonsynonymous shape parameter estimated with 3 rate categories was plotted against the nonsynonymous shape parameter using 5 rate categories (C). The red line represents the x = y line. . 80

Figure 3.5 **Estimated Synonymous Shape Parameter for Varying Rate Categories.** For all the mitochondrial data sets the estimated synonymous shape parameter using 10 rate categories plotted against the estimated synonymous shape parameter using 3 categories (A), 5 rate categories (B), and 7 rate categories (D). Additionally, the synonymous shape parameter estimated with 3 rate categories was plotted against the synonymous shape parameter using 5 rate categories (C). The red line represent the x = y line. 81

Figure 3.6 **Violin Plots.** Violin plots of the nonsynonymous CV (A) and nonsynonymous shape parameter (C) versus the number of nonsynonymous rate categories as well as the synonymous CV (B) and synonymous shape parameter (D) versus the number of synonymous rate categories. The height of the violin plots represents the range of the parameter over the data sets. The width of plots represents the density of data sets across each parameter. The horizontal lines split the violin plots into quintiles representing 10% of the data sets each. The red dot represents the median of each. 82

Figure 4.1 **Histograms of the P-Values According to BUSTED and BUSTED+SRV.** The p-values as calculated by BUSTED (red) and BUSTED+SRV (blue) are plotted with a range of $p = 0$ to $p = 1$ (A) and a range of $p = 0$ to $p = 0.15$ (B). The y-axis for both plots represent the number of data sets that fall within the 0.05 range of each bin. The data sets represented here are simulated with a CV of SRV = 0 and an $\omega_3 = 1$ but have a varying number of sites and sequences. . . . 96

Figure 4.2 **Fraction of Alignments Under Selection for the Selectome Data Set.** Fraction of alignments under selection ($P \leq 5.6e^{-6}$) versus the median of a sliding window for A) Sequences B) Coefficient of variation of synonymous substitution rates C) Number of Codons D) the ω_3 maximum likelihood estimate according to BUSTED. Lines are Loess fit lines for for BUSTED p-values (red) and BUSTED+SRV p-values (blue). 97

Figure 4.3 **Difference of AICc.** Histogram of the difference between BUSTED's AICc and BUSTED+SRV's AICc for each data set. 98

Figure 4.4 **Power Curves and P-Value Boxplots for Simulation of a Tree with 33 Sequences and with 5000 Codons Faceted by the True ω_3 Value.** The caluculated p-values for each simulation of 100 replicates is represented by the boxplots as the simulated CV value increases. The blue boxplots are the caluculated p-values according to BUSTED+SRV and the red are the caluculated p-values according to BUSTED. The power of each test for each value of synonymous CV is also plotted here for both BUSTED (Red line) and BUSTED+SRV (blue line). The graphs are split by the simulated ω_3. 99

Figure 4.5 **Power Curves and P-Value Boxplots for Simulation with $\omega_3 = 1$ for Trees with 16 and 31 Sequences and 100, 300, 500, 1000, 5000 Codons Respectively.** The caluculated p-values for each simulation of 100 replicates is represented by the boxplots as the simulated CV value increases. The blue boxplots are the estimates according to BUSTED+SRV and the red are according to BUSTED. The power of each test for each value of synonymous CV is also plotted here for both BUSTED (Red line) and BUSTED+SRV (blue line). The graphs are split horizontally by the number of sequences and vertically by the number of sites. 100

Figure 4.6 **Power Curves and P-Value Boxplots for Simulation with $\omega_3 = 2.077$ for Trees with 16 and 31 Sequences and 100, 300, 500, 1000, 5000 Codons Respectively.** The caluculated p-values for each simulation of 100 replicates is represented by the boxplots as the simulated CV value increases. The blue boxplots are the estimates according to BUSTED+SRV and the red are according to BUSTED. The power of each test for each value of synonymous CV is also plotted here for both BUSTED (Red line) and BUSTED+SRV (blue line). The graphs are split horizontally by the number of sequences and vertically by the number of sites. 101

Figure 4.7 **Power Curves and P-Value Boxplots for Simulation with $\omega_3 = 6$ for Trees with 16 and 31 Sequences and 100, 300, 500, 1000, 5000 Codons Respectively.** The caluculated p-values for each simulation of 100 replicates is represented by the boxplots as the simulated CV value increases. The blue boxplots are the estimates according to BUSTED+SRV and the red are according to BUSTED. The power of each test for each value of synonymous CV is also plotted here for both BUSTED (Red line) and BUSTED+SRV (blue line). The graphs are split horizontally by the number of sequences and vertically by the number of sites. 102

Figure 4.8 **Histogram Showing the Range of the Estimated Synonymous CV for the Selectome Data Set.** Note that the true maximum synonymous CV for the Selectome data sets was 21.29. 103

CHAPTER

1

INTRODUCTION

In this introduction, I outline the key discoveries, methods, and models that have led up to my thesis research. Much of molecular biology and evolution is incremental. By building on and refining past discoveries to advance our understanding of the process of sequence evolution. Here we look at what components comprise a sequence substitution model. This includes how models can incorporate site-to-site rate variation and why that is biologically relevant. We then discuss the three most common types of models and the extensions within those categories. Additionally, we discuss tests of selection and how the models are used for them.

1.1 Components of Substitution Models

In this section, I will lay out the framework for the most common sequence substitution models. While each type of model has unique characteristics with its own strengths and weaknesses, many

use the same framework to describe sequence substitution. Using the General Time Reversible (GTR) [Tavaré 1986] as an example I will describe the most common components of sequence substitution models.

Finite state, continuous time Markov chains are typically used to model the evolution of DNA and protein sequences. First-order Markov chains treat sites as independent random variables that change stochastically from one state to the next. A key feature of first-order Markov chains is that the probability of changing from one state to the next is only dependent on the current state and none of the previous states of a site. For model sequence evolution, the sites are either individual nucleotides, codons, or amino acids. The GTR is a nucleotide model with 4 possible states, A, C, G, and T, shown graphically in figure 1.1.

One of the main components of a sequence substitution model is the instantaneous rate matrix Q. The instantaneous rate matrix Q for the GTR is given as:

$$\mathbf{Q}_{GTR} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -\sum & \pi_C r_1 & \pi_G r_2 & \pi_T r_3 \\ \pi_A r_1 & -\sum & \pi_G r_4 & \pi_T r_5 \\ \pi_A r_2 & \pi_C r_3 & -\sum & \pi_T r_6 \\ \pi_A r_4 & \pi_C r_5 & \pi_G r_6 & -\sum \end{pmatrix} \end{array}$$

where each q_{ij} is the instantaneous rate of change from state i to state j. For substitution models, q_{ij} typically includes a substitution rate parameter which we will refer to as r as well as a parameter for equilibrium frequencies π. Since each change is assumed to occur at different rates, there are 6 individual substitution rate parameters, $r_1, r_2, r_3, r_4, r_5, r_6$. For the GTR model, it is assumed that the frequencies of each nucleotide are not equal, $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$. These parameters may be specified based on prior knowledge or estimated from real data. In general for the equilibrium frequencies, $\sum_{i=1}^{n} \pi_i = 1$, where n is the number of possible states. Frequency parameters can reflect the composition of a given sequence, such as higher GC content. The diagonals of Q are defined

such that each row sums to zero, thus: $q_{ii} = -\sum_{i \neq j} q_{ij}$. Also of note for most of the models we will discuss is that the instantaneous rate matrix only allows for a single change in state to occur.

The transition probability of a site s changing from state i to state j in time t is denoted $p_{ij}(t)$. The matrix of all the possible transition probabilities is known as the transition probability matrix and is represented as $P(t)$. The transition probability matrix for a 4 state nucleotide model is:

$$P(t) = \begin{bmatrix} p_{AA}(t) & p_{AC}(t) & p_{AG}(t) & p_{AT}(t) \\ p_{CA}(t) & p_{CC}(t) & p_{CG}(t) & p_{CT}(t) \\ p_{GA}(t) & p_{GC}(t) & p_{GG}(t) & p_{GT}(t) \\ p_{TA}(t) & p_{TC}(t) & p_{TG}(t) & p_{TT}(t) \end{bmatrix}.$$

The relationship between the transition probability matrix $P(t)$ and the instantaneous rate matrix Q is:

$$\frac{dP(t)}{dt} = QP(t),$$

which can be solved to give us the relationship

$$P(t) = e^{Qt}.$$

For most models, no closed-form expression of this can be found. But for some simple, special cases we can find one. These components form the basic building blocks of most substitution models. Finite state, continuous time Markov chains provide a simple and convenient way to model sequence evolution at the molecular level. In addition to the elements discussed here, substitution models can also incorporate other biologically relevant parameters. These individual parameters will be discussed as they are introduced by models.

1.2 Examples of Substitution Models

Sequence substitution models can be based on three main types of data: nucleotides, amino acids or codons. Nucleotide models were the first to be proposed. They allow us to examine how sequences evolve at the base level but lack information of how single nucleotide substitutions impact the sequence downstream. Amino acid models provide information about possible changes to coding regions and potential information on how proteins are impacted by nonsynonymous changes. However, they lack the ability to account for synonymous substitutions, as those changes are not reflected at the protein level. Here, codon models become useful, distinguishing between synonymous and nonsynonymous changes in the sequence. Additionally, while nucleotide models provide a simple way of modeling sequence evolution, they do not reflect the selective constraints placed on nucleotides in different positions of the codon while codon models do. In the following section, I will outline commonly used nucleotide, amino acid, and codon models.

1.2.1 Nucleotide Models

This section will cover the basics of several nucleotide-based models of substitution. Nucleotide models, as the name suggests, model the process of substitution at the nucleotide level. They are often used to estimate the evolutionary distance between sequences as well as the substitution rate. Over the course of several decades, nucleotide models have incorporated information about the underlying biological process of substitution as it became available. Here we present several nucleotide models and highlight how they build upon each other. What is presented here is meant to be an overview highlighting a few important models and not a comprehensive list.

1.2.1.1 Jukes-Cantor 1969

The first model of sequence evolution was published by Jukes & Cantor [1969]. The JC69 model assumes that the probability of changing from any one nucleotide (A,G,C, or T) to any other nucleotide is equally likely and that all base frequencies are equal, or $\pi_T = \pi_C = \pi_A = \pi_G = \frac{1}{4}$. Using

the notation from the GTR, this means $r_1 = r_2 = r_3 = r_4 = r_5 = r_6 = \beta$. This gives us the instantaneous rate matrix:

$$\mathbf{Q}_{\text{JC69}} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -\sum & \frac{1}{4}\beta & \frac{1}{4}\beta & \frac{1}{4}\beta \\ \frac{1}{4}\beta & -\sum & \frac{1}{4}\beta & \frac{1}{4}\beta \\ \frac{1}{4}\beta & \frac{1}{4}\beta & -\sum & \frac{1}{4}\beta \\ \frac{1}{4}\beta & \frac{1}{4}\beta & \frac{1}{4}\beta & -\sum \end{pmatrix} \end{array}$$

where the summation of each row is zero, in this case $\sum = \frac{3\beta}{4}$. The transition probabilities for this model are:

$$p_{ij}(t) = \begin{cases} \frac{1}{4} - \frac{1}{4}e^{-\beta t} & \text{if } i \neq j \\ \frac{1}{4} + \frac{3}{4}e^{-\beta t} & \text{if } i = j \end{cases}.$$

The JC69 model is called a "one parameter model" since it uses only a single rate parameter and is the simplest nucleotide model we will discuss. It provides a computationally quick way to calculate the distance between two sequences or other analyses such as phylogeny estimation. However, because of its simplicity it is often inaccurate as it does not account for underlying biological biases such as transition and transversion substitution rate bias. The models that follow address some of these weaknesses.

1.2.1.2 Kimura 1980

Kimura [1980] chose to use two parameters for the K80 or K2P model: the rate of transitions (α) and the rate of transversions (β). A transition occurs when a nucleotide substitution results in a purine being replaced by a purine or a pyrimidine being replaced by a pyrimidine. Transversions occur when a nucleotide substitution results in a purine being replaced by a pyrimidine or vice versa. Evidence later collected from sequence information shows that transitions occur more frequently than transversions [Brown et al. 1982], so the JC69 assumption of equal rates when modeling is not biologically accurate.

The K80 model builds on the JC69 as all bases are still assumed to occur with equal frequencies ($\pi_T = \pi_C = \pi_A = \pi_G$). The transition probability of changing from one nucleotide to another is no longer the same for all changes. It is dependent on the type of change occurring. Using our established notation we see that $r_1 = r_3 = r_4 = r_6 = \beta$ and $r_2 = r_5 = \alpha$. Here we present the instantaneous rate matrix, Q_{K80} for the model:

$$\mathbf{Q}_{\text{K80}} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -\sum & \frac{1}{4}\beta & \frac{1}{4}\alpha & \frac{1}{4}\beta \\ \frac{1}{4}\beta & -\sum & \frac{1}{4}\beta & \frac{1}{4}\alpha \\ \frac{1}{4}\alpha & \frac{1}{4}\beta & -\sum & \frac{1}{4}\beta \\ \frac{1}{4}\beta & \frac{1}{4}\alpha & \frac{1}{4}\beta & -\sum \end{pmatrix} \end{array}.$$

It is common now to see K80 parameterized with the ratio of transitions to transversions (κ) and the transversion rate equal to 1:

$$\mathbf{Q}_{\text{K80}} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -\sum & \frac{1}{4} & \frac{1}{4}\kappa & \frac{1}{4} \\ \frac{1}{4} & -\sum & \frac{1}{4} & \frac{1}{4}\kappa \\ \frac{1}{4}\kappa & \frac{1}{4} & -\sum & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4}\kappa & \frac{1}{4} & -\sum \end{pmatrix} \end{array}$$

which illustrates that only the relative values of the r_i matter. Following this the transition probabilities for K80 in terms of α and β are as follows:

$$p_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{1}{4}e^{-\beta t} + \frac{1}{2}e^{-\frac{1}{2}t[\alpha+\beta]} & \text{if } i = j \\ \frac{1}{4} + \frac{1}{4}e^{-\beta t} - \frac{1}{2}e^{-\frac{1}{2}t[\alpha+\beta]} & \text{if } i \longleftrightarrow j \text{ is a transition} \\ \frac{1}{4} - \frac{1}{4}e^{-\beta t} & \text{if } i \longleftrightarrow j \text{ is a transversion} \end{cases}$$

where $p_{ij}(t)$ is the probability of a nucleotide changing from i to j in time, t.

1.2.1.3 Felsenstein 1981

The F81 model includes unequal base frequencies, $\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$, but does not differentiate between transition and transversions, $r_1 = r_2 = r_3 = r_4 = r_5 = r_6 = \beta$. Therefore the model differs from both JC69 and the K80 models and has the following rate matrix:

$$\mathbf{Q}_{\text{F81}} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -\sum & \pi_C\beta & \pi_G\beta & \pi_T\beta \\ \pi_A\beta & -\sum & \pi_G\beta & \pi_T\beta \\ \pi_A\beta & \pi_C\beta & -\sum & \pi_T\beta \\ \pi_A\beta & \pi_C\beta & \pi_G\beta & -\sum \end{pmatrix} \end{array}.$$

The transition probabilities are:

$$p_{ij}(t) = \begin{cases} \pi_j - \pi_j e^{-\beta t} & \text{for } i \neq j \\ \pi_j - \pi_j e^{-\beta t} + e^{-\beta t} & \text{for } i = j \end{cases}.$$

It is interesting to note that of the models presented thus far, JC69 can be considered nested into both K80 and F81 but F81 and K80 cannot be nested in each other.

1.2.1.4 HKY 1985

The HKY85 [Hasegawa et al. 1985] incorporates the unequal base frequencies seen in the F81 model ($\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$) and allows for separate transition and transversion rate as in K80

($r_1 = r_3 = r_4 = r_6 = \beta$ and $r_2 = r_5 = \alpha$). The instantaneous rate matrix Q_{HKY85} is:

$$\mathbf{Q}_{\text{HKY85}} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -\sum & \pi_C\beta & \pi_G\alpha & \pi_T\beta \\ \pi_A\beta & -\sum & \pi_G\beta & \pi_T\alpha \\ \pi_A\alpha & \pi_C\beta & -\sum & \pi_T\beta \\ \pi_A\beta & \pi_C\alpha & \pi_G\beta & -\sum \end{pmatrix} \end{array}.$$

Then we can use the transition probabilities as denoted in Hasegawa et al.:

$$p_{ij}(t) = \begin{cases} \pi_j(1 + \frac{1-\eta_j}{\eta_j} e^{-\beta t}) + \frac{\eta_j - \pi_j}{\eta_j} e^{-t[\eta_j\alpha + (1-\eta_j)\beta]} & \text{if } i = j \\ \pi_j(1 + \frac{1-\eta_j}{\eta_j} e^{-\beta t} - \frac{1}{\eta_j} e^{-t[\eta_j\alpha + (1-\eta_j)\beta]}) & \text{if } i \longleftrightarrow j \text{ is a transition} \\ \pi_j - \pi_j e^{-\beta t} & \text{if } i \longleftrightarrow j \text{ is a transversion} \end{cases}$$

where η_j is the frequency of purines or pyrimidines based on the change occurring. For example if there is a change from $A \rightarrow T$, $\eta_T = \pi_C + \pi_T$. Here we see that the transition probabilities of the previous nucleotide models can be found from the HKY85 transition probabilities by setting $\alpha = \beta$ or $\pi_T = \pi_C = \pi_A = \pi_G = \frac{1}{4}$.

While we have outlined several nucleotide models there are others not discussed here such as F84 [Felsenstein 1984], TN93 [Tamura & Nei 1993], and UNREST [Yang 1994a]. The TN93 model and the F84 models are similar to the HKY85 model but their parameterization differs slightly and they are less commonly used. TN93 splits the α rate into α_1 and α_2 while F84 also splits the transition rates but does so using the equilibrium frequencies instead. The general unrestricted, or UNREST model, was introduced as a general model of nucleotide substitution similar to that of the GTR but is less well known. The UNREST model uses separate rate and frequency parameters for each substitution so that all nonidentical substitutions are free parameters. The vast majority of analyses with nucleotide models typically use one of JC69, K80, HKY85 or the GTR.

1.2.2 Amino Acid Models

Amino acid models of substitution allow us to better understand how change occurs at the protein level. Most amino acid models are empirical models made up of a 20 x 20 substitution matrix. While there are some parametric models of amino acid substitution, these are less common. Empirical amino acid models are relatively simple to compute. There is a risk with empirical models that the given matrix doesn't fully represent the data in question, which is why parametric models can be useful.

The first empirical model of amino acid substitution presented by Dayhoff et al. [1978] introduced what are widely known as point accepted mutation (PAM) or Dayhoff substitution matrices. The empirical models consisted of the 20 by 20 substitution matrices along with the 20 amino acid frequencies. Dayhoff et al. calculated the substitution matrices in such a way as to lessen the chance that an observed change between sequences was due to multiple mutations. This was done by using sequences with at least 85% similarity. The substitution matrix was constructed by counting the amino acid changes along a phylogeny. Each PAM_x matrix estimates the rate of substitution if x amino acid changes occur every 100 amino acids.

Jones et al. [1992] updated the matrices calculated in Dayhoff et al. [1978] using a larger database. Rare substitutions can be underrepresented or absent from the empirical matrices due to the way they are calculated. Other methods of estimating empirical substitution matrices have been introduced that are only suitable for smaller datasets. More recently Whelan et al. [2001] introduced a new method for estimating the substitutions matrices that combines the strength of empirical and parametric models. Unlike nucleotide and codon substitution models, parametric amino acid substitution models are less common than their empirical or semi-empirical counterparts though they do exist [Liberles et al. 2012]. Most analyses that utilize amino acid models use some updated form of the PAM or Jones et al. [1992] matrices.

1.2.3 Codon Models

Presented in this section is a brief overview of codon models and some of the subsequent extensions. For several informative reviews see: Anisimova & Kosiol [2009], Arenas [2015], and Delport et al. [2008]. Building upon nucleotide and amino acid models, codon models utilize the codon — a triplet of nucleotides — as the unit of evolution instead of individual nucleotides or amino acids. Codon models and amino acid models are both useful for investigating the evolution of proteins. Nucleotide and codon models can both account for changes in the sequence at the nucleotide or DNA level. A codon model's advantage lies in its ability to account for the redundancy of the genetic code and incorporate dependencies among sites within a codon. Codon models are able to account for the synonymous and nonsynonymous changes in the nucleotide sequence, while amino acid models can only describe nonsynonymous changes.

1.2.3.1 Muse and Gaut/Goldman and Yang 1994

Appearing back to back in a 1994 issue of Molecular Biology and Evolution, the MG94 [Muse & Gaut 1994] and the GY94 [Goldman & Yang 1994] models were the first codon models published. Here, I will outline the similarities and the key difference between these two models. Both models provide a framework for investigating the evolution of protein-coding regions in sequences. Individual codons are assumed to evolve independently according to a 61 state continuous time Markov chain, while nucleotide positions within each codon are not independent of each other. The major difference between the two models is the parameterization of the target frequency vector, π.

When it was first introduced, the instantaneous rate matrix for GY94 included both the transition/transversion ratio seen first in the K80 model (κ) and the ratio of synonymous to nonsynonymous substitutions (ω). The original GY94 model also included a parameter to account for the physicochemical distances between the 20 amino acids. Here we present the rate matrix in a

simplified form that is now more commonly used:

$$q_{ij} = \begin{cases} 0 & \text{if more than 1 substitution occurs} \\ \pi_j & \text{for a synonymous transversion} \\ \omega\pi_j & \text{for a nonsynonymous transversion} \\ \kappa\pi_j & \text{for a synonymous transition} \\ \kappa\omega\pi_j & \text{for a nonsynonymous transition} \end{cases}$$

where i and j are codons and π_j is the frequency parameter of the target codon j.

The MG94 model as it was originally parameterized used equal base frequencies at each codon position and had separate parameters for the synonymous substitution rate (α) and the nonsynonymous substitution rate (β). Now, these terms are often simplified to the ω ratio where $\omega = \frac{\beta}{\alpha}$ and the transition/transversion ratio κ is often included. A general form of the instantaneous substitution rate from codon i to codon j for MG94 is:

$$q_{ij} = \begin{cases} 0 & \text{if more than 1 substitution occurs (e.g. AAA} \rightarrow \text{AGC)} \\ \pi_{j_n} & \text{for a synonymous transversion (e.g. GTC} \rightarrow \text{GTA)} \\ \omega\pi_{j_n} & \text{for a nonsynonymous transversion (e.g. GTC} \rightarrow \text{GAC)} \\ \kappa\pi_{j_n} & \text{for a synonymous transition (e.g. GTC} \rightarrow \text{GTT)} \\ \kappa\omega\pi_{j_n} & \text{for a nonsynonymous transition (e.g. GTC} \rightarrow \text{ATC)} \end{cases}$$

where π_{j_n} is the frequency of the target nucleotide n of codon j. For the example listed above when GTC $\rightarrow$ GTA, $\pi_{j_n} = \pi_A$. This specification of the equilibrium frequencies is the key difference between MG94 and GY94. We can better understand the distinction between the two if we consider that under the MG94 model a given codon composed of nucleotides at n, l, and m has the frequency:

$$\pi_j = \frac{\pi_n \pi_l \pi_m}{1 - \pi_{\text{stop}}}$$

where π_{stop} is the sum of the frequencies for the stop codons.

1.2.3.2 Extensions of Codon Models

Most modern codon models are simply extensions of these two original models and are commonly termed GY-style or MG-style depending on which one they are based on [eg: Kosakovsky Pond & Muse 2005; Mayrose et al. 2007b; Yang et al. 2000]. However, nonparametric codon models exist as well [eg: Doron-Faigenboim & Pupko 2006; Schneider et al. 2005; Schöniger et al. 1990]. Here, we will discuss some of the basic codon model extensions and their uses.

Mutation-selection models are an extension of codon models that provide a framework for separating mutational bias and selective pressures. These can be useful in looking at codon bias. The first mutation-selection model that utilized codon models was introduced by Nielsen et al. [2006]. This model extended the GY94 and MG94 codon substitution models to include two categories of codons, preferred and unpreferred. From this mutation-selection model several others have arisen, including the FMutSel model from Yang and Nielsen 2008 that improves upon the previous model by using mutational bias to describe the frequency parameter and the codon bias separately.

Other codon models account for site dependencies. While the assumption that sites evolve independently of each other is a mathematically convenient one, it is often not the case in real data [Felsenstein & Churchill 1996]. Distant sites may have important interactions at the protein level and sequences often have dependencies among sites. The majority of codon models that include some sort of site dependency differ greatly in structure to the MG and GY type models we have discussed. However, one recent extension of codon models includes a GY-type model designed to account for the contextual differences in broadly neutralizing antibody lineages [Hoehn et al. 2017]. Other models like Mayrose et al. [2007a], account for site dependencies at the codon level using hidden Markov models. Models also exist that account for dependencies among codons due to protein folding [Robinson et al. 2003].

Empirical models differ from parametric models in that they use estimates of the substitution matrix from a real data set. In the case of codon models, empirical models require a large data set

from which to estimate as the substitution matrix is 61 by 61. The first widely accepted empirical codon matrix was introduced by Schneider et al. [2005] and was built off of 17,502 alignments of vertebrate DNA totaling 8.3 million codons. There was a previous attempt at building an empirical codon model in 1990, but due to the limited number of sequences available it was not widely used [Schöniger et al. 1990].

Similar to the ideas behind empirical amino acid models, Schneider et al. [2005] developed this empirical model as more protein-coding DNA sequences became readily available in the early 2000s. The data set they used was carefully selected to include only accurate alignments with few uncertainties. The sequences were also filtered so that they covered large enough distances between them that rare substitutions would be accounted for while not being too distant as to include too many multiple and reversus substitutions. The number of synonymous substitutions was also used as a filter. Vertebrate DNA was chosen specifically as vertebrates have the least variable range of codon usage bias. In simulations, a comparison between this codon model and an amino acid model derived from it showed that the codon model more accurately aligned sequences. However, because of the narrow phylogenetic scope of data used to construct the empirical matrix, it is not widely applicable. Data sets that are not comprised of vertebrate DNA or that cover a different range of distances cannot readily use this matrix. Therefore, we will discuss two semi-empirical models that followed this one and allow for more flexibility.

Semi-empirical models combined aspects of parametric and empirical models. They typically use an empirical data set to establish the substitution matrix and then estimate parameters from there. One such model is known as the mechanistic-empirical combined (MEC) model [Doron-Faigenboim & Pupko 2006]. This model uses an empirical amino acid substitution matrix and integrates the empirical amino acid replacement rates into a mechanistic codon model. The parametric codon model used is similar to GY94 in that each substitution is weighted by two parameters, the transition/transversion ratio and a parameter that accounts for the similarities between codons. However, while this second parameter is represented by physicochemical distance in the original GY94 model, the MEC model uses the empirical amino acid transition probabilities. Additionally, the

MEC model allows for multiple nucleotide changes to occur simultaneously while most mechanistic models only allow a single instantaneous nucleotide change. A comparison to both a full empirical model (Schneider as presented above) and a full mechanistic model [Nielsen & Yang 1998] showed the MEC improved on the log-likelihood score of both. This suggests that mechanistic models may benefit from the information included in the empirical models while the empirical model may benefit from the addition of parameters. The MEC also has the added benefit that by substituting in different amino acid matrices it allows for more flexibility in the types of data it can handle.

Another model that mixes aspects of empirical models and parametric models is the empirical codon model (ECM) [Kosiol et al. 2007]. While the previous models used matrices formed via a counting method, the empirical codon matrix of the ECM model is estimated directly from a database of sequence alignments using a method introduced in Whelan et al. [2001]. The empirical substitution matrix is then combined with parameters found in parametric models. However, the parameters in the ECM are not directly comparable to those from parametric models as they represent deviations from the averages and not the ratios themselves. The ECM was compared to standard parametric models, and an increase in likelihoods was seen.

1.3 Likelihood

Likelihood is an approach to statistical inference that is commonly used in molecular evolution. In 1981, Felsenstein published a landmark paper that introduced the framework for using maximum likelihood to estimate evolutionary trees from DNA sequence data [Felsenstein 1981]. This likelihood framework can be used to give maximum likelihood estimates (MLEs) of parameters of interest from substitution models such as the substitution rate, κ, or ω.

The likelihood itself, L, is the probability of an observed dataset D given the model parameters θ, $L = \Pr(\theta; D)$. From the likelihood equation, we can find the parameter values that maximize the likelihood of the observed data. These are maximum likelihood estimates or MLEs.

Here, we will present an example of computing the likelihood for a single site, h, in a sequence

along the given tree in figure 1.2:

$$L_h = \pi_{h_0} P_{h_0 h_1}(t_1) P_{h_1 h_2}(t_2) P_{h_1 h_3}(t_3) P_{h_0 h_4}(t_4) P_{h_4 h_5}(t_5) P_{h_4 h_6}(t_6),$$

where π_{h_0} is the the chance of observing a specific nucleotide at the root of the tree, node 0, and $P_{h_i h_j}(t_i)$ is the probability that a site originally in state h_i in state h_j after t units of time. However, we do not know the ancestral sequences excepting very rare cases. Thus to calculate the likelihood we have to sum over all of the possible states at each unknown ancestral node h_0, h_1, h_4 such that our likelihood becomes:

$$L_h = \sum_{h_0} \sum_{h_1} \sum_{h_4} \pi_{h_0} P_{h_0 h_1}(t_1) P_{h_1 h_2}(t_2) P_{h_1 h_3}(t_3) P_{h_0 h_4}(t_4) P_{h_4 h_5}(t_5) P_{h_4 h_6}(t_6).$$

However, with many data sets we are unsure of where the root of the tree truly belongs. Thus, we can use an unrooted tree instead, as shown in figure 1.3. Note here that t_1^* for the unrooted tree is simply $t_1 + t_4$ from the rooted tree [Felsenstein 1981]. The likelihood of the unrooted tree can be calculated as follows:

$$L_h = \sum_{h_1} \sum_{h_4} \pi_1 P_{h_1 h_4}(t_1^*) P_{h_1 h_2}(t_2) P_{h_1 h_3}(t_3) P_{h_4 h_5}(t_5) P_{h_4 h_6}(t_6).$$

The likelihood for the entire tree for a sequence of length n is the product of the site likelihoods:

$$L_T = \prod_{h=1}^{n} L_h.$$

We are able to compare the fits of models by using the likelihood ratio test or LRT. We compare a simpler null model to a more general alternative model if the two are nested. The LRT compares the likelihood of the null (L_0) to the likelihood of the alternative (L_A) using the test statistic $2(ln(L_A) - ln(L_0))$. Under the null model, this statistic follows a χ^2 distribution with degrees of freedom equal to the difference in free parameters between the two models. This provides us with one criterion to

determine if the alternative model is a better fit for the data than the null. For instance, this test may be used to compare two models, one with and one without selection in order to determine if there is evidence of selection.

Another possible criterion to compare model fit is the Akaike information criterion (AIC), $AIC = 2k - 2ln(L)$. The AIC takes into account the number of parameters, k, present in a given model. When choosing between two models the model with the lowest AIC is said to be the preferred model as it best represents the data with the fewest parameters. One advantage of AIC over LRT is that the models do not need to be nested.

1.3.1 Maximum Likelihood Estimation of D

As noted, the likelihood framework provides a convenient method of estimating parameters from substitution models. It is one of the primary procedures used throughout this dissertation and in the field of molecular evolution to estimate parameters and test hypotheses. Here, we present an example of using maximum likelihood to estimate a parameter from our simplest nucleotide model JC69. Specifically, we estimate the JC69 distance, D, between two sequences A and B of length n. The distance, D is the expected number of substitutions between sequences in time t:

$$D = -t \sum \pi_j Q_{jj}.$$

We see the distance, D, between two sequences for JC69:

$$D = \frac{3}{4}\beta t.$$

For this pairwise comparison a site h will either be a match or a mismatch. We know from the JC69 transition probabilities that the probability of a site being a match between the two sequences can be written with respect to D:

$$P_+ = \frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}D}.$$

From there we know the probability of a mismatch is $1-P(+)$:

$$P_{-} = \frac{3}{4} - \frac{3}{4} e^{-\frac{4}{3}D}.$$

The log likelihood of D for n sites will then be:

$$ln(L(D;A,B)) = \sum_{h=1}^{n} ln(L_h) = n_{+} ln(P_{+}) + n_{-} ln(P_{-})$$

where $n = n_{+} + n_{-}$. To estimate D we take the derivative of the log likelihood with respect to D and set it equal to zero which can be solved:

$$\widehat{D} = -\frac{3}{4} ln(1 - \frac{4}{3} d)$$

where d is the observed proportion of differences between sequences A and B.

This is just one example of how to use the likelihood framework to estimate a parameter. It should be noted that while we can find a closed form estimate for the JC69 D, as models become more complex it becomes impossible to derive these estimates.

1.4 Counting Methods

Counting methods allow us to estimate the evolutionary distance between a pair of sequences. These methods predate likelihood applications to molecular evolution [Fitch & Margoliash 1967]. They tend to be faster to compute and like likelihood methods, can be used with different sequence evolution models. Here we present the distance calculation for the JC69 model based on counting methods using the same notation.

From JC69 we see that the instantaneous rate of change is $\frac{3}{4}\beta$. The expected number of substitutions occurring in time t is $D = \frac{3}{4}\beta t$ as previously defined. For counting methods, we take two sequences A and B with n sites and count the number of sites that match (n_{+}) and the number of sites that do not match (n_{-}). We see that the observed proportion of differences between sequences

is:

$$d = \frac{n_{\_}}{n}.$$

We know the probability of a mismatch for JC69 is:

$$P_{\_} = \frac{3}{4} - \frac{3}{4} e^{-\frac{4}{3}D}.$$

Using the method of moments we set $d = P_{\_}$ and solve to estimate D:

$$\widehat{D} = \frac{3}{4} l\,n(1 - \frac{4}{3} d).$$

Note that for the JC69 distance the result from the counting method and the result from the likelihood method match. This is a special case and for other distance estimations the calculations become more complex and the counting and likelihood methods no longer agree. Fitch & Margoliash [1967] used counting methods to estimate the number of invariable sites in sequences before a likelihood framework had been introduced. These counting methods are still in use today and form the basis of popular methods like neighbor joining [Saitou & Nei 1987].

1.4.1 Codon Models and Counting Methods

The ability to categorize changes as synonymous or nonsynonymous makes codon models useful when selection is of interest. Prior to the introduction of codon models tests of selection relied on estimating the expected number of nonsynonymous substitutions per nonsynonymous site (d_N) and the expected number of synonymous substitutions per synonymous site (d_S). These estimates are typically achieved using counting methods [Li et al. 1985; Miyata & Yasunaga 1980b; Nei & Gojobori 1986].

Counting methods typically involve a similar set of steps. First, they classify and count the synonymous (s) and nonsynonymous (n) sites. Next, they count the number of sites at which a synonymous (s_d) or nonsynonymous (n_d) change occurs. Finally, they use that information to

calculate the ratio of the number of sites to the number of changes, $p = \frac{\text{changes}}{\text{sites}}$. Most counting methods also try to correct for multiple substitutions.

The first counting method for estimating d_S and d_N was published by Miyata & Yasunaga [1980a] shortly after DNA sequencing became easily accessible. The original Miyata and Yasunaga method assumed that all nucleotide changes were equally probable and used physicochemical distances between amino acids to weight pathways when more than one change occurred between codons. A simplified version of this method was introduced by Nei and Gojobori 1987 that does not weight each path and assumes they happen with the same probability.

Both Nei and Gojobori's and Miyata and Yasunaga's methods are similar to the JC69 and do not account for the transition/transversion bias. A third method introduced by Li et al. [1985] included the transition/transversion bias as they found ignoring it led to biased estimations. For instance, by not including this bias d_S was often overestimated leading to an underestimation of the $\frac{d_N}{d_S}$ ratio. Thus they incorporated the transition/transversion bias by partitioning the sites into different degeneracy classes.

Early tests of selection relied on these counting estimates of d_S and d_N. The ratio of $\omega = \frac{d_N}{d_S}$ was used to determine selective pressures. It has been shown that the number and location of synonymous and nonsynonymous sites change over time, and counting methods have difficulty accommodating for these shifts [Muse 1996]. Muse [1996] showed that the Nei and Gojobori estimates of both parameters were biased and that d_S estimates in particular suffered. He showed that d_S according to Nei and Gojobori method was, in fact, a function of both the synonymous and nonsynonymous rates. Despite these noted biases, counting methods still remain a popular way of inferring selection pressure among other estimates with the Nei and Gojobori method being the most widely used.

1.5 Rate Variation

The models we have discussed up until now have all assumed a constant substitution rate across sites. Biologically, we know that substitution rates can vary spatially from site-to-site and temporally along the branches of a given phylogeny [Hodgkinson & Eyre-Walker 2011]. Varying substitution rates can be attributed to mutational hotspots or selective pressures acting on sites. Not accounting for rate variation, no matter the source, can lead to potentially incorrect inferences [Kosakovsky Pond & Muse 2005; Wakeley 1994; Yang 1994b]. Here we will introduce the concept of rate variation as it pertains to substitution models.

One of the first instances of partitioning sites into different rate categories came from Fitch & Margoliash [1967]. In this paper, they tested amino acid sequences of cytochrome c to determine if conserved codon sites are unchanged because a substitution occurring in them would be deleterious. To do this they first categorized the codons based on whether a substitution had occurred or not. From the unchanged category, they further categorized the codons based on how a potential substitution would impact the sequence. If a substitution in these sites would be non-deleterious they are considered mutable but unchanged, but if a substitution would cause an immediate deleterious response the sites are referred to as invariable sites. By partitioning the sites into these categories, they showed that the substitution rate for every codon in the sequence was not the same as invariable sites do not experience substitutions. This model was improved upon later by Fitch & Markowitz [1970]. Eventually, substitution models that partitioned sites into variable and invariable were developed [Shoemaker & Fitch 1989]. This is now commonly added to most models of sequence evolution.

Another common way of accounting for rate heterogeneity across sites treats the substitution rate at site s, z_s, as a random variable drawn from a statistical distribution with probability density function $f(z_s;\theta)$. Here, θ is a parameter of the density function and may also be a vector. Recall that when we estimate rates, we cannot separate the rate and time parameters. So, using the HKY85 model as an example, we can estimate κt but not κ or t separately. When we take this approach

to model site-to-site rate heterogeneity, mathematically we are actually letting t vary across sites. However, since t and the substitution rate parameter θ cannot be separated, this gives us the variation of θt across sites, which lets sites change at different rates.

The most commonly used distribution for rate variation is the gamma distribution, although others have been tested and used [e.g. Golding 1984; Kosakovsky Pond & Frost 2005; Waddell & Steel 1997]. In this setting, the gamma with a shape and scale parameter equal to λ is used for technical reasons. The gamma distribution is chosen because varying the shape parameters allows us a wide range of variability in the shape of the distributions, see figure 1.5. The probability density function for $z_s \sim Gamma(\lambda, \lambda)$ is:

$$f(z_s;\lambda) = \frac{1}{\Gamma(\lambda)\lambda^{\lambda}} {z_s}^{\lambda-1} e^{-\frac{z_s}{\lambda}},$$

where $\Gamma(\lambda)$ is the gamma function. The sitewise likelihood function for a specific site,s, in an alignment D with a tree T is

$$L(\theta; D_s, T) = \int_0^{\infty} Pr(D_s|z_s;T) f(z_s;\theta) dz_s$$

where $Pr(D_s|z_s;T)$ is the probability of the observed data at site s given the random variable z_s. Note that this is simply integrating the likelihood function over all possible z_s rates. The full likelihood function, as before, is the product across sites:

$$L(\theta; D, T) = \prod_{s=1}^{S} L(\theta; D_s, T) = \prod_{s=1}^{S} \int_0^{\infty} Pr(D_s|z_s;T) f(z_s;\theta) dz_s$$

where S is the total number of sites in the alignment.

Using the continuous gamma distribution to estimate the rate variation across sites is computationally intense because it is impossible to find a closed form expression of the likelihood, and thus it is only suited for small data sets. In 1994, Yang published a method of including rate variation in a model that utilized a discrete gamma distribution to approximate the continuous distribution [Yang 1994b]. This reduced the computational burden and was better suited for larger datasets. While what is described in the original paper is the gamma distribution, the same methodology can be generalized to any continuous distribution [Felsenstein 2001]. Here, if we have $f(z_s;\theta)$ as a discrete

distribution with k rate classes the complete likelihood function is:

$$L(\theta;D,T)=\prod_{s=1}^{S}L(\theta;D_s,T)=\prod_{s=1}^{S}\sum_{i=1}^{k}Pr(D_s|z_s=z_i;T)Pr(z_s=z_i;\theta).$$

Notice that the integral from the continuous method is replaced by a discrete sum of k terms. With this discrete method, rates over sites are random variables drawn from k rate categories approximating the gamma distribution. If we define the i^{th} category to be the region from x_i to y_i, the probability of a site falling in category i is:

$$p_i=\int_{x_i}^{y_i}h(z_s=z_i|\lambda)dz_s$$

where x_k are the left boundaries of the rate categories and y_k are the right boundaries. In figure 1.4 these left and right bounds are represented by the boundaries of each histogram bin. If we choose to have the rate categories contain equal proportions, $p_i=\frac{1}{i}$. If we choose to let the rate categories have unequal frequencies, then they are usually estimated separately from the data. The General Discrete Distribution [Kosakovsky Pond & Frost 2005] can be used instead of the gamma or the quadrature methods introduced in Felsenstein [2001] can be used to approximate gamma or any other specified distribution. In figure 1.4 we choose to use four equiprobable rates to approximate the gamma distribution with $\lambda=2$. When calculating a rate, z_i, for interval i we can choose to use either the mean or median value of the interval. The dashed lines in figure 1.4 represent the means of each interval and give us the z_i for each.

Using several empirical data sets, Yang [1994b] found that for $k\geq 3$ the increase in the likelihood was negligible, especially when accounting for the additional computational cost at the time. Most modern phylogenetic estimation software uses a default setting of either $k=3$ or $k=4$.

1.5.1 Nonsynonymous Rate Variation

Codon models can be modified to account for site-to-site variation of nonsynonymous rates. Recall that in the MG94 and GY94 models, $\omega=\frac{\beta}{\alpha}$. The most common practice when modeling nonsyn-

onymous rate variation is to assume a constant synonymous substitution rate from site to site and set $\alpha = 1$ for all sites [Nielsen & Yang 1998; Yang et al. 2000]. If we define ω_s to be $\frac{\beta_s}{\alpha_s}$, then any variation that occurs in ω_s can be attributed to the nonsynonymous substitution rates. Here, we draw nonsynonymous rates from $\beta_s \sim Gamma(\lambda, \lambda)$:

$$f(\beta_s) = \frac{1}{\Gamma(\lambda)\lambda^\lambda} \beta_s^{\lambda-1} e^{\frac{\beta_s}{\lambda}}$$

where the shape parameter of the distribution is λ. The site-wise likelihood of β_s given an alignment D_s is:

$$L(\beta_s; D_s) = \prod_{s=1}^{S} \int_0^\infty Pr(D_s|\beta_s) f(\beta_s) d\beta_s$$

where we cannot find the closed form expression of this integral due to its mathematic complexity. Instead we use the discrete distribution calculation so that:

$$ln(L(\beta_s; D_s)) = \prod_{s=1}^{S} \sum_{i=1}^{k} Pr(D_s|\beta_s = \beta_i) Pr(\beta_s = \beta_i)$$

where k is the number of discrete rate categories and β_i is the rate for the i^{th} rate category. Again, note that by using the discrete method modeling rate variability, we simplify the integral from the continuous into a sum. Accounting for the site-to-site nonsynonymous rate variation improves model fit [e.g. Mayrose et al. 2005; Yang 1996; Yang et al. 2000] and is supported by biological evidence [Echave et al. 2016]. It has become a standard component of sequence substitution models and is a step towards making inferences based on more realistic models.

1.5.2 Synonymous Rate Variation

The codon models we have mentioned thus far include a standard assumption that the synonymous substitution rate does not vary from site-to-site. Perhaps the most common reason for this assumption is that synonymous changes, which are also referred to as "silent", are thought to be invisible when it comes to selection at the protein level. This silent assumption is also computationally

convenient as it cuts down on computational cost. Models have to estimate fewer parameters if the synonymous rate is assumed constant across sites.

The site models introduced in Yang et al. [2000] allow ω to vary from site to site but attribute all of that variation to the nonsynonymous substitution rate (β) by assuming that the synonymous substitution rate is constant across sites ($\alpha = 1$). This constant synonymous rate assumption is found in the majority of other models incorporating rate variation as well. However, several studies show synonymous rates vary among genes [eg: Gaut et al. 1996; Hanada et al. 2004; Mouchiroud et al. 1995; Ohta & Ina 1995; Wolfe & Sharp 1993; Zhu et al. 2014]. There is also evidence that synonymous rates vary between gene types and different gene regions [eg: Miyata & Yasunaga 1980b; Wolfe et al. 1987; Wolfe et al. 1989]. Site-to-site synonymous rate variation within genes has also been reported [Hurst & Pál 2001; Kosakovsky Pond & Muse 2005].

A major result of this dissertation, detailed in chapter 2, shows the prevalence of synonymous rate variation in Metazoan mitochondrial DNA (mtDNA). Using the coefficient of variation (CV $= \frac{\sigma}{\mu}$) to describe the magnitude of site-to-site variation, figure 1.6 shows that the synonymous CV is non-zero for all 721 alignments analyzed and that it is of a comparable magnitude to that of the nonsynonymous CV. This is just one more piece of evidence showing synonymous rates do in fact vary from site to site. This is similar to the results in [Kosakovsky Pond & Muse 2005], where 9 out of the 10 data sets tested rejected the null hypothesis of a constant synonymous rate and all had non-zero synonymous CVs. Our study in chapter 2 is one of the largest investigations of this phenomenon to date.

Kosakovsky Pond & Muse [2005] were the first to simultaneously model site-to-site variation of α and β. They let α and β vary across sites s with independent gamma distributions such that α_s and β_s are modeled by a bivariate distribution with density functions:

$$f(\beta_s, \alpha_s) = f(\beta_s) f(\alpha_s).$$

Using the same notation as in previous sections, the likelihood using this distribution is

$$L = \prod_{s=1}^{S} \int_0^\infty \int_0^\infty Pr(D_s|\beta_s, \alpha_s) f(\beta_s, \alpha_s) d\beta_s d\alpha_s$$

for which there is no closed form expression. Using the same approach described above they approximated the likelihood with a discretized version:

$$L = \prod_{s=1}^{S} \sum_{i=1}^{k} \sum_{j=1}^{m} Pr(D_s|\beta_s = q_{ij}, \alpha_s = r_{ij}) Pr(\beta_s = q_{ij}, \alpha_s = r_{ij}). \tag{1.1}$$

If α_s and β_s are sampled from independent distributions, equation 1.1 can be simplified to:

$$L = \prod_{s=1}^{S} \sum_{j=1}^{m} Pr(\alpha_s = r_j) \sum_{i=1}^{k} Pr(D_s|\alpha_s = r_j, \beta_s = q_i) Pr(\beta_s = q_i).$$

where r_j and q_i are the possible values of α_s and β_s. Note that discretizing the independent continuous functions results in two independent summations over k and m possible rate categories. Again, as shown in figure 1.6, we see that the synonymous CV and the nonsynonymous CV are generally of different magnitudes and thus estimating them separately is advised. Also note that as was the case when accounting for nonsynonymous rate heterogeneity, accounting for synonymous rate heterogeneity has been shown to improve model fit [e.g. Kosakovsky Pond & Muse 2005; Mayrose et al. 2007a].

1.6 Tests of Selection

One of the useful applications of codon models is the detection and measurement of measure selective pressures. These methods typically rely on three major categories of selection: neutral, negative and positive, and inferences about selection are based on the value of the $\omega = \frac{\beta}{\alpha}$ ratio. When $\omega = 1$ and synonymous and nonsynonymous changes are occurring at equal rates, we say there is evidence of neutral selection. There is evidence of purifying or negative selection when $\omega < 1$. When nonsynonymous changes result in nonfunctional or deleterious amino acids they are

fixed less often in the population, hence the synonymous changes outnumber the nonsynonymous changes. Positive selection is typically the result of helpful or positive nonsynonymous changes being fixed in the population, thus when positive selection occurs the nonsynonymous changes may outnumber the synonymous changes and we see evidence of this when $\omega > 1$.

Yang et al. [2000] showed that using maximum likelihood to estimate the ω ratio gave the best results out of the available methods at the time. As codon models became more widely used, new tests of selection emerged as well [e.g. Kosakovsky Pond et al. 2011; Murrell et al. 2015; Yang & Rannala 2012]. Many tests of selection compare a null model, one that typically restricts the $\omega = 1$, to an alternative model, one that allows for either positive or negative selection depending upon which type of selective pressure is of interest.

As previously mentioned, rate variation is a feature that can be incorporated into models of selection. When using codon models to test for selection, the ω ratio is often varied in three ways that make up the three major types of tests for selection. The first of these use "site models" where ω varies across the sites of the sequences. Next, "branch models" allow ω to vary along the branches of the lineages being tested. Finally, combining the two previous strategies we get "branch-site models". Branch-site models allow ω to vary both spatially across sites and temporally across branches. The following sections will describe each of these strategies and give some examples of common implementations.

1.6.1 Site Models

Here we will discuss two major strategies for site models, random effects likelihood (REL) and fixed effects likelihood (FEL). REL methods use a random variable drawn from a statistical distribution to describe the rate variation across sites. FEL methods rely on specifying a prior partitioning of codon sites. Each method has its unique benefits and drawbacks and for a detailed comparison of these methods and counting methods please see Kosakovsky Pond et al. [2005]. As previously noted, the tests of selection discussed in this section allow rates to vary only across sites and not across branches.

REL methods model synonymous and nonsynonymous rates using a predefined distribution and treat the parameters, such as ω, as random variables estimated from a discretized distribution. This discretized distribution is most often the gamma or general discrete distribution, although any distribution could be used. For REL methods a test of positive selection typically involves the likelihood ratio test (LRT) between two models. As previously mentioned using three rate categories is a standard of the field so here we will discuss two models with three ω rate categories where $\omega_1 \leq \omega_2 < \omega_3$. Depending on the model used each rate category may have equiprobable proportions or unequal proportions. The null model will often disallow positive selection, either by restricting $\omega_3 \leq 1$ or setting the proportion of sites with $\omega_3 > 1$ to zero. The alternative model will often allow for positive selection by setting $\omega_3 > 1$. Tests then look like: $H_O : \omega_3 \leq 1$ versus $H_A : \omega_3 > 1$.

Yang et al. [2000] introduces the first series of random effect models, known as the M-series of models. This series of models includes over 8 different models with different approaches to varying ω across sites. The M5 model draws ω from a gamma distribution. The M3 model is known as the discrete model and uses the common $k = 3$ rate categories to approximate ω varying across sites. Due to the way these models are proposed they can be easily used as REL methods and paired for a LRT depending on the hypothesis of interest. For example, if you wish to test for variation of selective pressure across sites you might pair the M3 model with the M0 model, which is the one ratio model and has constant ω across sites. Several studies test the pairings of these models and propose additions [Swanson et al. 2003; Wong et al. 2004].

FEL methods allow sites to be assigned to separate rate classes based on protein folding, protein domain information, or inferred recombination breakpoints. Partitioning sites to better test for selection predates ML methods, as Hughes and Nei partitioned sites into binding and non-binding sites for the Major Histocompatibility Complex and used a pairwise comparison of synonymous and nonsynonymous sites to test for selection [Hughes & Nei 1988]. Building on that study but using FEL methods, Yang & Nielsen [2002] partitioned sites in MHC genes based on whether or not they bound to foreign peptides to test for positive selection. However, when partitioning sites, treating each site as its own partition can quickly lead to overparameterization [Felsenstein 2001]. One of

the major drawbacks of FEL methods is that the *a priori* knowledge needed to split the sites into partitions is not often available. If this is the case but FEL methods are still of interest then it is possible to use the maximum likelihood framework to partition the data.

1.6.2 Branch Models

Branch models, also referred to as lineage models, allow ω to vary across branches but not across sites. Branch models are particularly useful if you want to determine if a specific taxon or clade is experiencing selective pressures. The first two tests that specifically measured selective pressures across branches were those of Messier & Stewart [1997] and Crandall & Hillis [1997]. These tests did not use codon models. Instead, ancestral sequences were reconstructed using a combination of parsimony and maximum likelihood techniques. Then, d_N and d_S were estimated for each branch using pairwise counting methods and the ratio, $\frac{d_N}{d_S}$ was used to measure selective pressures. Bias from the phylogenetic reconstruction of the ancestral sequences was likely reflected in the inferences made about selection on lineages using these two methods.

In Yang & Nielsen [1998] codon-based likelihood models that allow ω to vary for individual branches were introduced. By using a maximum likelihood framework, they considered all possible paths from ancestral sequences to descendants; therefore there is less inherent bias than in the methods employed by Messier & Stewart and Crandall & Hillis. Like the M-series of models in Yang et al. [2000], branch models can incorporate different numbers of parameters based on foreground and background branches. Consider the case where we have reason to suspect selection occurring along the branch leading to hominoids. We could set up a two ratio branch model where the hominoid branch has ω_h and all other branches have ω_0. This would allow us to test against the one ratio model ($\omega_h = \omega_o$) to see if there is evidence of selection along the hominoid branch. Branch models are not limited to only two ratios; in fact, a free ratio branch model allows for a separate ω ratio along each branch of the tree [Yang 1998].

Yang & Nielsen [1998] applied these codon-based likelihood models to the same lysozyme data used by Messier & Stewart [1997]. Both studies found evidence of selection along the hominoid

branch. However, Yang & Nielsen found no evidence of positive selection along the colobine branch while Messier & Stewart did. Yang & Nielsen points out that this lack of evidence of positive selection along the branch does not mean positive selection did not occur as both methods assumed unrealistically that rates do not vary along sites.

One of the major drawbacks of branch models is that they may require prior knowledge to partition the phylogeny into foreground and background branches [Yang & Nielsen 1998]. However, as was the case with site models, there are a few methods that do not rely on a prior biological information to categorize branches into rate categories. A fully Bayesian approach is one such method [Kosiol et al. 2007] as is the genetic algorithm [Kosakovsky Pond & Frost 2005] which assigns the ω ratio by using maximum likelihood to optimize the fit to the data. A third option is using clustering or dynamic programming to assign ω ratios to branches without any prior information [Dutheil et al. 2012; Zhang et al. 2011]. Here we have presented an overview of branch models as tests of selection. Note that these models will only measure selective pressures across branches.

1.6.3 Branch-Site Models

The first widely recognized model that allowed selective pressures to vary over both branches and sites was published by Yang & Nielsen [2002]. In branch-site models the phylogenetic tree is split into two categories: foreground and background. Foreground branches are the branches that are being tested for positive selection while background branches are not. In the original, paper branches were categorized as foreground or background based on prior knowledge of where positive selection was likely to have occurred. The branch-site models in Yang & Nielsen [2002] are built on GY94 codon substitution models. Like site and branch models, branch-site models generally rely on a discretized distribution to estimate a number of ω categories. However, unlike site and branch models, branch-site models use a combination of foreground and background categories.

In the original branch-site model of Yang & Nielsen [2002] there are four site classes and only three ω parameters; see table 1.1 for a breakdown of the parameter combinations and proportions. The first class of sites contains mostly highly conserved sites and has an ω_1 parameter that is very

small. The second class of sites contains sites that are neutral or weakly constrained. They have an ω_2 value that is or is close to 1. The third class of sites refers to sites that when on foreground branches have an ω_3 parameter that is greater than 1 but on background branches the site has a value of ω_1. Similarly, the fourth class of sites includes those that are neutral or nearly neutral when on background branches (ω_2) but when on branches in the foreground lineage are positively selected with a value of ω_3. For each of these latter two site classes, this means that somewhere on the foreground branches, the site has undergone an event of positive selection. Each site class has an associated weight or proportion. We assume that sites under positive selection on foreground branches are equally as likely to come from a site with ω_1 or ω_2 in the background branches. The two models proposed in this paper are known as MA and MB. The MA model sets $\omega_1 = 0$ and $\omega_2 = 1$ and is an extension of the M1 model introduced in Nielsen & Yang [1998]. The MB model is an extension of the M3 model of Nielsen & Yang [1998] with $k = 2$ rate categories as ω_1 and ω_2 are estimated freely. Zhang [2004] found that for MA when the assumptions of the model are violated, inferences may not be correct. This finding led to improved MA models being introduced [Zhang et al. 2005]. One such improved MA model assumed $\omega_2 \leq 1$ instead of the previously assumed $\omega_2 = 1$.

One drawback to branch-site models is again the requirement of prior knowledge to specify which branches are in the foreground and background. Similar to branch models, you can use the genetic algorithm [Kosakovsky Pond & Frost 2005] or Bayesian methods [Kosiol et al. 2007] to assign branches to categories but these become more computationally intense for branch-site models. A branch by branch LRT approach is less computationally intense [Anisimova & Yang 2007]. However, this multiple LRT method suffers from a lack of power in detecting selection.

Since their introduction branch-site models have been examined and refined. Fletcher & Yang [2010] investigated the impact of indels on inferences made by branch-site models. Similarly, clade models were introduced [Bielawski & Yang 2004; Forsberg & Christiansen 2003; Weadick & Chang 2012]. Clade models are essentially branch-site models applied to specific clades. As when categorizing the tree into foreground and background branches, clade models require prior knowledge to categorize the trees into clades. By looking at selection on the clade level, we are able to get a

better understanding of functional divergence, possible gene duplication events, speciation, or even parasitic adaptation to hosts [Anisimova & Kosiol 2009].

A key branch-site model framework is the Branch-Site Random Effects Likelihood (BSREL) method [Kosakovsky Pond et al. 2011]. The BSREL method allows branches, as well as sites, to vary independently in a REL framework. Therefore at each site, the rate is chosen independently of any other site or branch. An extension of the BSREL method, the adaptive Branch-Site Random Effects Likelihood (aBSREL) method allows each branch to have a varying number of rate categories. For example, BSREL may assume that all branches have three ω categories to which sites can be assigned. However, aBSREL may assign one, two, or three ω categories to a branch, choosing the optimal number of rate categories using a small sample AIC ($AICc$). Both BSREL and aBSREL use the LRT to test for branches with evidence of selection.

Also built on the BSREL framework, BUSTED allows the detection of gene-wide positive selection events [Murrell et al. 2015]. Basically, BUSTED tells us whether there is evidence of selection anywhere on foreground branches of a gene tree but does not tell us the location of that selection. Like most branch-site tests of selection, BUSTED allows for the partitioning of branches into foreground and background if a biological hypothesis indicates the necessity. It can also be used without one, which allows the test to be performed over the entire phylogeny. For example, one study of HIV-1 genes [Lorenzo-Redondo et al. 2016] used BUSTED to see if selective pressures on spatially distinct sites differ. A study on virus interacting proteins utilized BUSTED to help identify proteins that experienced selection in both human and greater mammalian lineages [Enard et al. 2016]. BUSTED is often used as a preliminary test before trying to identify site-specific selective pressure. For example, Pacheco et al. [2018] used BUSTED to determine if selection acted on any proportion of sites for the Haemosporidian mitochondrial genome before using a site-specific test.

The instantaneous rate matrix for the codon model used in BUSTED is:

$$q_{ij} = \begin{cases} 0 & \text{if more than 1 substitution occurs} \\ \theta_{ij}\pi_{j_n} & \text{for a 1 nucleotide synonymous substitution} \\ \theta_{ij}\omega_k\pi_{j_n} & \text{for a 1 nucleotide nonsynonymous substitution} \end{cases},$$

where θ_{ij} are the underlying nucleotide substitution rate parameters. Here, they assume θ_{ij} follow the GTR but other nucleotide substitution models can be used as well. π_{j_n} is the target nucleotide n in codon j as in MG94 and ω_k is the ω ratio for rate category k. Since they used a discrete distribution with k rate categories, there are Q_k instantaneous rate matrices as well. BUSTED defaults to three rate categories ($k = 1,2,3$) and shares rates and parameters across sites and branches within each partition. The ω rates are estimated separately for foreground and background branches and are restricted such that $\omega_1 \leq \omega_2 \leq 1 \leq \omega_3$. As ω_3 is the rate of the category with the largest magnitude, we use its estimate to infer the level of selection present. BUSTED differs from the standard branch-site model because it estimates these rate parameters separately instead of assigning prefixed parameters to the branches. Similarly to the branch-site model, BUSTED uses likelihood ratios and a null model with no positive selection allowed ($\omega_3 = 1$), to determine if positive selection occurs anywhere within the lineage. When the null model is rejected, that tells us that there is positive selection at least one site on at at least one foreground branch. Because there are fewer restrictions on how the rate parameters are assigned, this does not rule out the possibility of positive selection having occurred on a background branch. If evidence of positive selection is present, then in order to determine the specific sites that are under positive selection pressures, a site-wise likelihood ratio is calculated. It should be noted that BUSTED in the form presented here and in its original paper does not include synonymous rate variation across sites. This will be addressed later in Chapter 4.

1.6.4 Effects of Synonymous Rate Variation

As previously mentioned, synonymous substitutions are often considered silent as they were thought to have no impact on selection. However, there is growing evidence to suggest that this is not the case. One of the first articles to show evidence of selection working on silent sites predates codon models and the constant synonymous rate assumption [Shields et al. 1988]. Shields et al. [1988] found that genes with higher G+C content at synonymous sites were more highly expressed and suggested that selection acts among synonymous codons in Drosophila. Another study of Drosophila found evidence of strong purifying selection on four-fold degenerate sites [Lawrie et al. 2013]. However, Drosophila are not the only organisms that experience selection on synonymous sites. There is evidence of purifying selection acting on the BRCA1 gene [Hurst & Pál 2001] in humans. Broader mammalian genes have also shown evidence of synonymous selection on exonic splicing enhancers [Schattner & Diekhans 2006]. Several reviews dealing with the non-silent nature of synonymous sites and their impact on selection and potential downstream impacts have been published [Chamary et al. 2006; Plotkin & Kudla 2011; Sauna & Kimchi-Sarfaty 2011; Sharp et al. 1995].

While it has been shown that site-to-site synonymous rate variation (SRV) does, in fact, exist there is still much not known about its impacts. The biological mechanisms behind it remain a mystery with a few common suggestions gaining popularity in the literature. Codon bias is one such theory and suggests that synonymous rates vary from site-to-site because certain tRNAs are favored over others [Plotkin & Kudla 2011]. The concept of optimal or favored codons was first presented by Ikemura [Ikemura 1981] who saw that genes in *E. coli* tended to use codons that corresponded to major tRNAs. Thus by using favored tRNAs, the translation process runs normally while using non-favored tRNAs slows the process and can have a potential downstream impact on protein folding for example. In Shields et al. [1988] paper, high G+C content is linked to the preferred codon usage. Another paper investigating the impact of silent substitutions in voltage-gated ion channels genes found that optimal codons were used in structurally and functionally important regions. They suggested this was to avoid a loss of function or potential protein misfolding [Zhou et al. 2012]. Optimal codon bias has also been suggested in relation to rare synonymous mutations associated

with autism spectrum disorders [Poliakov et al. 2014]. In bacteria, specifically salmonella, a study has shown that selection is acting on synonymous codon choice in highly expressed genes [Brandis & Hughes 2016]. Thus, evidence for biased codon usage has been found in a wide range of organisms and functional genes.

Another potential mechanism by which synonymous mutations may create downstream effects is via mRNA stability. One such case is found in the human *DRD2* gene [Duan et al. 2003]. They found evidence that synonymous substitutions had a negative impact on the mRNA stability of the gene and led to lower translational efficiency and synthesis of the receptor. Other studies have shown a possible relationship between selection acting on synonymous sites and a change in mRNA stability [Parmley & Hurst 2007; Resch et al. 2007].

Potential constraints on regulatory elements, such as splicing sites and transcription factor binding sites, present yet another mechanism through which synonymous sites may act through selection [Chamary et al. 2006; Hurst & Pál 2001]. Synonymous substitutions have been shown to impact splicing in several ways including potentially creating new sites, or impacting existing splicing sites like enhancers or silencers. These disruptions have even been linked to diseases [Chamary & Hurst 2005; Sauna & Kimchi-Sarfaty 2011]. Evidence also shows that synonymous substitutions in transcription factor binding sites can disrupt binding [Stergachis et al. 2013]. This can have potentially deleterious effects as it impacts translational efficiency.

The impact of synonymous rate variation on statistical inferences commonly used in molecular evolution also warrants more investigation. This area is less well investigated than the potential mechanism or fitness impacts of synonymous rate variation. Like many parameters of substitution models, synonymous rate estimates themselves can be impacted by assumptions about the underlying biology of molecular evolution [eg: Wakeley 1994]. It has been previously suggested that relying on the synonymous substitution rate as a proxy for the overall mutation rate may not be in the best interest [Shields et al. 1988].

As previously touched on, codon models of sequence evolution as they were originally introduced assumed a constant synonymous rate (α) across sites. However, with the growing evidence

that synonymous rate variation is biologically relevant, a few models that allow α to vary across sites have been introduced (see Section 1.5.2). The first model to do so was introduced in 2005 by Kosakovsky Pond & Muse. The paper presented several models but here we will focus on what they called the "Dual" and "Nonsynonymous" models. The Nonsynonymous model is the M5 parameterization of the codon model found in Yang et al. 2000, with varying nonsynonymous rates from site to site and a constant synonymous site rate. The Dual model is an MG94 x GTR model with independent distributions describing the synonymous and nonsynonymous rates. In this paper, they found that by comparing the Dual model to the Nonsynonymous model inferences did in fact change. They saw instances where sites that showed evidence of being under selection according to the Nonsynonymous model no longer did so when the Dual model was applied and vice versa. Additionally, the Dual model was found to be a better fit for the majority of the data sets. This indicates that the impacts of including SRV in models of sequence evolution should be investigated further.

Two years later, Mayrose et al 2007 introduced another method of accounting for synonymous rate variation. The model in that paper accounts not only for synonymous rate variation from site to site but accounts for adjacent site dependencies in an attempt to decrease the chance of inaccurately estimating either rate. In order to do this they propose a model that uses two Hidden Markov models (HMM) to model site-to-site variation. One HMM models the site-to-site dependencies of the nonsynonymous substitution rate while the other models the site-to-site dependencies of adjacent synonymous substitution rates. They find that by accounting for both dependencies and synonymous rate variation their model provides more accurate estimates of positive selection.

Thus we see that synonymous rate variation can be accommodated into sequence evolution models. We also see that there is some evidence of its impact on inferences of selection [Davydov et al. 2018; Kosakovsky Pond & Muse 2005; Mayrose et al. 2007a]. However, the widespread impact of doing so still remains to be seen.

1.7 Description of Data Sets Used in this Thesis

In this section, I will give an overview of the data sets that I used throughout my dissertation. I will describe the makeup of the mitochondrial DNA data sets and the Selectome data sets. Specifically, I will focus on the similarities and differences we see when these data sets are analyzed.

1.7.1 General Composition

The first data used in this dissertation is what we will refer to as the mitochondrial data set. This data set was originally derived from NCBI's GenBank database of mitochondrial sequences by Dr. Rachel Marceau West. The data set is comprised of 56 vertebrate and invertebrate Metazoan orders containing 5 to 25 representative species for each order. Orders with less than five represented species were excluded as they would not provide enough statistical power. See figure 1.7 for a tree representing the relationships between the orders. For each order, the sequences from the 13 protein-coding mitochondrial genes were used, although some orders were missing certain genes. Evidence of duplication was seen for some genes in certain species as well and these duplicate genes were retained in the alignments. Once the raw sequences were compiled they were aligned by gene and order using Mesquite v.2.74 [Maddison & Maddison 2007] and trees were generated using Mr.Bayes v.3.2.1 [Ronquist et al. 2012]. This resulted in a collection of 721 gene order combinations or alignments.

The second data set we use throughout this dissertation is derived from the Selectome database of Moretti et al. [2014]. Selectome is a database of positive selection constructed by using CODEML [Yang 2007] to detect positive selection across branches of gene trees. Specifically, we use a subset of 13,311 gene trees from the Euteleostomi set. Alignments of fewer than six sequences were excluded as they would decrease the statistical power, similar to the five species cut off for the mitochondrial data set. The details of filtering and multiple sequence alignment for the Selectome database can be found at: https://selectome.unil.ch/cgi-bin/methods.cgi. We choose to use Selectome because it is already being used for studies of positive selection inferences [e.g. Murrell et al. 2015; Rallapalli et al.

2014; Roux et al. 2014].

The first obvious difference between these two data sets is the total number of alignments in each. Selectome is over 18 times larger than the mitochondrial data set. We see in figure 1.8 that the mitochondrial data set typically has fewer sequences per alignment than the Selectome data set (1.8a). We also see that the mitochondrial data set has less variation in the number of codons per sequence than the Selectome data set but that the majority of the alignments are around the same length for both (1.8b). The mitochondrial data set also represents a total of 869 species while the Selectome data set represents 54 species.

1.7.2 Comparison of Analysis on the Data Sets

We analyzed both data sets using BUSTED [Murrell et al. 2015] and our new method, BUSTED+SRV, introduced in Chapter 4. For both BUSTED and BUSTED+SRV the number of nonsynonymous rate categories used is $k = 3$ and for BUSTED+SRV the number of synonymous rate categories is also $k = 3$. Here we provide a comparison of what we saw from the results of that analysis.

In figure 1.9 histograms for the estimated ω_3 are given for the mitochondrial and Selectome data sets according to both BUSTED and BUSTED+SRV. Note that both analyses are set so that $\omega_3 \geq 1$. We see that for both analyses the ranges of the estimated ω_3 values are similar for the mitochondrial and Selectome data sets. This suggests similar amounts of selective pressure may be present in the data sets. It does look like the ω_3 estimates are more similar between the two data sets for BUSTED+SRV. The median ω_3 according to BUSTED for the mitochondrial data set is at $\omega_3 = 1.99$ while for Selectome it is $\omega_3 = 2.10$. Similarly, the median ω_3 according to BUSTED+SRV for the mitochondrial data set is at $\omega_3 = 1.59$ while for Selectome it is $\omega_3 = 1.77$. For both analyses the range of the estimated ω_3 is similar.

For both data sets the coefficient of variation of the synonymous substitution rate (CV of SRV) is estimated using BUSTED+SRV. Since the CV is $\frac{\sigma}{\mu}$, it tells us the amount of variation relative to the mean. This means that the CV is useful for comparing the amount of variation between separate distributions which is why we use it here. The histogram in figure 1.10 shows the distribution of the

estimated CV of SRV. We see that for the majority of the alignments in both data sets the estimated CV of SRV lies between 0.50 and 1.30. For the mitochondrial data sets the CV of SRV estimates are centered 1.00 and the maximum estimated CV of SRV is 6.39. For the Selectome data set estimates center around 0.64 but the maximum CV of SRV estimated is 21.29. These numbers give us a good reference for what is realistic for empirical data sets.

1.8 Overview

In the following chapters, I provide evidence of site-to-site synonymous rate variation as a widespread phenomenon and delve into its potential impacts on methods of statistical inferences commonly used in molecular evolution. Chapter 2 details a study of Metazoan mitochondrial DNA. In this chapter, we see that at the codon level the site-to-site synonymous rate variation (SRV) is of a comparable magnitude to that of the nonsynonymous rate variation. Importantly, we note that the SRV is not restricted to a single mitochondrial gene or a single order and is a widespread phenomenon.

Chapter 3 is an extension of work done by Dr. Frank Mannino in his 2005 thesis. He originally derived an upper bound for the variation of a discretized rate distribution. In Chapter 3, we show how this bound impacts analysis of empirical data and how it may impact interpretations of the analyses. The potential bias of upper bound on variation is of particular interest given our work in Chapter 2 relies on accurate estimations of the synonymous and nonsynonymous rate variation.

Our results in Chapters 2 and 3 indicate that SRV and its impacts on statistical inferences require further study. Therefore, in Chapter 4 we develop a new method for detecting selection and use it to test the effect of including SRV on both empirical and simulated data. The simulation results are of particular interest as they reveal a high false positive rate for BUSTED when it fails to accommodate SRV.

1.9 Figures

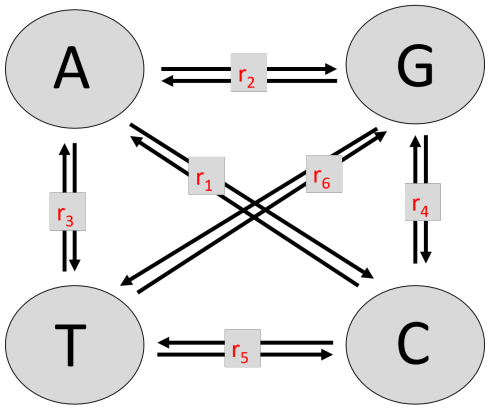


Figure 1.1 General Time Reversible Model. Four state, nucleotide, continuous time reversible Markov Chain. Arrows represent changes from one state to the next. The variable r_i on each arrow represents the rate of substitution from one state to the next.

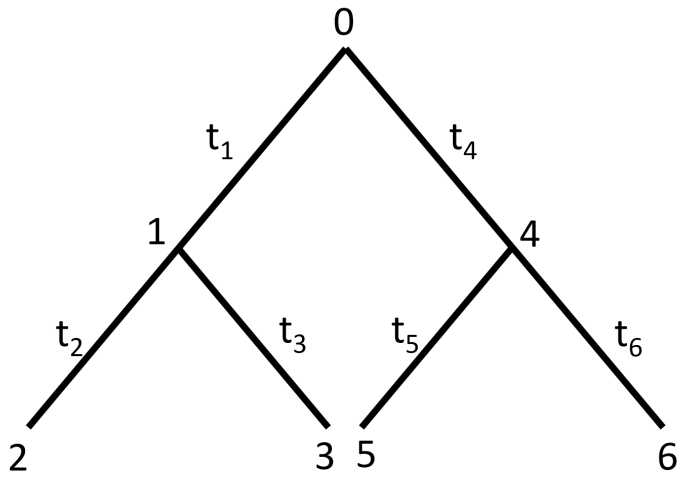


Figure 1.2 Simple, Rooted Tree. A simple rooted tree with the root at node 0 of four sequences.

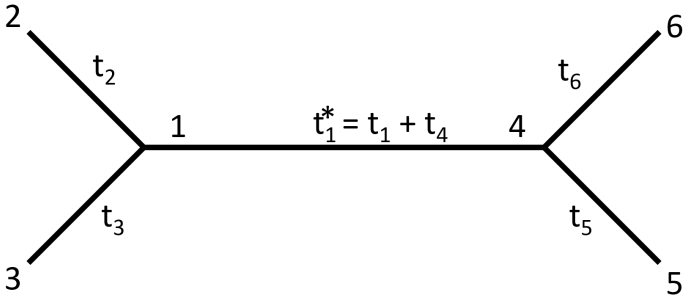


Figure 1.3 Unrooted Tree. Unrooted tree with 4 descendant sequences.

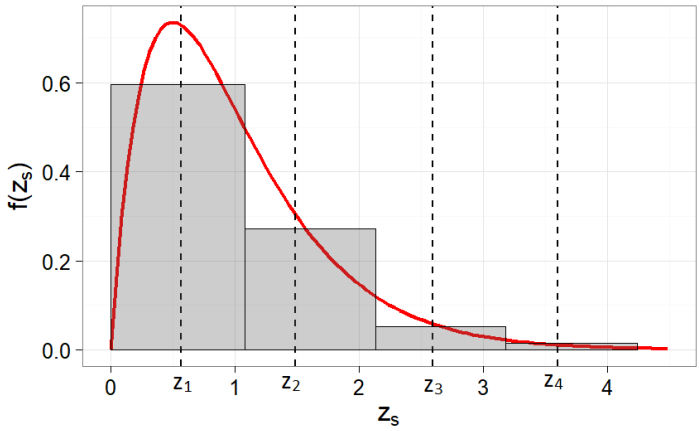


Figure 1.4 Example of Discrete Distribution. Gamma distribution with $\lambda_\alpha = 2$ and $k = 4$ represented by the red line. The histogram represents the 4 rate categories. Dashed lines represent the mean of each category, z_k.

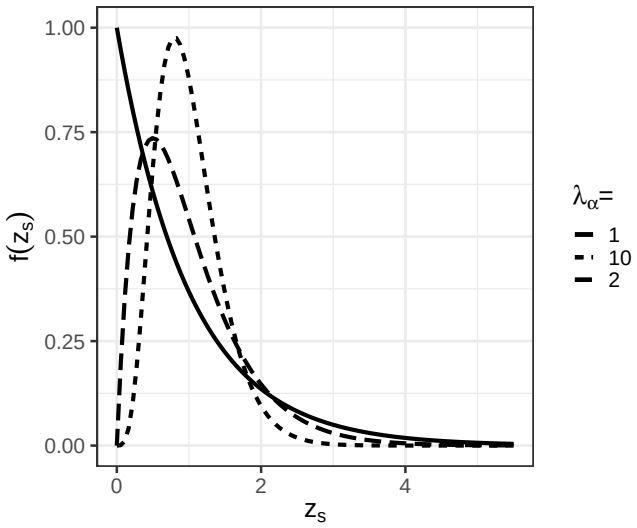


Figure 1.5 Varying Shape Parameters of the Gamma Distribution. Lines represent the continuous gamma distribution with $\lambda_\alpha = 1$, $\lambda_\alpha = 2$, $\lambda_\alpha = 10$ where the scale parameter equals the shape parameter.

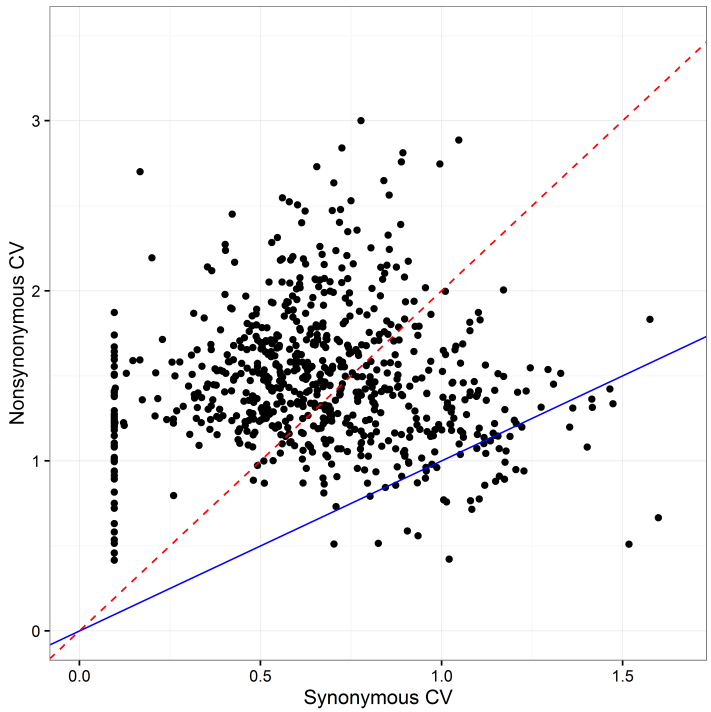


Figure 1.6 Comparison of Synonymous and Nonsynonymous Coefficients of Variation (CV). For each of the 721 datasets, we plot its estimated synonymous and nonsynonymous CV. Points below the blue line are datasets where the synonymous CV exceeds that of nonsynonymous CV; points below the red line had a synonymous CV of at least half the nonsynonymous CV. The vertical line of points on the left represents a numerical artifact for datasets with synonymous CV effectively zero.

Table 1.1 Branch-Site Rate Categories. A table describing the combination of foreground and background branch-site categories and their proportions.

Class	Background	Foreground	Proportion
1	ω_1	ω_1	p_1
2	ω_2	ω_2	p_2
3	ω_1	ω_3	$p_3 = (1-p_1-p_2)p_1/(p_1+p_2)$
4	ω_2	ω_3	$p_4 = (1-p_1-p_2)p_2/(p_1+p_2)$

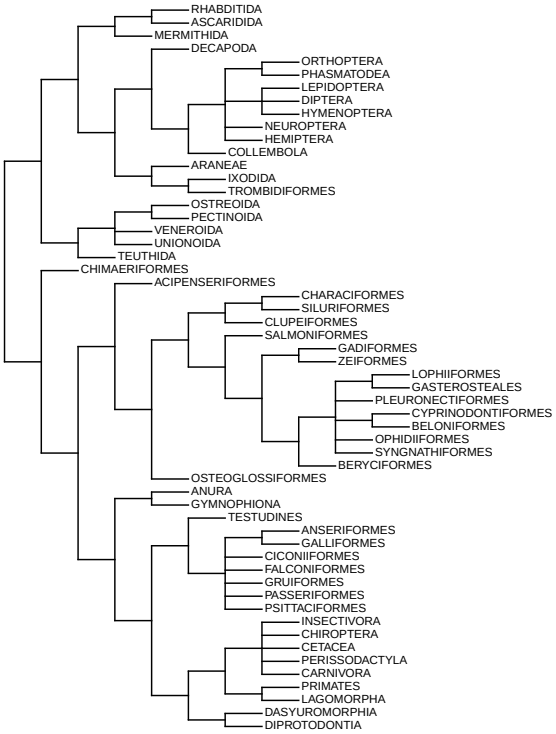
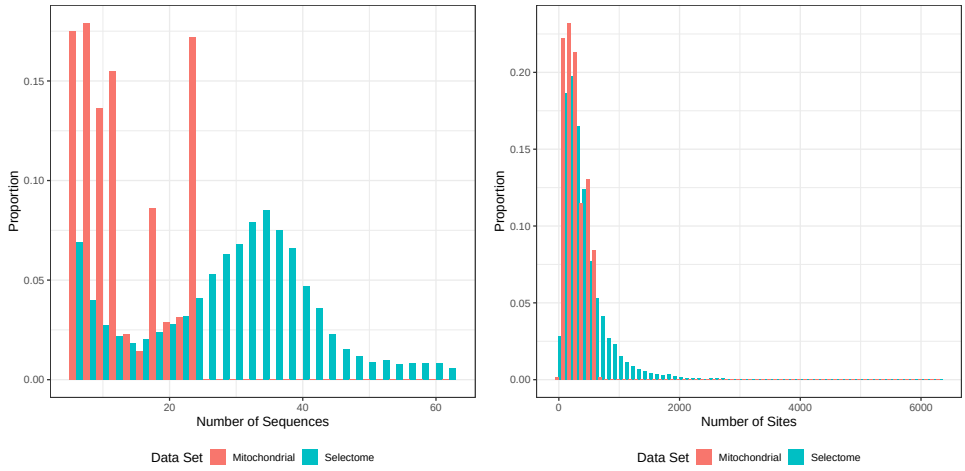


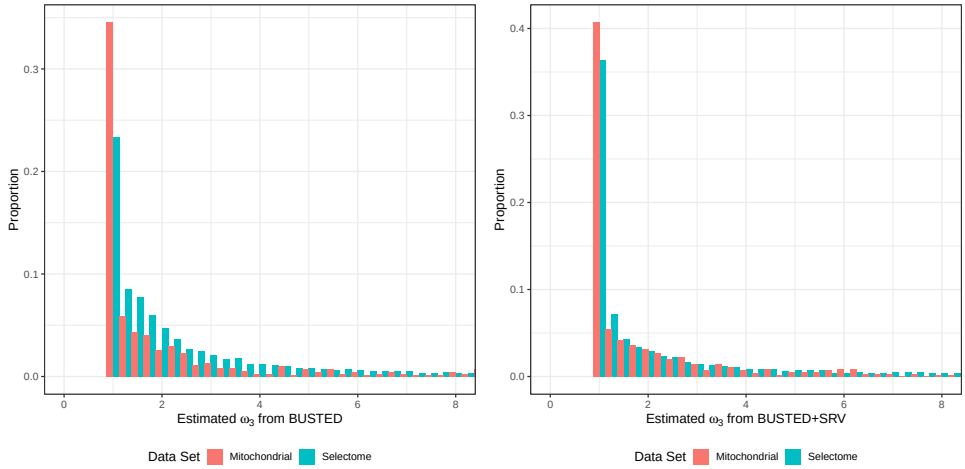
Figure 1.7 Tree of Metazoan Order for the Mitochondrial Data Set. This tree of the 56 orders comprised into the mitochondrial data set was generated using NCBI's taxonomy common tree tool.



(a) Histogram of the Number of Sequences per Alignment.

(b) Histogram of the Number of Sites per Alignment.

Figure 1.8 Range of Sites and Sequences for Data Sets. Histograms describing the number of sequences (a) and sites (b) per alignment. Histograms are split for the Mitochodrial (red) and Selectome (blue) data sets.



(a) Histogram of the ω_3 Estimate from BUSTED.

(b) Histogram of the ω_3 Estimate from BUSTED+SRV.

Figure 1.9 Range of Estimated ω_3 of Data Sets. Histograms describing the estimated ω_3 for the Mitochodrial (red) and Selectome (blue) data sets according to BUSTED (a) and BUSTED+SRV (b).

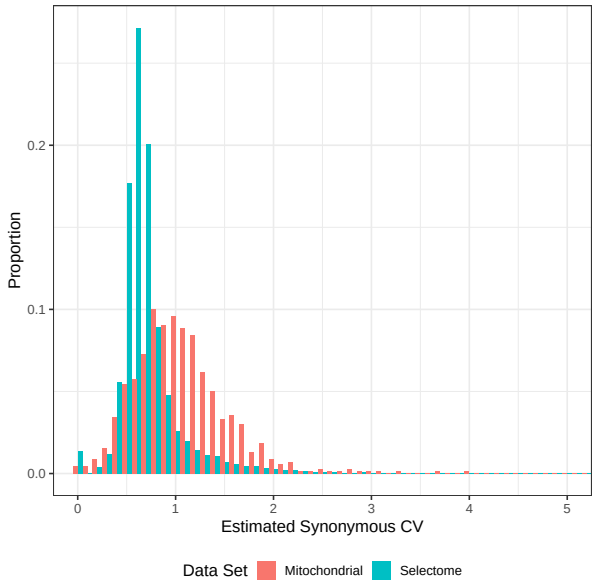


Figure 1.10 Histogram of the Estimated Synonymous Coefficient of Variation (CV). The distribution of estimated synonymous CVs according to BUSTED+SRV for the Selectome (blue) and mitochondrial (red) data sets.

BIBLIOGRAPHY

Anisimova, M. & Kosiol, C. (2009). "Investigating Protein-Coding Sequence Evolution with Probabilistic Codon Substitution Models". *Molecular Biology and Evolution* **26**.2, pp. 255–271.

Anisimova, M. & Yang, Z. (2007). "Multiple Hypothesis Testing to Detect Lineages under Positive Selection that Affects Only a Few Sites". *Molecular Biology and Evolution* **24**.5, pp. 1219–1228.

Arenas, M. (2015). "Trends in substitution models of molecular evolution". *Frontiers in Genetics* **6**.OCT, p. 319.

Bielawski, J. & Yang, Z. (2004). "A Maximum Likelihood Method for Detecting Functional Divergence at Individual Codon Sites, with Application to Gene Family Evolution". *Journal of Molecular Evolution* **59**.1, pp. 121–132.

Brandis, G. & Hughes, D. (2016). "The Selective Advantage of Synonymous Codon Usage Bias in Salmonella". *PLoS Genetics* **12**.3. Ed. by Ibba, M., e1005926.

Brown, W. M. et al. (1982). "Mitochondrial DNA sequences of primates: Tempo and mode of evolution". *Journal of Molecular Evolution* **18**.4, pp. 225–239.

Chamary, J. V. & Hurst, L. D. (2005). "Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals". en. *Genome Biology* **6**.9, R75.

Chamary, J. V. et al. (2006). "Hearing silence: non-neutral evolution at synonymous sites in mammals". en. *Nature Reviews Genetics* **7**.2, pp. 98–108.

Crandall, K. A. & Hillis, D. M. (1997). "Rhodopsin evolution in the dark". *Nature* **387**.6634, pp. 667–668.

Davydov, I. I. et al. (2018). "Modeling Codon Rate Variation: Robust Inference of Protein and Nucleotide Selection". *bioRxiv*, p. 174839.

Dayhoff, M. et al. (1978). "A model of evolutionary change in proteins". *Atlas of protein sequence and structure*, pp. 345–352.

Delport, W. et al. (2008). "Models of coding sequence evolution". *Briefings in Bioinformatics* **10**.1, pp. 97–109.

Doron-Faigenboim, A. & Pupko, T. (2006). “A Combined Empirical and Mechanistic Codon Model”. *Molecular Biology and Evolution* **24**.2, pp. 388–397.

Duan, J. et al. (2003). “Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor”. en. *Human Molecular Genetics* **12**.3, pp. 205–216.

Dutheil, J. Y. et al. (2012). “Efficient Selection of Branch-Specific Models of Sequence Evolution”. *Molecular Biology and Evolution* **29**.7, pp. 1861–1874.

Echave, J. et al. (2016). “Causes of evolutionary rate variation among protein sites”. *Nature Reviews Genetics* **17**.2, pp. 109–121.

Enard, D. et al. (2016). “Viruses are a dominant driver of protein adaptation in mammals”.

Felsenstein, J. & Churchill, G. A. (1996). “A Hidden Markov Model approach to variation among sites in rate of evolution”. *Molecular Biology and Evolution* **13**.1, pp. 93–104.

Felsenstein, J. (1981). “Evolutionary trees from DNA sequences: A maximum likelihood approach”. *Journal of Molecular Evolution* **17**.6, pp. 368–376.

— (1984). “DISTANCE METHODS FOR INFERRING PHYLOGENIES: A JUSTIFICATION”. *Evolution* **38**.1, pp. 16–24.

— (2001). “Taking variation of evolutionary rates between sites into account in inferring phylogenies”. *Journal of Molecular Evolution* **53**.4-5, pp. 447–455.

Fitch, W. M. & Margoliash, E. (1967). “A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case”. *Biochemical Genetics* **1**.1, pp. 65–71.

Fitch, W. M. & Markowitz, E. (1970). “An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution”. *Biochemical Genetics* **4**.5, pp. 579–593.

Fletcher, W. & Yang, Z. (2010). “The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection”. *Molecular Biology and Evolution* **27**.10, pp. 2257–2267.

Forsberg, R. & Christiansen, F. B. (2003). “A Codon-Based Model of Host-Specific Selection in Parasites, with an Application to the Influenza A Virus”. *Molecular Biology and Evolution* **20**.8, pp. 1252–1259.

Gaut, B. S. et al. (1996). “Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL.” *Proceedings of the National Academy of Sciences of the United States of America* **93**.19, pp. 10274–9.

Golding, G. B. (1984). “Estimates of DNA and protein sequence divergence: an examination of some assumptions.” *Molecular Biology and Evolution* **1**.1, pp. 125–142.

Goldman, N & Yang, Z (1994). “A codon-based model of nucleotide substitution for protein-coding DNA sequences.” *Molecular biology and evolution* **11**.5, pp. 725–736.

Hanada, K. et al. (2004). “A Large Variation in the Rates of Synonymous Substitution for RNA Viruses and Its Relationship to a Diversity of Viral Infection and Transmission Modes”. *Molecular Biology and Evolution* **21**.6, pp. 1074–1080.

Hasegawa, M. et al. (1985). “Dating of the human-ape splitting by a molecular clock of mitochondrial DNA”. *Journal of Molecular Evolution* **22**.2, pp. 160–174.

Hodgkinson, A. & Eyre-Walker, A. (2011). “Variation in the mutation rate across mammalian genomes”. *Nature Reviews Genetics* **12**.11, pp. 756–766.

Hoehn, K. B. et al. (2017). “A Phylogenetic Codon Substitution Model for Antibody Lineages.” *Genetics* **206**.1, pp. 417–427.

Hughes, A. L. & Nei, M. (1988). “Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection”. *Nature* **335**.6186, pp. 167–170.

Hurst, L. D. & Pál, C. (2001). “Evidence for purifying selection acting on silent sites in BRCA1”. *Trends in Genetics* **17**.2, pp. 62–65.

Ikemura, T (1981). “Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system.” *Journal of molecular biology* **151**.3, pp. 389–409.

Jones, D. T. et al. (1992). “The rapid generation of mutation data matrices from protein sequences”. *Bioinformatics* **8**.3, pp. 275–282.

Jukes, T. H. & Cantor, C. R. (1969). “Evolution of protein molecules”. *Mammalian Protein Metabolism*, pp. 21–123.

Kimura, M. (1980). “A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences”. *Journal of Molecular Evolution* **16**.2, pp. 111–120.

Kosakovsky Pond, S. L. & Frost, S. D. W. (2005). “A Simple Hierarchical Approach to Modeling Distributions of Substitution Rates”. *Molecular Biology and Evolution* **22**.2, pp. 223–234.

Kosakovsky Pond, S. L. & Muse, S. V. (2005). “Site-to-site variation of synonymous substitution rates.” en. *Molecular Biology and Evolution* **22**.12, pp. 2375–2385.

Kosakovsky Pond, S. L. et al. (2005). “Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection”. en. *Molecular Biology and Evolution* **22**.5, pp. 1208–1222.

Kosakovsky Pond, S. L. et al. (2011). “A random effects branch-site model for detecting episodic diversifying selection”. *Molecular Biology and Evolution* **28**.11, pp. 3033–3043.

Kosiol, C. et al. (2007). “An Empirical Codon Model for Protein Sequence Evolution”. *Molecular Biology and Evolution* **24**.7, pp. 1464–1479.

Lawrie, D. S. et al. (2013). “Strong Purifying Selection at Synonymous Sites in D. melanogaster”. *PLoS Genetics* **9**.5. Ed. by Plotkin, J. B., e1003527.

Li, W. H. et al. (1985). “A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes.” *Molecular biology and evolution* **2**.2, pp. 150–74.

Liberles, D. A. et al. (2012). *The interface of protein structure, protein biophysics, and molecular evolution*.

Lorenzo-Redondo, R. et al. (2016). “Persistent HIV-1 replication maintains the tissue reservoir during therapy”. *Nature* **530**.7588, pp. 51–56.

Maddison, W. P. & Maddison, D. (2007). *Mesquite: a modular system for evolutionary analysis.*

Mayrose, I. et al. (2007a). “Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates”. *Bioinformatics* **23**.13, pp. i319–i327.

Mayrose, I. et al. (2005). “A Gamma mixture model better accounts for among site rate heterogeneity”. *Bioinformatics* **21**.Suppl 2, pp. ii151–ii158.

Mayrose, I. et al. (2007b). “Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates”. *Bioinformatics* **23**.13, pp. i319–i327.

Messier, W. & Stewart, C.-B. (1997). “Episodic adaptive evolution of primate lysozymes”. *Nature* **385**.6612, pp. 151–154.

Miyata, T. & Yasunaga, T. (1980a). “Journal of Molecular Evolution Molecular Evolution of mRNA: A Method for Estimating Evolutionary Rates of Synonymous and Amino Acid Substitutions from Homologous Nucleotide Sequences and Its Application”. *J. Mol. Evol* **16**.2, pp. 3–3.

— (1980b). “Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application”. *Journal of Molecular Evolution* **16**.1, pp. 23–36.

Moretti, S. et al. (2014). “Selectome update: quality control and computational improvements to a database of positive selection.” en. *Nucleic Acids Research* **42**.Database issue, pp. 917–21.

Mouchiroud, D. et al. (1995). “Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions”. *Journal of Molecular Evolution* **40**.1, pp. 107–113.

Murrell, B. et al. (2015). “Gene-Wide Identification of Episodic Selection”. en. *Molecular Biology and Evolution* **32**.5, pp. 1365–1371.

Muse, S. V. (1996). “Estimating synonymous and nonsynonymous substitution rates.” en. *Molecular Biology and Evolution* **13**.1, pp. 105–114.

Muse, S. V. & Gaut, B. S. (1994). “A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome.” *Molecular Biology and Evolution* **11**.5, pp. 715–724.

Nei, M & Gojobori, T (1986). “Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.” *Molecular Biology and Evolution* **3**.5, pp. 418–426.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Vol. 17, p. 512.

Nielsen, R & Yang, Z (1998). “Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.” *Genetics* **148**.3, pp. 929–36.

Nielsen, R. et al. (2006). “Maximum Likelihood Estimation of Ancestral Codon Usage Bias Parameters in Drosophila”. *Molecular Biology and Evolution* **24**.1, pp. 228–235.

Ohta, T. & Ina, Y. (1995). “Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences”. *Journal of Molecular Evolution* **41**.6.

Pacheco, M. A. et al. (2018). “Mode and Rate of Evolution of Haemosporidian Mitochondrial Genomes: Timing the Radiation of Avian Parasites”. *Molecular Biology and Evolution* **35**.2, pp. 383–403.

Parmley, J. L. & Hurst, L. D. (2007). “How do synonymous mutations affect fitness?” en. *BioEssays* **29**.6, pp. 515–519.

Plotkin, J. B. & Kudla, G. (2011). “Synonymous but not the same: the causes and consequences of codon bias”. en. *Nature Reviews Genetics* **12**.1, pp. 32–42.

Poliakov, E. et al. (2014). “Impairment of translation in neurons as a putative causative factor for autism.” *Biology direct* **9**.1, p. 16.

Rallapalli, P. M. et al. (2014). “Positive Selection during the Evolution of the Blood Coagulation Factors in the Context of Their Disease-Causing Mutations”. *Molecular Biology and Evolution* **31**.11, pp. 3040–3056.

Resch, A. M. et al. (2007). “Widespread Positive Selection in Synonymous Sites of Mammalian Genes”. *Molecular Biology and Evolution* **24**.8, pp. 1821–1831.

Robinson, D. M. et al. (2003). “Protein Evolution with Dependence Among Codons Due to Tertiary Structure”. *Molecular Biology and Evolution* **20**.10, pp. 1692–1704.

Ronquist, F. et al. (2012). “MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space”. *Systematic Biology* **61**.3, pp. 539–542.

Roux, J. et al. (2014). “Patterns of Positive Selection in Seven Ant Genomes”. *Molecular Biology and Evolution* **31**.7, pp. 1661–1685.

Saitou, N & Nei, M (1987). “The neighbor-joining method: a new method for reconstructing phylogenetic trees.” *Molecular Biology and Evolution* **4**.4, pp. 406–425.

Sauna, Z. E. & Kimchi-Sarfaty, C. (2011). “Understanding the contribution of synonymous mutations to human disease”. en. *Nature Reviews Genetics* **12**.10, pp. 683–691.

Schattner, P. & Diekhans, M. (2006). “Regions of extreme synonymous codon selection in mammalian genes”. en. *Nucleic Acids Research* **34**.6, pp. 1700–1710.

Schneider, A. et al. (2005). “Empirical codon substitution matrix”. *BMC Bioinformatics* **6**.1, p. 134.

Schöniger, M. et al. (1990). “Stochastic traits of molecular evolutionâĂŤAcceptance of point mutations in native actin genes”. *Journal of Theoretical Biology* **143**.3, pp. 287–306.

Sharp, P. M. et al. (1995). “DNA sequence evolution: the sounds of silence.” *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **349**.1329, pp. 241–247.

Shields, D. C. et al. (1988). “"Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons.” *Molecular biology and evolution* **5**.6, pp. 704–716.

Shoemaker, J. S. & Fitch, W. M. (1989). “Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated.” *Molecular Biology and Evolution* **6**.3, pp. 270–289.

Stergachis, A. B. et al. (2013). “Exonic transcription factor binding directs codon choice and affects protein evolution.” *Science (New York, N.Y.)* **342**.6164, pp. 1367–72.

Swanson, W. J. et al. (2003). “Pervasive Adaptive Evolution in Mammalian Fertilization Proteins”. *Molecular Biology and Evolution* **20**.1, pp. 18–20.

Tamura, K & Nei, M (1993). “Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.” *Molecular Biology and Evolution* **10**.3, pp. 512–526.

Tavaré, S (1986). *Some probabilistic and statistical problems in the analysis of DNA sequences*.

Waddell, P. J. & Steel, M. (1997). "General Time-Reversible Distances with Unequal Rates across Sites: Mixing Γ and Inverse Gaussian Distributions with Invariant Sites". *Molecular Phylogenetics and Evolution* **8**.3, pp. 398–414.

Wakeley, J (1994). "Substitution-rate variation among sites and the estimation of transition bias." *Molecular biology and evolution* **11**.3, pp. 436–442.

Weadick, C. J. & Chang, B. S. (2012). "An Improved Likelihood Ratio Test for Detecting Site-Specific Functional Divergence among Clades of Protein-Coding Genes". *Molecular Biology and Evolution* **29**.5, pp. 1297–1300.

Whelan, S. et al. (2001). "Molecular phylogenetics: state-of-the-art methods for looking into the past". *Trends in Genetics* **17**.5, pp. 262–272.

Wolfe, K. H. et al. (1987). "Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs." *Proceedings of the National Academy of Sciences of the United States of America* **84**.24, pp. 9054–8.

Wolfe, K. H. et al. (1989). "Mutation rates differ among regions of the mammalian genome". *Nature* **337**.6204, pp. 283–285.

Wolfe, K. & Sharp, P. (1993). "Mammalian gene evolution: Nucleotide sequence divergence between mouse and rat". *Journal of Molecular Evolution* **37**.4, pp. 441–456.

Wong, W. S. W. et al. (2004). "Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites." *Genetics* **168**.2, pp. 1041–51.

Yang, Z. (1998). "Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution". *Molecular Biology and Evolution* **15**.5, pp. 568–573.

— (2007). "PAML 4: Phylogenetic Analysis by Maximum Likelihood". *Molecular Biology and Evolution* **24**.8, pp. 1586–1591.

Yang, Z. (1994a). "Estimating the pattern of nucleotide substitution". *Journal of Molecular Evolution* **39**.1, pp. 105–111.

Yang, Z. (1994b). “Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods”. *Journal of Molecular Evolution* **39**.3, pp. 306–314.

— (1996). “Among-site rate variation and its impact on phylogenetic analyses”. *Trends in Ecology & Evolution* **11**.9, pp. 367–372.

Yang, Z. & Nielsen, R. (1998). “Synonymous and nonsynonymous rate variation in nuclear genes of mammals”. *Journal of Molecular Evolution* **46**.4, pp. 409–418.

— (2002). “Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages.” en. *Molecular biology and evolution* **19**.6, pp. 908–17.

— (2008). “Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage”. *Molecular Biology and Evolution* **25**.3, pp. 568–579.

Yang, Z. & Rannala, B. (2012). “Molecular phylogenetics: principles and practice”. *Nature Reviews Genetics* **13**.5, pp. 303–314.

Yang, Z. et al. (2000). “Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites”. *Molecular biology and evolution* **19**.1, pp. 49–57.

Zhang, C. et al. (2011). “Dynamic programming procedure for searching optimal models to estimate substitution rates based on the maximum-likelihood method.” *Proceedings of the National Academy of Sciences of the United States of America* **108**.19, pp. 7860–5.

Zhang, J. (2004). “Frequent False Detection of Positive Selection by the Likelihood Method with Branch-Site Models”. *Molecular Biology and Evolution* **21**.7, pp. 1332–1339.

Zhang, J. et al. (2005). “Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level.” en. *Molecular biology and evolution* **22**.12, pp. 2472–9.

Zhou, T. et al. (2012). “Non-Silent Story on Synonymous Sites in Voltage-Gated Ion Channel Genes”. *PLoS ONE* **7**.10. Ed. by Uversky, V. N., e48541.

Zhu, A. et al. (2014). “Unprecedented heterogeneity in the synonymous substitution rate within a plant genome.” en. *Molecular biology and evolution* **31**.5, pp. 1228–36.

CHAPTER

2

WIDESPREAD SITE-TO-SITE SYNONYMOUS SUBSTITUTION RATE VARIATION

2.1 Authors

Sadie Wisotsky; Bioinformatics Research Center, North Carolina State University, Rachel Marceau West; Department of Statistics, North Carolina State University, Spencer Muse; Bioinformatics Research Center & Department of Statistics, North Carolina State University

2.2 Contribution

The following article, Widespread Site-to-Site Synonymous Substitution Rate Variation, is currently out for review. The aim of this paper was to further investigations of how widespread the phenomenon of synonymous rate variation is. We were also interested in quantifying the amount of SRV present. For this paper, I ran the analysis described within, interpreted the results, created the figures and wrote the paper. Rachel Marceau West assembled and cleaned the data set. Presented here is the paper as it is submitted with additional content.

2.3 Abstract

We investigate the presence and magnitude of site-to-site variability of synonymous substitution rates in mitochondrial DNA (mtDNA) using a codon-based model of sequence evolution. Analysis of 13 protein-coding genes from the complete mitochondrial genomes of 869 species representing 56 metazoan orders reveals extensive synonymous substitution rate variability (SRV). Over half of the analyzed datasets showed statistically significant levels of SRV. Furthermore, the magnitude of SRV is comparable to that of nonsynonymous rates, including 43 datasets where there is more synonymous rate heterogeneity than nonsynonymous heterogeneity. These findings raise interesting questions about underlying mechanisms of selection at silent sites and have important implications for data analysis methods, especially those used for studies of adaptive evolution based on the assumption that silent rates are equal across all sites in a gene.

2.4 Introduction

The use of substitution models to identify and quantify sources of rate heterogeneity is key to our understanding of molecular evolution. In order to be computationally tractable and statistically powerful enough to make inferences, these models must make assumptions regarding the underlying biological process of sequence change. Thus, there is a trade-off between accurately representing

the biological complexity of the process and managing the computational complexity. Since the introduction of the earliest model [Jukes & Cantor 1969] several decades ago, researchers continue to refine these assumptions to more accurately reflect underlying biological processes and to take advantage of more powerful computational hardware now available.

One persistent assumption is that the synonymous substitution rate in many models of sequence evolution is constant across all sites in a gene [eg: Goldman & Yang 1994; Mayrose et al. 2007; Muse & Gaut 1994; Nielsen & Yang 1998; Parto & Lartillot 2017; Yang & Nielsen 2008]. Synonymous substitutions, often referred to as "silent", do not change the resulting amino acid and are thought to have no effect on fitness, rendering them invisible to natural selection. However, a growing number of studies have shown selection acting on synonymous substitution sites [eg: Goymer 2007; Hurst & Pál 2001; Parmley et al. 2006; Schattner & Diekhans 2006; Singh et al. 2007] challenging this historic belief. Furthermore, it has been demonstrated that biological mechanisms are impacted by synonymous changes. Synonymous substitutions have been shown to impact the stability of mRNA [Duan et al. 2003; Hurst & Pál 2001], affect the downstream translation of proteins [Zhou et al. 2012], and potentially disrupt splicing and cause exon skipping [eg: Iida & Akashi 2000; Pagani et al. 2005]. Synonymous substitutions have even been associated with several diseases in humans [Bhardwaj 2014; Chen & Shapiro 2015]. While potential mechanisms underlying selection of synonymous sites are not well understood, it is evident that not all synonymous substitutions are neutral. Their potential impacts should be considered when modeling sequence evolution.

Mitochondrial genomes are often used in evolutionary studies to reconstruct ancestry, barcode or label species, and search for sites influenced by selection [Cao et al. 1994; Hoelzer 1997; Parham et al. 2006; Peng et al. 2017; Yu et al. 2014; Zardoya & Meyer 1996]. Their lack of introns, lack of recombination, and maternal inheritance make them excellent tools for evolutionary studies. Their relatively small size (about 16 kb) and the limited number of protein-coding genes (12-13) makes metazoan mtDNA sequencing simple and cheap. Notably, the evolutionary rate of mitochondrial genomes is rapid in many lineages [Brown et al. 1979; Oliveira et al. 2008], making mitochondrial genes ideal tools for studies of relatively recent evolutionary events. With over 7000 complete

mitochondrial genome sequences available now on GenBank and more constantly being added, the widespread use of mitochondrial genomes is evident. Previous studies have shown that the D-loop region of human mtDNA experiences extreme site-to-site nonsynonymous rate heterogeneity [Endicott & Ho 2008; Excoffier & Yang 1999; Wakeley 1994] and that neglecting this among-site rate heterogeneity results in incorrect estimations for parameters ranging from time to the most recent common ancestor to age estimations. It was this lack of knowledge about SRV in metazoan mtDNA combined with the use of mtDNA in inferences that could be impacted by SRV that led to our choice of data sets.

Our study set out to identify and quantify site-to-site SRV in metazoan mtDNA genomes. We fit two models, one with SRV and one without, and compare the fits in order to determine if SRV is present in the data. We used the estimated coefficients of variation (CV) for both synonymous and nonsynonymous rates to quantify the amount of SRV. Our findings indicate that while the synonymous CV is typically lower than the nonsynonymous, it has a comparable magnitude to that of nonsynonymous rates. Additionally, we find that SRV is not restricted to a single gene or order but rather is a widespread phenomenon.

2.5 Results and Discussion

In the following sections, we addressed three main questions. First, we investigated how frequently we were able to detect evidence of SRV across sites. Then we needed to know the magnitude of SRV when it was present and how it compared to nonsynonymous rate variation. Finally, we examined how much levels of SRV varied across genes and lineages. For these studies, we assembled 721 datasets (alignments) from complete metazoan mitochondrial genomes in GenBank. Each dataset consists of the sequence alignment of a single protein-coding gene for 5 to 25 representatives of a metazoan order.

2.5.1 Prevalence of Synonymous Rate Variation

As seen in Figure 2.1, 412 of 721 (57%) alignments rejected the null hypothesis of no SRV across sites using a strict Bonferroni correction. All genes had multiple orders that reject the null at the 0.01 level. The widespread finding of lack of SRV suggests that SRV is common across sites in mitochondrial genes and in metazoan orders. It is important to note that virtually all widely used statistical methods assume that synonymous rates are equal across all sites within a gene.

Only two metazoan orders, Araneae and Venerodia, had no alignments that rejected the null. These orders each had fewer than ten species per alignment. Alignments with a lower number of species were less likely to reject the null than those that had ten or more species representing an order. The discrepancy between alignments with fewer sequences and alignments with more sequences may indicate an underlying lack of power when using smaller datasets.

2.5.2 Magnitude of Synonymous Rate Variation

The CV provides us with a measure of how much the synonymous and nonsynonymous rates varied across sites and allows us to compare the magnitude of that variation across multiple datasets. The majority of alignments (61%) had a synonymous CV less than half of the nonsynonymous CV (Fig. 2.2). However, 43 alignments (6%) had a synonymous CV greater than the nonsynonymous CV, and an additional 226 alignments (31%) had a synonymous CV at least half the respective nonsynonymous CV. All alignments had a nonzero synonymous CV, indicating that the synonymous rates did vary, however minutely, across sites in all the datasets. This directly contradicts the assumption made by most current models of no SRV. Additionally, the evidence showed the magnitude of SRV is not only comparable to but is sometimes greater than the variation of the nonsynonymous substitution rates.

2.5.3 Synonymous Rate Variation is Widespread

In order to determine if the occurrence of SRV was widespread, we looked at the distributions of the CV for both synonymous and nonsynonymous rates among genes and orders separately

(Figure 2.3 and 2.4). Figure 2.3 shows boxplots of estimated CV values across lineages for each mitochondrial gene. The level of SRV is appreciable for each of the 13 genes included, indicating that SRV is not restricted to a single gene and is also not absent from any genes. The same is true of Figure 2.4, which pictures boxplots of estimated CV ranges across each gene for each metazoan order. There is not a single order that lacks multiple genes with synonymous CVs greater than zero. This indicates that SRV is not restricted to a single order. Similarly to what we saw in Figure 2.2 as well, the nonsynonymous CV ranges are typically greater than the synonymous ranges for the same gene and order groupings, indicating that across genes and orders the nonsynonymous CV is typically greater than the synonymous CV. Also worth noting, when grouped by order (Figure 2.4.A) the ranges of the synonymous CV appear noisier than when grouped by gene.

2.6 Implications

The implications of widespread site-to-site SRV across genes and orders are twofold, with potential impacts regarding the underlying biological mechanisms of sequence evolution and the statistical inference made using models. As noted in the Introduction, there is mounting evidence from studies that selection does in fact act on synonymous sites. Several proposed mechanisms have been suggested to explain synonymous sites undergoing positive selection, including preferential tRNA usage [Ikemura 1981], induction of translational pausing [eg: Parmley & Hurst 2007], impacts on mRNA stability [eg: Chamary & Hurst 2005] and changes to splicing patterns [eg: Xing & Lee 2005]. However, there are still question about the underlying mechanisms that cause synonymous rates to vary across sites. In fact, there are a few studies that suggest it is likely some combination of mechanisms responsible for allowing synonymous substitutions to impact phenotypes [Agashe et al. 2016; Knöppel et al. 2016].

The evidence presented here indicates widespread synonymous rate variation across mtDNA of a comparable magnitude to that of the among-site nonsynonymous rate variation. Given that previous studies have shown that a number of inferences on mtDNA are sensitive to unaccounted for site-to-site rate variation [Endicott & Ho 2008; Excoffier & Yang 1999], it stands to reason the same

may be true for the impact of unaccounted for SRV. There is already evidence of misidentification of positive selection acting on sites when SRV is not accounted for, potentially due to the variation in the synonymous rate being attributed solely to nonsynonymous rate [Hurst & Pál 2001; Kosakovsky Pond & Frost 2005; Kosakovsky Pond & Muse 2005]. Other studies have also shown that with the addition of SRV to codon models, better fits and more accurate estimates of positive selection were achieved [Kosakovsky Pond et al. 2008; Kosakovsky Pond et al. 2006; Lemey et al. 2007; Mayrose et al. 2007; Ngandu et al. 2008]. Most methods for detecting positive selection could be readily modified to include SRV in their parameterizations and several have even mentioned it as a possibility in passing [eg: Kosakovsky Pond & Frost 2005; Murrell et al. 2015; Smith et al. 2015; Yang 1993; Zhang et al. 2005].

One study of positive selection on a wide array of protein-coding genes found evidence of SRV in 42% of the protein families investigated [Dimitrieva & Anisimova 2014], providing the first indication of how widespread the phenomenon was. Interestingly, the majority of the mitochondrial genes we examined are classified as having oxidoreductase activity. Genes with oxidoreducatase activity, according to the the Dimitrieva & Anisimova [2014] study were more likely than expected to show evidence of SRV. Here, we have shown SRV is present in the majority of the data sets (57%) tested, further proving SRV is widespread. SRV was not absent from a single order or gene, indicating that it is not restricted to a specific lineage or gene. Additionally, we have shown that the SRV present is of a substantial magnitude comparable to that of the nonsynonymous substitution rate variation. Therefore, our results show that SRV is common and of appreciable magnitude.

2.7 Materials and Methods

2.7.1 Sequences

To form our dataset, we first searched NCBI's GenBank ["GenBank"] for sequences with all mitochondrial protein-coding genes present. Then we separated those sequences into orders. By looking at the order level we ensured that datasets had enough diversity to estimate the substitution rates

while not having so much history that there would be a large variance of the estimates. Orders with fewer than 5 species were not considered as that would result in a lack of statistical power. Those with more than 25 species were pared down to 20 representative species in order to reduce computational costs. Typically, metazoan mitochondrial genomes contain 13 protein-coding genes: NADH dehydrogenase subunits 1-6 and 4L (ND1-ND6, ND4L), cytochrome c subunits I, II, and III (COX1-COX3), ATP synthase F0 subunits 6 and 8 (ATP6, ATP8), and cytochrome b (CYTB). While 50 orders had all 13 genes, 5 orders (Ascaridida, Mermithida, Ostreoida, Rhabditida, Veneroida) were missing one gene (ATP8) and one order (Pectinoida) was missing two genes (ATP8 and COX2). Therefore we ended up with a dataset consisting of 721 gene and order combinations representing 869 species across 56 orders. (See A.1 for Accession Numbers.)

2.7.2 Alignments

Sequences were aligned using the MUSCLE [Edgar 2004] algorithm as implemented in Mesquite v. 2.74 [Maddison & Maddison 2007].

2.7.3 Trees

FASconCAT [Kück & Meusemann 2010], a Perl script, was used to generate concatenated gene alignments and a Nexus block for each order that was then used with MrBayes v3.2.1 [Huelsenbeck et al. 2001] to create trees with separate partitions for each gene. These trees were further modified using the burntrees.pl Perl script that appended the actual species names on the nodes of the trees [*Burntrees*]. The trees generated with this method are not likely to be the correct phylogenies. However, previous research has shown that inferences about substitution rates are not sensitive to minor errors in the estimated tree [Yang et al. 2000].

2.7.4 Statistical Methods

For each alignment we tested the null hypothesis of no site-to-site rate heterogeneity as describe in Kosakvosky Pond and Muse 2005 using the "Dual" and "Nonsynonymous" models described there.

We ran the HyPhy dNdSRateAnalysis batch file for both the Dual and Nonsynonymous Variable Rates models for all the alignments separately. The appropriate genetic code was specified as either vertebrate or invertebrate mtDNA. We selected a codon model using the MG94xREV rate matrix with gamma distributions for both the synonymous and nonsynonymous rate categories and choose to use the default initial values. We choose to use 7 synonymous and 10 nonsynonymous rate categories to avoid biasing out estimations of the CV due to the bound (See Chapter 3). Maximum likelihood optimization was performed using HyPhy [Kosakovsky Pond et al. 2005]. Additionally, the magnitude of the rate variation was quantified by computing the CV for both synonymous and nonsynonymous rate.

2.8 Figures

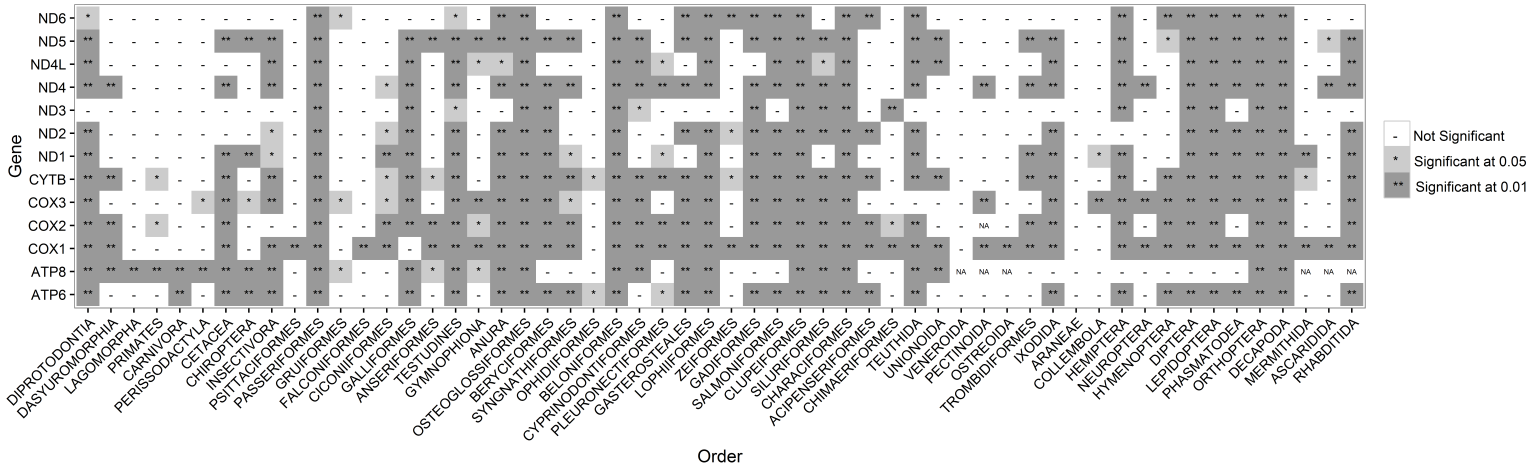


Figure 2.1 Results from the Likelihood Ratio Test for the Presence of Synonymous Rate Variability within Genes. Datasets for each gene × order combination are found to be significant (at 0.01 or 0.05) or nonsignificant after a Bonferroni correction. The null hypothesis of no synonymous rate variation (SRV) from site to site is rejected for 57% of the combinations. Orders are arranged in phylogenetic relation to each other on the x-axis. NAs represent gene × order combinations that were unavailable for analysis.

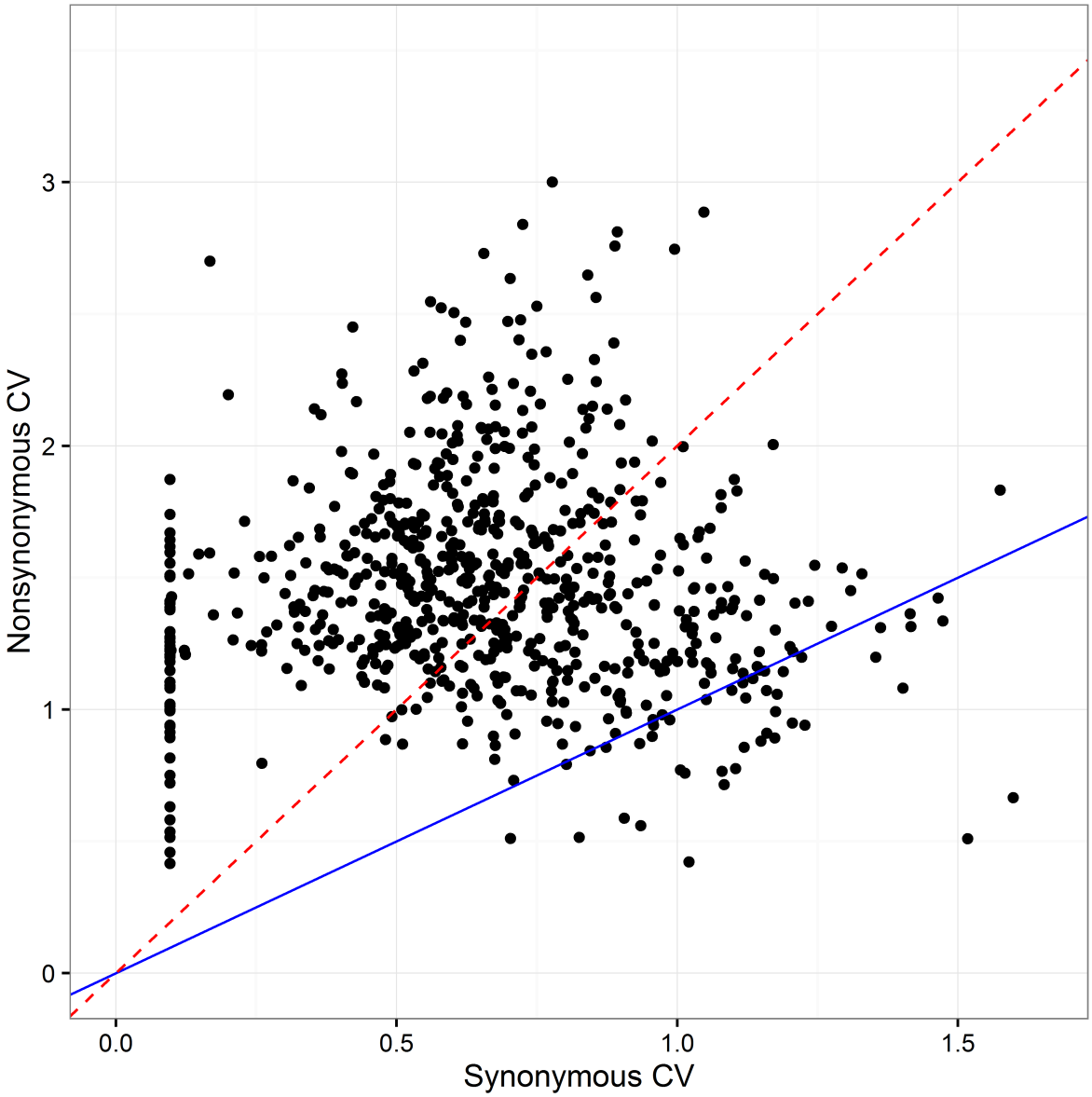


Figure 2.2 Comparison of Synonymous and Nonsynonymous Coefficients of Variation (CV). For each of the 721 datasets, we plot its estimated synonymous and nonsynonymous CV. Points below the blue line are datasets where the synonymous CV exceeds that of nonsynonymous rates; points below the red line had a synonymous CV of at least half the nonsynonymous CV. The vertical line of points on the left represents a numerical artifact for datasets with synonymous CV effectively zero.

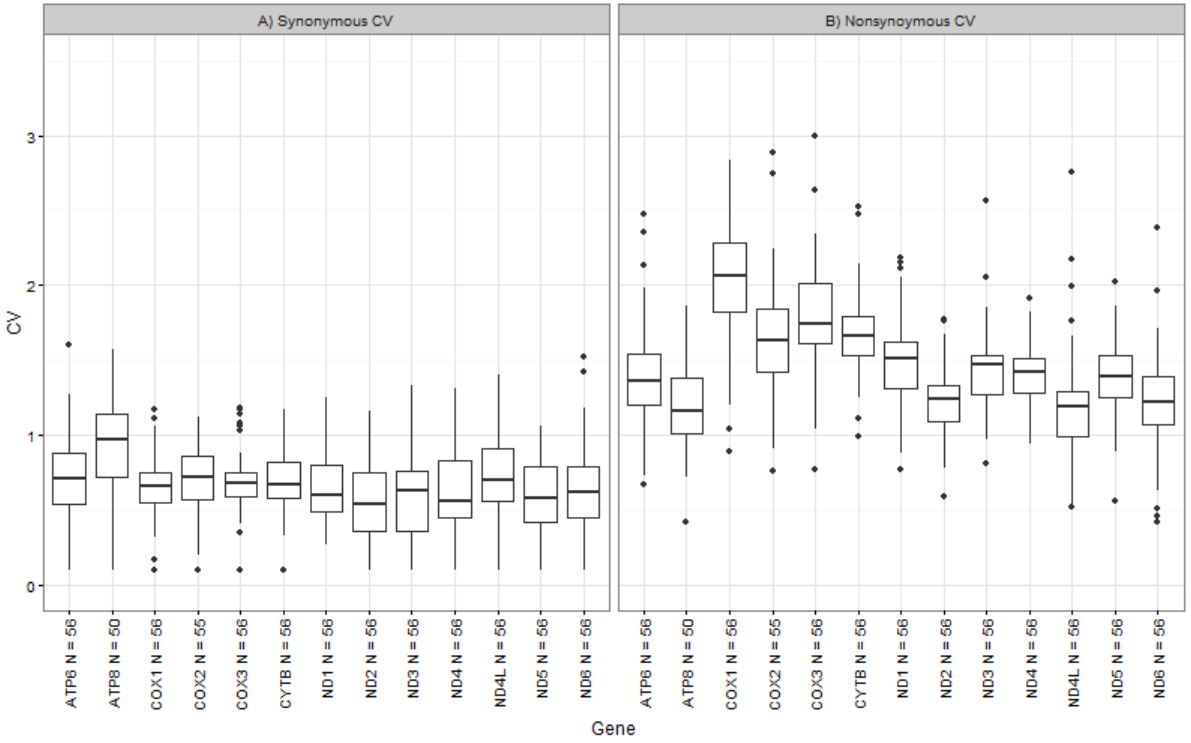


Figure 2.3 Boxplots of the Ranges of Synonymous (A) and Nonsynonymous (B) Coefficients of Variation (CV) for each Gene Group. Boxplots in A indicate the range of synonymous rate variation (SRV) across the metazoan orders for each gene. Boxplots in B represent the range of nonsynonymous rate variation across the metazoan orders for each gene. While the range of synonymous CVs is generally lower than that of the nonsynonymous CVs they are of the same order of magnitude.

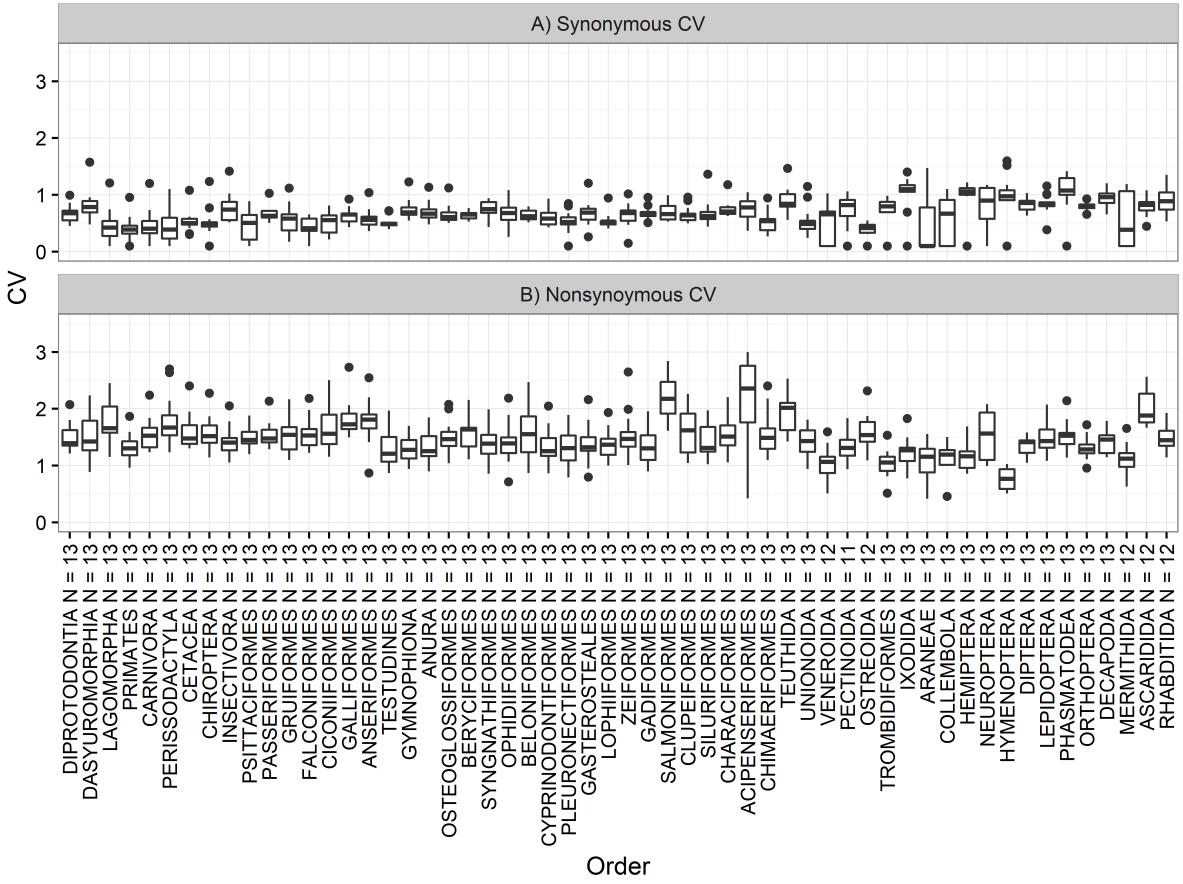


Figure 2.4 Boxplots of the Ranges of Synonymous (A) and Nonsynonymous (B) Coefficients of Variation (CV) for each Metazoan Order. The boxplots in A and B represent the synonymous and nonsynonymous CV ranges respectively across the mitochondrial genes for each Metazoan order. The x-axis of orders is arranged in phylogenetic relation.

CHAPTER

3

UNEXPECTED CONSEQUENCES STEMMING FROM THE SELECTION OF THE NUMBER OF DISCRETE RATE CATEGORIES

3.1 Authors

Frank Mannino[1], Sadie Wisotksy[1], Sergei L. Kosakovsky Pond[2], and Spencer V Muse[1,3]

[1] Bioinformatics Research Center and [3] Department of Statistics, North Carolina State University, Raleigh, NC, USA

[2]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA

3.2 Contribution

This chapter consists of a result which while previously unpublished was first described by Dr. Frank Mannino in his thesis. He derived the result. I ran the analysis described within and created the figures. I contributed to writing the manuscript along with Dr. Mannino, Dr. Kosakovsky Pond, and Dr. Muse. Presented here is an extended version of the manuscript followed by the manuscript itself as the manuscript was subject to a strict word count.

3.3 Abstract

The standard way to model the site-to-site variability of nucleotide substitution rates is to discretize a continuous distribution— most commonly the Gamma distribution— into a small number of equiprobable rate categories. We show that the variance of this discretized distribution is subject to an upper bound determined entirely by the choice of the number of rate categories. This bound can introduce substantial biases into statistical inferences, especially in settings where it is important to accurately estimate parameters governing site-to-site variability in substitution rates. The use of more flexible discrete distributions does not suffer from the same undesirable behavior, and should, therefore, be encouraged when rate inference is of primary interest.

3.4 Introduction

The discrete gamma approach popularized by Yang [1994] has become the standard technique for modeling site-to-site variability of substitution rates in molecular evolutionary studies. This approach bypasses the computationally demanding integration over the unobserved site-specific substitution rates drawn from a continuous distribution, by replacing the continuous density for a discrete one with K values (this is essentially a Riemann sum method for numerical quadrature). Empirical studies [Yang 1994; Yang et al. 2000] and comparative approximation analyses [Felsenstein

2001] have revealed that in terms of goodness of fit (log likelihood), a small number of discrete values, e.g. $K = 4-8$ is sufficient for most applications. The default value of K is set at this level in a number of computer programs, e.g, PAML = 5 [Yang 2007], HyPhy = 4 [Kosakovsky Pond et al. 2005], MrBayes = 4 [Ronquist et al. 2012], codonPHYML = 4 [Gil et al. 2013]. However, when the estimation of the actual value of the variance is of direct interest, the choice of K imposes an upper bound on the estimate of that variance (proof below), which does not depend on the higher moments of the underlying continuous distribution. Examples where this might be a concern include not only the estimation of the variance of substitution rates across sites but also estimates of quantities that are functions of the variance such as the coefficient of variation (CV, the ratio of standard deviation to mean) of the distribution of rates across sites. When inference calls for a direct estimate of variance (or similar quantity), using more categories will produce estimates with less bias. Yang [1994] suggested using $K = 8$ categories if the estimation of the shape parameter of the gamma distribution is the goal. We demonstrate below that maximum likelihood estimates of such parameters improve with increasing K.

3.5 Materials and Methods

3.5.1 Data

The data set used here is the same mitochondrial data set as described in chapter 2. It is comprised of 721 protein-coding mitochondrial DNA gene alignments from the Metazoans. For a detailed description of how the data set was assembled see section 2.7.

3.5.2 Analysis

For each alignment, we ran the "Dual" model [Kosakovsky Pond & Muse 2005], a model with independent synonymous and nonsynonymous substitutions rates estimated by a discretized gamma distribution. We ran this analysis with 3, 4, 5, 7, and 10 rate categories for both the synonymous and nonsynonymous substitution rates.

3.6 Results and Discussion

3.6.1 Upper Bound Derivation

Consider a non-negative valued discrete random variable X having mean μ and possible values $X_i \geq 0, i = 1..K$, where $Pr\{X = X_i\} = 1/K$.

$$\begin{aligned} \text{Var}(X) &:= E\left(X^2\right) - \mu^2 \\ &= \tfrac{1}{K}\sum_{i=1}^{K} X_i^2 - \mu^2 \\ &\leq \tfrac{1}{K}\left(\sum_{i=1}^{K} X_i\right)^2 - \mu^2 \\ &= K\left(\sum_{i=1}^{K} X_i/K\right)^2 - \mu^2 \\ &= (K-1)\mu^2 \end{aligned}$$

The inequality holds because all X_i are non-negative. Some simple algebra further shows that the CV of X is bounded by $\sqrt{K-1}$. Consequently, if the true variance of a continuous distribution exceeds the upper bound imposed by choice of K, estimates of that variance derived from the discretized version of that distribution will be negatively biased.

3.6.2 Bias of Estimated Discrete CV

In figure 3.1 we plot the estimated CVs of synonymous and nonsynonymous substitution rates for a collection of 721 mitochondrial gene alignments (in prep Wisotsky 2018). It is clearly seen that the upper bound for the CV estimate is reached for many datasets, especially when estimating the CV of the nonsynonymous rate distribution. As K increases from 3 to 10 we see that fewer and fewer data sets hit the upper bound; in fact, we see 13.7% of the datasets are at the maximum CV for $K = 3$ while 0% of the datasets hit the upper bound with $K = 10$.

In figure 3.2, we see the nonsynonymous CV as estimated by 3, 5, 7, and 10 rate categories compared to each other along with the x = y line (red). In 3.2.A the nonsynonymous CV estimate with $K = 3$ is plotted verse the nonsynonymous CV estimate with $K = 10$. It is evident from the shift of the points away from the x = y line that there is a severe bias across all the data sets and not just those

which are near the upper bound. In figure 3.2.D we see that the estimates of the nonsynonymous CV are closest to having a linear relationship when comparing 7 and 10 rate categories. However, even in the 7 and 10 rate category comparison, there is still an obvious bias in the estimates.

We see this same bias less drastically in figure 3.3, specifically figure 3.3.A where we compare the synonymous CV estimated with $K = 3$ and $K = 10$. For the synonymous CV, the relationship between estimates with $K = 7$ and $K = 10$ for the synonymous rate categories appears to follow the x = y line. Thus fewer rate categories may be necessary to reach an unbiased estimate of the synonymous CV than the number of rate categories needed for an unbiased nonsynonymous CV estimate.

3.6.3 Bias of Shape Parameter

The upper bound on rate variation creates a bias in discrete estimations of variance and CV with a clear relationship to the number of K rate categories used. The shape parameter (α), which is often used to measure variance is not impacted by this upper bound. However, α is not without its own biases which have also been shown to relate to the number of rate categories used in estimations.

We see the value of the nonsynonymous (figure 3.4) and synonymous (figure 3.5) shape parameters as estimated by 3, 5, 7, and 10 rate categories plotted against each other along with the x = y line (red). When comparing the nonsynonymous shape parameter estimation with $K = 3$ and $K = 10$ (figure 3.4.A) there is a tendency to underestimate the shape with $K = 3$. While not as dramatic as the bias imposed on the discrete estimation of variance seen in 3.2.A there is still a bias. When comparing the nonsynonymous shape parameter estimates for $K = 7$ and $K = 10$ we see this bias is greatly reduced. Additionally, much like with the discrete estimations of CV the estimates of the synonymous shape parameter appear less sensitive to this bias.

3.7 Implications

The most commonly used model for site-to-site rate variation is the gamma distribution with mean 1 and shape parameter α, and it is common to report the estimate of α to indicate the degree of

site-to-site variability. One could estimate the variance (or CV) of this distribution in several ways. Some quantities, such as the variance of the rate distribution, are simple functions of α. If these are estimated by plugging in an estimate of α into the desired function, the upper bound bias problem is largely avoided. There is evidence, however, that the α estimate itself is negatively biased when fewer rate categories are used [Excoffier & Yang 1999; Mannino 2006]. Using the α as an estimate also implies that we assume a continuous gamma distribution truly approximates the distribution of rates across sites. However, estimating these quantities directly from the discrete distribution will lead to significant bias. Such a situation arises whenever there is no simple function relating α to the quantity of interest, or when one prefers not to use a fully parametric description of the rate distribution.

Thus, if a study requires an accurate estimation of the magnitude of CV in any meaningful way, it is essential to use more rate categories, a more general discrete rate distribution [e.g. Kosakovsky Pond & Frost 2005], or a discretization scheme that does not use equiprobable rate classes [e.g. Felsenstein 2001; Kosakovsky Pond & Muse 2005; Yang et al. 2000]. However, if the accurate estimation of the CV is not central to the study, as is the case for most investigations of selection and inference of phylogeny, then the standard 3 or 4 rate categories will likely suffice.

The computational expense for likelihood calculations increases linearly as a function of the number of discrete rate categories K. While this cost was once significant, it is no longer prohibitive. The robustness and reduction in bias granted by using a larger number of categories are likely justified for any analysis targeting actual estimates of properties of the underlying rate distributions.

3.8 Acknowledgments

This work was supported in part by grants R01 GM093939 (NIH/NIGMS).

3.9 Supplementary Materials

3.9.1 Shifting Distribution of Estimates

In figure 3.6 we plot the estimates of the nonsynonymous and synonymous CV (A and B) and shape parameter (C and D) verse the number of rate categories (K) used in the estimation. We see in figure 3.6.A that as the number of K increases the distribution of the estimate spreads. We see that the red points representing the medians of the distribution and the horizontal lines representing the 10% percentiles all shift up slightly. This taken together indicates that the upper bound on variation does not impact only the estimates of those data sets that hit the upper bound but impacts the estimation of all data sets. We see a similar result in the graph with the synonymous CV (B) albeit to a lesser extent.

3.10 Brief Communication Text

Presented here is the text of the brief communication as submitted.

Equiprobable discretized models of site-specific substitution rates underestimate the extent of rate variability

Frank Mannino[1], Sadie Wisotksy[1], Sergei L. Kosakovsky Pond[2], and Spencer V Muse[1,3,*]
[1] Bioinformatics Research Center and [3] Department of Statistics, North Carolina State University, Raleigh, NC, USA
[2]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA
***Corresponding author:** E-mail: muse@stat.ncsu.edu
Associate Editor: TBD

Abstract

It is standard to model site-to-site variability of substitution rates via discretization of a continuous distribution into a small number, K, of equiprobable rate categories. We show that the variance of this discretized distribution is subject to an upper bound determined by the choice of K and the mean of the distribution. This bound can introduce biases into statistical inference, especially when it is important to accurately estimate parameters governing site-to-site variability of substitution rates. When parameter estimation is of primary interest, the use of additional rate categories or the use of more flexible approximation methods should be encouraged.

Key words: Rate estimation, smoothing, bias

Brief Co

The discrete gamma approach popularized by Yang (1994) has become the standard technique for modeling site-to-site variability of substitution rates in molecular evolutionary studies. This approach approximates the computationally demanding integration over the unobserved site-specific substitution rates drawn from a continuous distribution a summation over a discretized version of this distribution with K values. Empirical studies (Yang, 1994; Yang *et al.*, 2000) and comparative approximation analyses (Felsenstein, 2001) have revealed that in terms of model goodness-of-fit, a small number of rate categories is sufficient for most applications. Thus, the default value of K is set at a low level in many popular computer programs: e.g. PAML = 5 (Yang, 2007), HyPhy = 4 (Kosakovsky Pond *et al.*, 2005), MrBayes = 4 (Ronquist *et al.*, 2012), codonPHYML = 4 (Gil *et al.*, 2013), and is rarely modified by the users. However, the choice of K imposes an upper bound on the variance of the discrete distribution (proof below) that does not depend on the higher moments of the underlying continuous distribution. Examples where this might be a concern include not only estimation of the variance of substitution rates across sites, but also estimates of quantities that are functions of the variance such as the coefficient of variation

of the distribution of rates across sites (CV, the ratio of standard deviation to mean). When inference calls for a direct estimate of the variance or similar quantity from the discrete distribution, using more categories, or allowing categories with unequal weights, will produce estimates with less bias.

Consider a non-negative valued discrete random variable X having mean μ and possible values $0 \leq X_1 \leq X_2 \ldots \leq X_K$, where $Pr\{X = X_i\} = 1/K$.

$$\begin{aligned} \mathrm{Var}(X) &:= E(X^2) - \mu^2 \\ &= \tfrac{1}{K}\textstyle\sum_{i=1}^{K} X_i^2 - \mu^2 \\ &\leq \tfrac{1}{K}\left(\textstyle\sum_{i=1}^{K} X_i\right)^2 - \mu^2 \\ &= K\left(\textstyle\sum_{i=1}^{K} X_i/K\right)^2 - \mu^2 \\ &= (K-1)\mu^2 \end{aligned}$$

The inequality holds because all X_i are non-negative. The coefficient of variation (CV) of X, defined as $\sqrt{\mathrm{Var}(X)}/E(X)$, is therefore bounded by $\sqrt{K-1}$. Consequently, if the true variance of a continuous distribution exceeds the upper bound imposed by choice of K, estimates of that variance derived from the discretized version of that distribution will be negatively biased. Indeed, all higher moments of the discrete distribution are subject to similar bounds

$$\begin{aligned} \mathrm{E}(X^n) &= \tfrac{1}{K}\textstyle\sum_{i=1}^{K} X_i^n \\ &\leq \tfrac{1}{K}\left(\textstyle\sum_{i=1}^{K} X_i\right)^n \\ &= K^{n-1}\left(\textstyle\sum_{i=1}^{K} X_i/K\right)^n \\ &= K^{n-1}\mu^n \end{aligned}$$

The most commonly used model for site-to-site rate variation is the gamma distribution with mean 1 and shape parameter α, and it is common to report the estimate of α to quantify the degree of site-to-site variability. One could estimate the variance (or CV) of this distribution in several ways. Some quantities, such as the variance of the rate distribution, are simple functions of α. If these are estimated by plugging in an estimate of α into the desired function, the upper bound bias problem is largely avoided. There is evidence, however, that the α estimate itself is negatively biased (Excoffier and Yang, 1999; Mannino, 2006). Using the α as an estimate also implies that we assume a continuous gamma distribution truly approximates the distribution of rates across sites. However, estimating these quantities directly from the discrete distribution will lead to significant bias. Such a situation arises whenever there is no simple functional relationship between α and the quantity of interest, or when one prefers not to use a fully parametric description of the rate distribution. One common application would be empirical Bayes analysis to infer rates at individual sites.

In Figure 1 we plot the estimated CVs of synonymous and nonsynonymous substitution rates for a collection of 721 mitochondrial gene alignments. It is clearly seen that the discretization induced upper bound is reached for many datasets, especially when estimating the CV of the nonsynonymous rate distribution.

As K increases from 3 to 10, progressively fewer data sets hit the upper bound: 13.7% of the datasets for $K=3$, and only one data set with $K=10$. Thus, if a study requires accurate estimation of the magnitude of CV in any meaningful way, it is essential to use more rate categories, a more general discrete rate distribution (Kosakovsky Pond and Frost, 2005, e.g.), or a discretization scheme that does not use equiprobable rate classes the general (Felsenstein, 2001; Kosakovsky Pond and Muse, 2005; Yang *et al.*, 2000, e.g.). However, if the accurate estimation of the CV is not central to the study, or when the rate distribution is a nuisance parameter, e.g., in phylogenetic inference, then the standard 4 or 5 rate categories will likely suffice.

Acknowledgments

This work was supported in part by grants R01 GM093939 (NIH/NIGMS) .

References

Excoffier, L. and Yang, Z. 1999. Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Molecular Biology and Evolution*, 16(10): 1357–1368.

Felsenstein, J. 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *Journal of Molecular Evolution*, 53(4-5): 447–455.

Gil, M., Zanetti, M. S., Zoller, S., and Anisimova, M. 2013. CodonPhyML: Fast maximum likelihood phylogeny estimation under codon substitution models. *Molecular Biology and Evolution*, 30(6): 1270–1280.

Kosakovsky Pond, S. L. and Frost, S. D. W. 2005. A Simple Hierarchical Approach to Modeling Distributions of Substitution Rates. *Molecular Biology and Evolution*, 22(2): 223–234.

Kosakovsky Pond, S. L. and Muse, S. V. 2005. Site-to-site variation of synonymous substitution rates. *Molecular Biology and Evolution*, 22(12): 2375–2385.

Kosakovsky Pond, S. L., Frost, S. D. W., and Muse, S. V. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5): 676–679.

Mannino, F. V. 2006. *Site-to-Site Rate Variation in Protein Coding Genes*. Ph.D. thesis, North Carolina State University.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology*, 61(3): 539–542.

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

Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8): 1586–1591.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M. K. 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Molecular biology and evolution*, 19(1): 49–57.

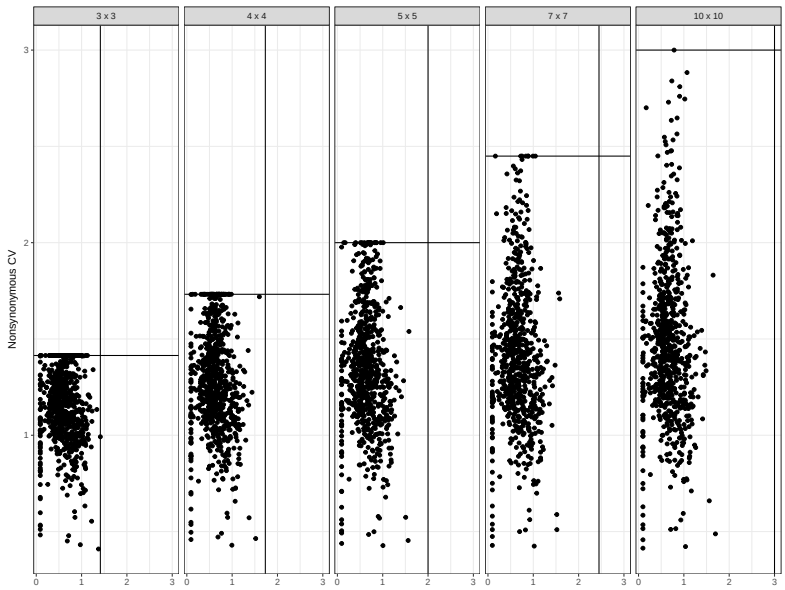


FIG. 1. For each of 13 mtDNA genes for 56 metazoan orders, the estimated coefficient of variation (CV) for nonsynonymous rates is plotted against that for the synonymous rates using an increasing number of rate categories, K= 3, 4, 5, 7, 10.

The horizontal line in each sub-plot is the maximum possible CV estimate according to equation $\sqrt{K-1}$. Note that as the number of rate categories increase fewer points reach the theoretical upperbound. The upper bound does not appear to be consequential for the synonymous substitution rate estimates in this study.

3.11 Figures

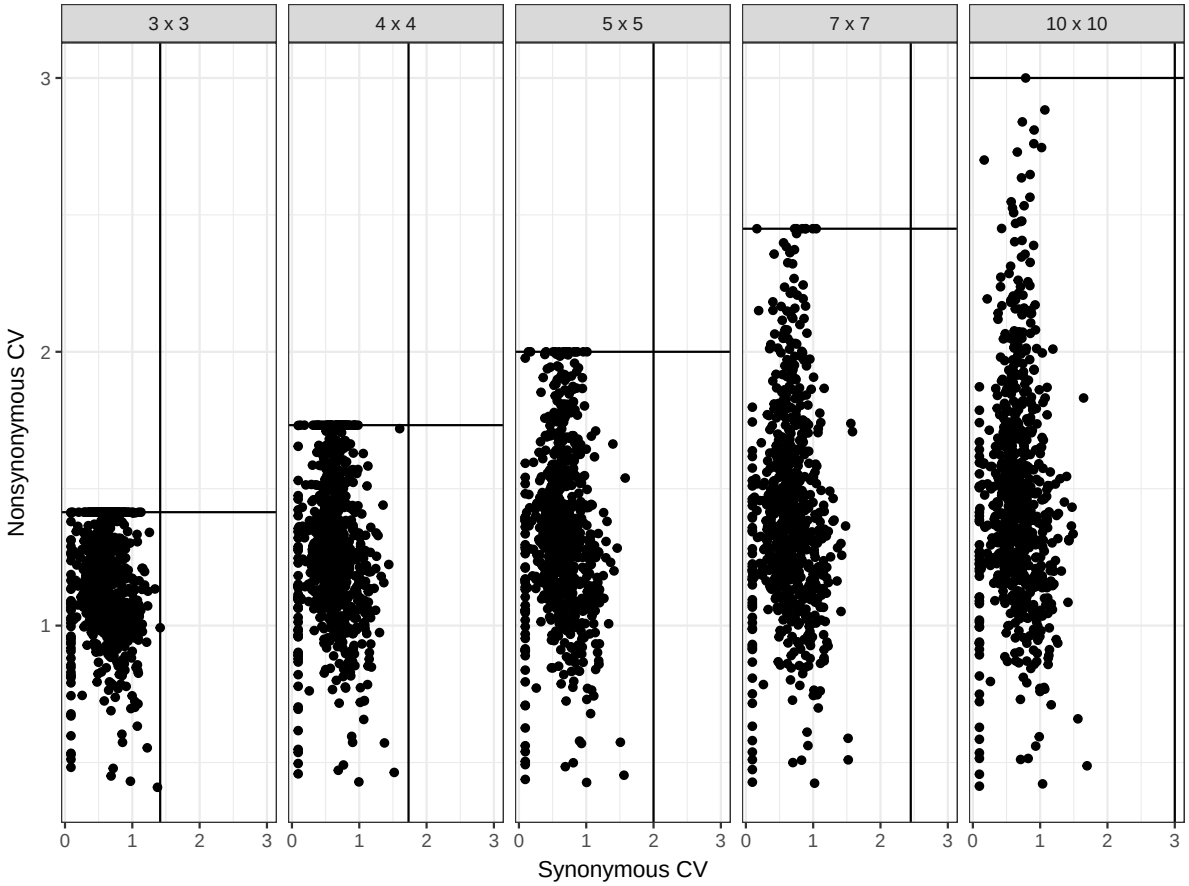


Figure 3.1 Variation Bound Plots. For each of 13 mtDNA genes for 56 metazoan orders, the estimated coefficient of variation (CV) for nonsynonymous rates is plotted against that for the synonymous rates using an increasing number of rate categories, K= 3, 4, 5, 7, 10. The horizontal line in each sub-plot is the maximum possible CV estimate according to the equation $\sqrt{K-1}$. Note that as the number of rate categories increase fewer points reach the theoretical upper bound. The upper bound does not appear to be consequential for the synonymous substitution rate estimates in this study.

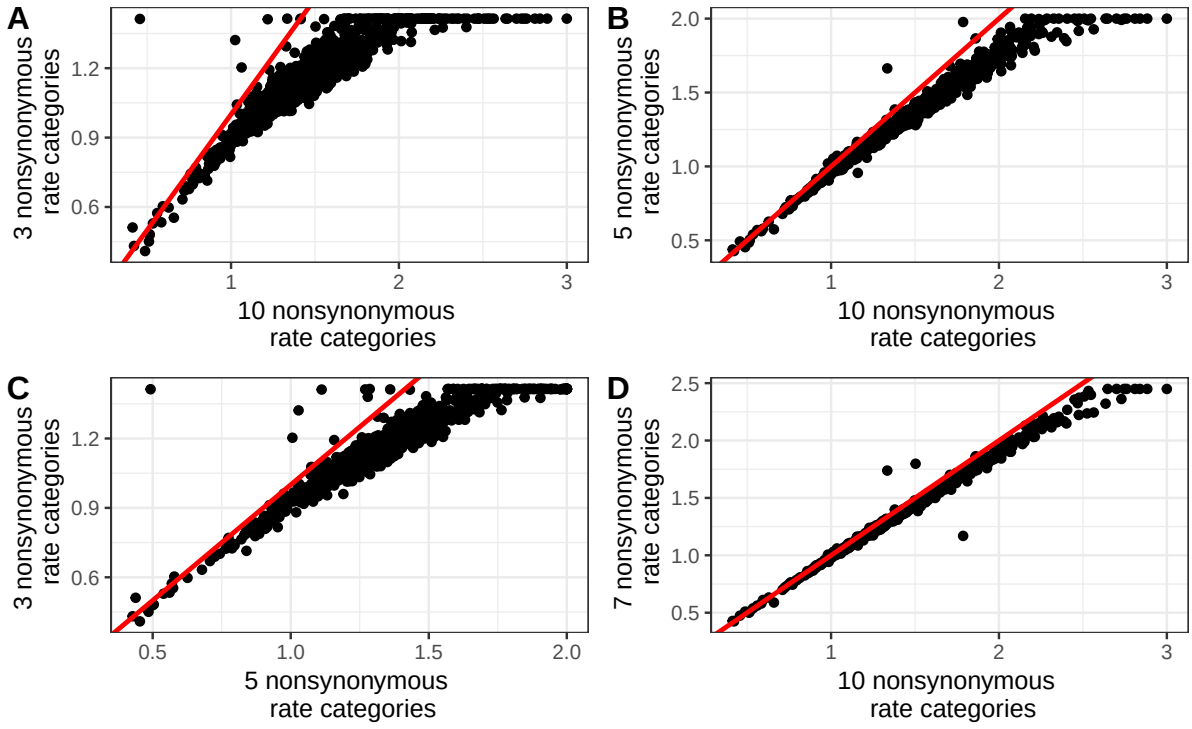


Figure 3.2 Estimated Nonsynonymous CV for Varying Rate Categories. For all the mitochondrial data sets the estimated nonsynonymous coefficient of variation (CV) using 10 rate categories plotted against the estimated nonsynonymous CV using 3 categories (A), 5 rate categories (B), and 7 rate categories (D). Additionally, the nonsynonymous CV estimated with 3 rate categories was plotted against the nonsynonymous CV using 5 rate categories (C). The red line represents the x = y line.

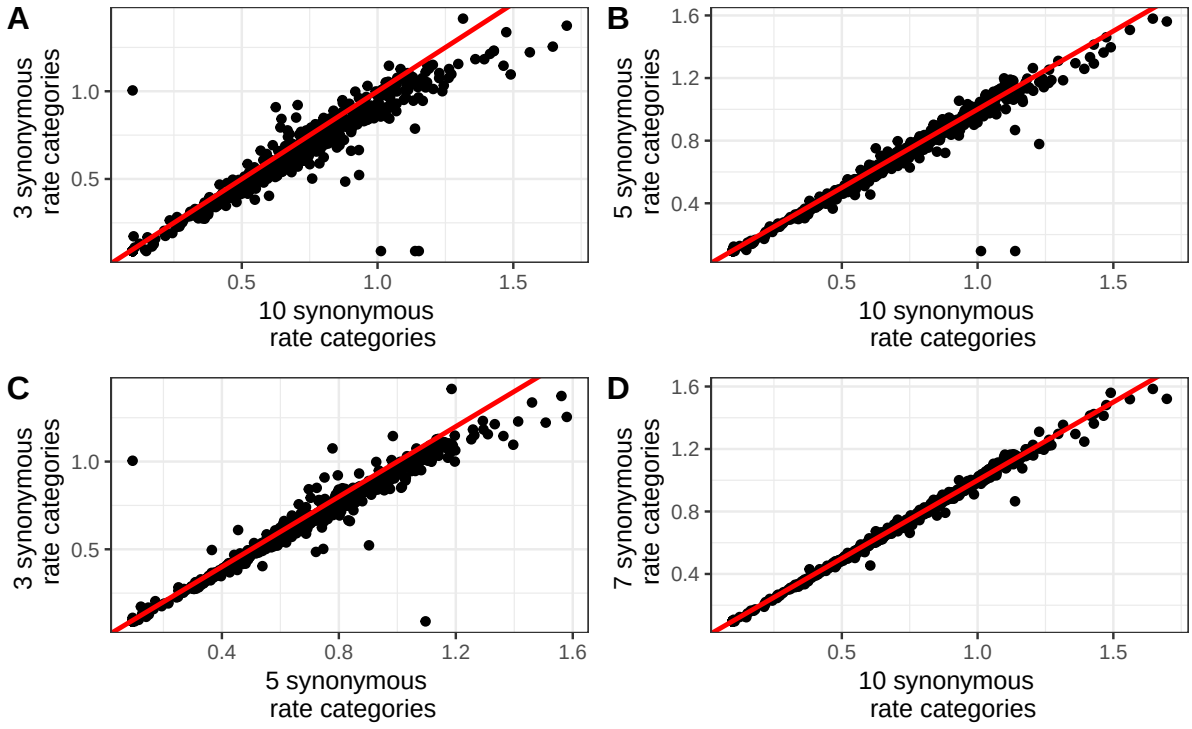


Figure 3.3 Estimated Synonymous CV for Varying Rate Categories. For all the mitochondrial data sets the estimated synonymous coefficient of variation (CV) using 10 rate categories plotted against the estimated synonymous CV using 3 categories (A), 5 rate categories (B), and 7 rate categories (D). Additionally, the synonymous CV estimated with 3 rate categories was plotted against the synonymous CV using 5 rate categories (C). The red line represents the x = y line.

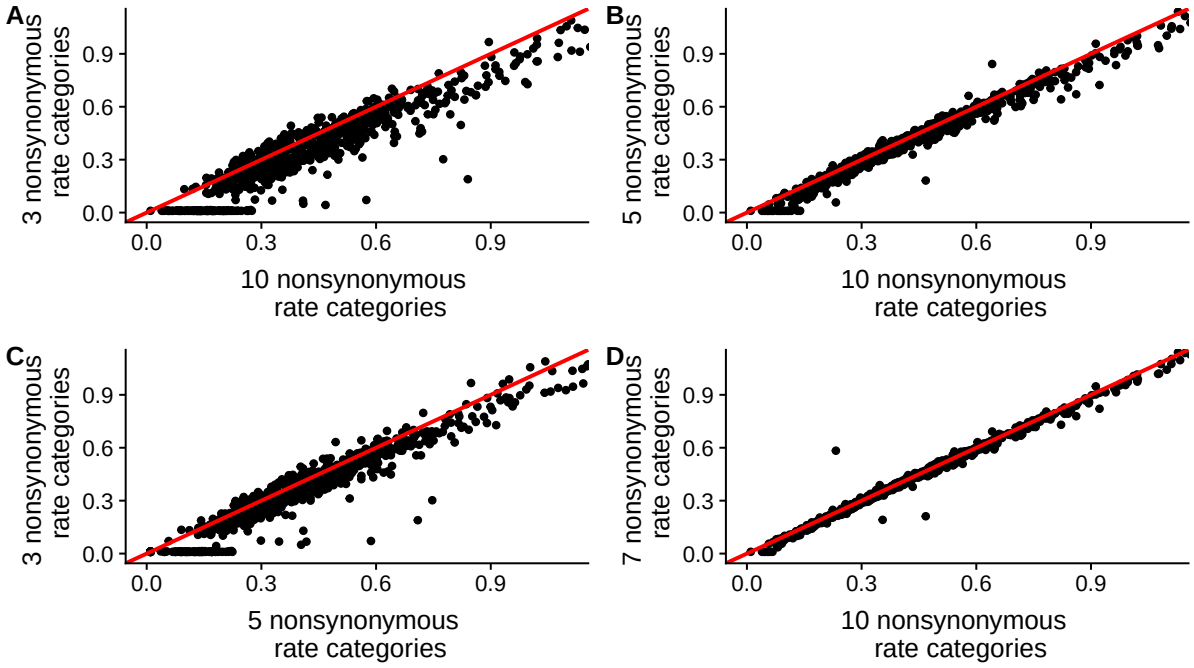


Figure 3.4 Estimated Nonsynonymous Shape Parameter for Varying Rate Categories. For all the mitochondrial data sets the estimated nonsynonymous shape parameter using 10 rate categories plotted against the estimated nonsynonymous shape parameter using 3 categories (A), 5 rate categories (B), and 7 rate categories (D). Additionally, the nonsynonymous shape parameter estimated with 3 rate categories was plotted against the nonsynonymous shape parameter using 5 rate categories (C). The red line represents the x = y line.

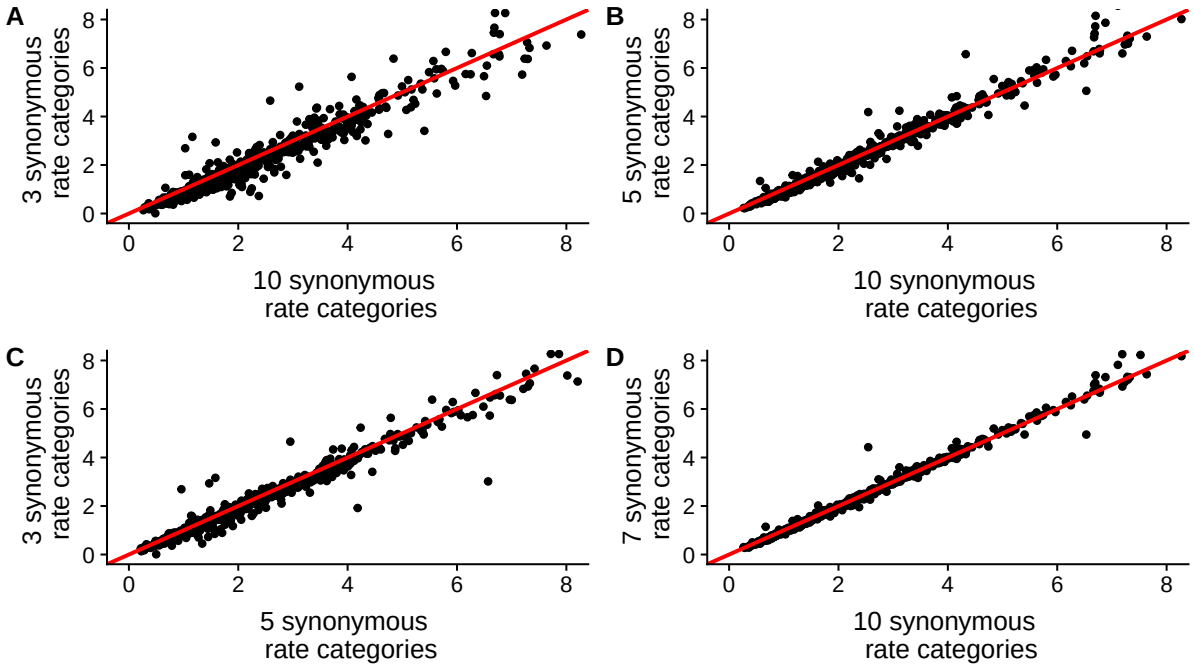


Figure 3.5 Estimated Synonymous Shape Parameter for Varying Rate Categories. For all the mitochondrial data sets the estimated synonymous shape parameter using 10 rate categories plotted against the estimated synonymous shape parameter using 3 categories (A), 5 rate categories (B), and 7 rate categories (D). Additionally, the synonymous shape parameter estimated with 3 rate categories was plotted against the synonymous shape parameter using 5 rate categories (C). The red line represent the x = y line.

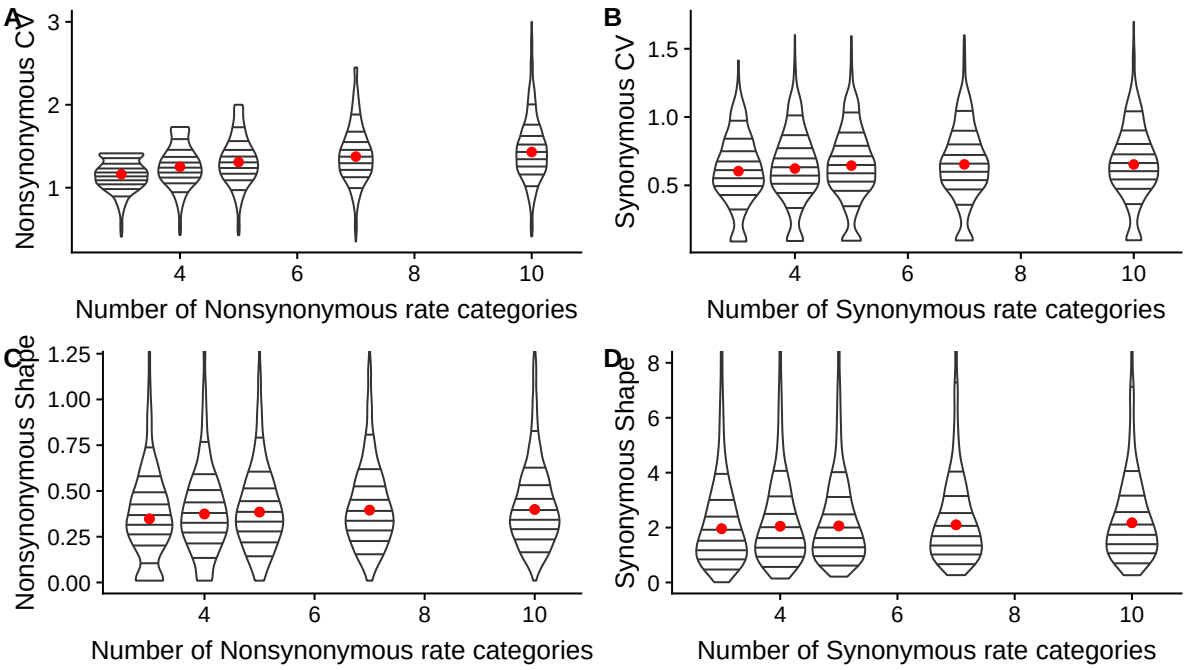


Figure 3.6 Violin Plots. Violin plots of the nonsynonymous CV (A) and nonsynonymous shape parameter (C) versus the number of nonsynonymous rate categories as well as the synonymous CV (B) and synonymous shape parameter (D) versus the number of synonymous rate categories. The height of the violin plots represents the range of the parameter over the data sets. The width of plots represents the density of data sets across each parameter. The horizontal lines split the violin plots into quintiles representing 10% of the data sets each. The red dot represents the median of each.

BIBLIOGRAPHY

Excoffier, L. & Yang, Z. (1999). "Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees". *Molecular Biology and Evolution* **16**.10, pp. 1357–1368.

Felsenstein, J. (2001). "Taking variation of evolutionary rates between sites into account in inferring phylogenies". *Journal of Molecular Evolution* **53**.4-5, pp. 447–455.

Gil, M. et al. (2013). "CodonPhyML: Fast Maximum Likelihood Phylogeny Estimation under Codon Substitution Models". *Molecular Biology and Evolution* **30**.6, pp. 1270–1280.

Kosakovsky Pond, S. L. & Frost, S. D. W. (2005). "A Simple Hierarchical Approach to Modeling Distributions of Substitution Rates". *Molecular Biology and Evolution* **22**.2, pp. 223–234.

Kosakovsky Pond, S. L. & Muse, S. V. (2005). "Site-to-site variation of synonymous substitution rates." en. *Molecular Biology and Evolution* **22**.12, pp. 2375–2385.

Kosakovsky Pond, S. L. et al. (2005). "HyPhy: Hypothesis testing using phylogenies". *Bioinformatics* **21**.5, pp. 676–679.

Mannino, F. V. (2006). "Site-to-Site Rate Variation in Protein Coding Genes". PhD thesis. North Carolina State University, p. 153.

Ronquist, F. et al. (2012). "MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space". *Systematic Biology* **61**.3, pp. 539–542.

Yang, Z. (2007). "PAML 4: Phylogenetic Analysis by Maximum Likelihood". *Molecular Biology and Evolution* **24**.8, pp. 1586–1591.

Yang, Z. (1994). "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods". *Journal of Molecular Evolution* **39**.3, pp. 306–314.

Yang, Z. et al. (2000). "Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites". *Molecular biology and evolution* **19**.1, pp. 49–57.

CHAPTER

4

ACCOUNTING FOR SITE-TO-SITE SYNONYMOUS RATE VARIABILITY REVEALS HIGH FALSE POSITIVE RATE IN TEST OF SELECTION

4.1 Authors

Sadie R. Wisotsky*[,1], Sergei L. Kosakovsky Pond[3], and Spencer V. Muse[1,2]

[1] Bioinformatics Research Center and [2] Department of Statistics, North Carolina State University, Raleigh, NC, USA

[3]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA

4.2 Contribution

For this paper, I wrote a new method BUSTED+SRV in HyPhy with the assistance of Dr. Sergei Kosakovsky Pond based on a previous method by Murrell et al. [2015]. I performed the analysis described within on the empirical and simulated data sets, created the simulated data sets, the figures, and tables within. Dr. Sergei Kosakovsky Pond provided the Selectome [Moretti et al. 2014] data set as it was used in the paper where BUSTED was introduced [Murrell et al. 2015]. I wrote the paper along with Dr. Spencer Muse and Dr. Sergei Kosakovsky Pond.

4.3 Abstract

Most standard methods of detecting selection use substitution models with the assumption of a constant synonymous substitution rate from site-to-site. However, a growing body of literature fails to support this assumption from a biological perspective. Previous work has shown that synonymous substitution rates do vary from site to site at a magnitude comparable to that of nonsynonymous substitution rate variation, which is accounted for in current methods. Here, we present a new method, BUSTED+SRV, of gene-wide selection identification that includes site-to-site synonymous rate variation (SRV). BUSTED+SRV is based off the previously published BUSTED method. Analysis under the null reveals BUSTED+SRV behaves similarly to that of BUSTED indicating it is a viable inference method. A comparison of the inferences on the Selectome data set reveals a large reduction in the number of data sets with evidence of positive selection when synonymous rate variation is accounted for. The results from the simulated data reveal this reduction is likely due to a high false positive rate for BUSTED when SRV is present.

4.4 Introduction

Models of sequence evolution aim to accurately reflect the underlying biological processes involved while balancing the computational costs associated with modeling such a complex problem. These models are often used as the basis of methods to infer information about sequence evolution, such as phylogenies, the presence of selection, the strength of selection, and the molecular clock rate. As our knowledge of the biological process of sequence evolution changes, so should our models change to reflect the new information. For instance, synonymous substitutions have long been considered neutral or silent when it comes to evolution. However, more recent studies [eg. Chamary & Hurst 2005; Chamary et al. 2006; Hurst & Pál 2001; Kubatko et al. 2016; Lawrie et al. 2013; Shields et al. 1988] show that this is not the case and selective pressures do act on synonymous substitutions. Synonymous substitutions at sites have been implicated in shifting codon bias, affecting the stability of mRNA, causing alternative splicing events, changing the translational speed, and have even been associated with several human diseases [Agashe et al. 2016; Bhardwaj 2014; Brandis & Hughes 2016; Duan et al. 2003; Eyre-Walker 1996; Supek et al. 2014; Takata et al. 2016]. Additional studies have shown that the synonymous substitution rates vary across sites within genes in a variety of organisms and genes [Dimitrieva & Anisimova 2014; Kosakovsky Pond & Muse 2005, in prep Wisotsky 2018]. Prior research has also shown accounting for variable synonymous rates across sites improves the model fit [Kosakovsky Pond & Muse 2005; Mayrose et al. 2007]. Additionally, those studies show that incorporating synonymous rate variation can reduce the number of positively selected sites detected while also indicating previously unaccounted for positively selected sites. However, despite this evidence, most models of sequence evolution assume synonymous substitution rates are constant across all sites and that only the nonsynonymous substitution rates vary.

Here, we introduce a new method for inferring positive selection anywhere on a tree while accounting for SRV called BUSTED+SRV, based on the previously published BUSTED [Murrell et al. 2015]. BUSTED, which stands for branch-site unrestricted statistical test for episodic diversification, allows users to specify foreground and background branches on a tree and then runs an analysis to

determine if at least one site on at least one foreground branch has experienced positive selection. BUSTED and BUSTED+SRV provide gene-wide results and not site-specific ones. In its current implementation, BUSTED does not have the ability to account for SRV. It instead assumes, as most methods typically do, that the synonymous substitution rate across sites is constant and thus attributes any variation in the overall ω rate to nonsynonymous rate variation alone. Our new method, BUSTED+SRV, allows both the synonymous and nonsynonymous substitution rates to vary across sites independently. We present a large-scale investigation of how adding SRV to this method of selection detection impacts its performance and inferences. We use simulated data under the null hypothesis of no selection and no SRV to determine if the new method itself is a viable means of inference. We then compare the selection inferences on our empirical data set for the SRV inclusive and exclusive methods to determine where they agree and disagree. Further simulations are analyzed to determine the possible underlying causes of the high rate of disagreement we see between the two methods on the empirical data sets.

4.5 Materials and Methods

4.5.1 Empirical Data

For our empirical data set, we used the Selectome database [Moretti et al. 2014], a database of positive selection. Specifically, we looked at the Eusteleostomes gene trees collected from version 6 of the database. There were initially 13,714 trees total but the number was reduced to 13,311 trees due to certain trees being prohibitively computationally costly to run. There were 6543 subtrees that showed evidence of positive selection according to the initial analysis run by Selectome, which implemented a branch-site model [Zhang et al. 2005] in CodeML [Yang 2007] to detect selection. We chose this dataset as we have evidence of SRV being present in Metazoans (see Chapter 2) and it is a curated data set geared towards the detection of positive selection.

4.5.2 Model

BUSTED+SRV stands for branch site unrestricted statistical test for episodic diversification with the addition of synonymous substitution rate variation across sites. The original BUSTED method [Murrell et al. 2015] did not account for synonymous rate variation across sites. BUSTED+SRV has an additional α rate parameter that is allowed to vary across codons independently of ω accounting for SRV. Like BUSTED, we use a BS-REL framework to model the process. Here, the rate matrix used is a modification of the MG94 model [Muse & Gaut 1994] and the instantaneous rate matrix is q_{ij} where q is the rate of change from codon i to j:

$$q_{ij}(c,\theta,\pi) = \begin{cases} \alpha_c \theta_{ij} \pi_j & \delta_{ij} = 1,\ AA(i) = AA(j), \\ \omega_c \theta_{ij} \pi_j & \delta_{ij} = 1,\ AA(i) \neq AA(j), \\ 0 & \delta_{ij} > 1 \end{cases}$$

where $c \in \{1,2,3\}$ is the rate category, $\delta(i,j)$ is the number of nucleotide differences between codon i and codon j and $AA(i)$ is the amino acid coded for by codon i. Here, θ stands for the nucleotide substitution parameters that follow the General Time Reversible model [Tavaré 1986], although any nucleotide model can be used, and π are the equilibrium codon frequencies. The α_c is the rate of synonymous substitution variation associated with category c while ω_c is the rate of nonsynonymous substitution variation associated with category c. This allows the synonymous substitution rate to vary from site to site but not from branch to branch. This also means the ω values from BUSTED and BUSTED+SRV allow for a direct rate comparison. Like BUSTED, branches are split into foreground and background partitions. The ω rates are estimated independently and are restricted such that $\omega_1 \leq \omega_2 \leq 1 \leq \omega_3$. While here we use the default of $c = 3$ rate categories, BUSTED+SRV can accommodate fewer or more rate categories for both the synonymous and nonsynonymous rates.

4.5.3 Simulation

In order to explore the rate of false positives along with the power of both BUSTED and BUSTED+SRV, I simulated data sets under the BUSTED+SRV framework varying the number of codons, sequences, the amount of SRV present, and the strength of selection present. See table B.1 for a list of parameters. Many of the values for these simulation parameters were based on the estimates from the empirical data set. Each simulation used 3 synonymous and 3 nonsynonymous rate categories. Each combination of parameters simulated consisted of 100 replicates. The replicates were analyzed by both BUSTED and BUSTED+SRV.

4.5.4 Implementation

Both BUSTED and BUSTED+SRV are implemented in the HyPhy batch language [Kosakovsky Pond et al. 2005]. BUSTED's original implementation is available at HTTP://test.datamonkey.org/busted with additional data visualization at HTTP://vision.hyphy.org/BUSTED. A copy of the BUSTED+SRV code can be found at: HTTP://github.com/srwis/BUSTED-SRV/blob/hyphy-update/HBL/BUSTED-SRV.bf.

4.6 Results

4.6.1 Performance of BUSTED+SRV

In order to test the performance of BUSTED+SRV, we simulated data under the null ($\omega_3 = 1$ and coefficient of variation (CV) of SRV = 0) and analyzed it with both BUSTED and BUSTED+SRV. For our analysis, we use $p \leq 0.05$ as a measure of significance. Therefore, we would expect to see about 5% of the data sets with $p \leq 0.05$. We would also expect that the p values under the null of no selection would be distributed uniformly. Here, we look at the behavior of all the data sets simulated with $\omega_3 = 1$ and CV of SRV = 0 representing 1000 replicates with varying numbers of sites and sequences. In figure 4.1.A, we see the range of p values for BUSTED (red bars) and BUSTED+SRV (blue bars) from $p = 0$ to $p = 1$ on the x-axis and the number of data sets that fall within each bin on

the y-axis. We see that the distribution of p values for BUSTED and BUSTED+SRV do not follow a uniform distribution and in fact have a large spike of data sets around $p = 1$. This is due to the two-step test BUSTED and BUSTED+SRV perform. In figure 4.1.B we see a close up version of the histogram in figure 4.1.A showing the p values from $p = 0$ to $p = 0.15$ on the x-axis. Here we see that BUSTED and BUSTED+SRV behave similarly and that we do not see the expected 5% of the data sets with a $p \leq 0.05$. This is due to the conservative nature of the test statistic for both analyses and the behavior has been noted previously for BUSTED [Murrell et al. 2015].

4.6.2 Accounting for SRV Improves Model Fit

The Akaike information criterion with a correction for small sample size (AICc) provides a measure of relative goodness of fit between models. In figure 4.3, the x-axis represents the difference between the AICc for BUSTED and the AICc for BUSTED+SRV from the Selectome analysis. We see that the overwhelming majority of alignments have a positive difference between the two indicating that relative to each other the BUSTED+SRV has the smaller AICc and by this measure provides a better fit for the empirical data.

4.6.3 Selectome Analysis Reveals High False Positive Rate

BUSTED and BUSTED+SRV follow similar trends when it comes to the fraction of alignments experiencing selection as seen in figure 4.2, excepting that BUSTED+SRV always shows a lower fraction of data sets under selection. The fraction under selection stays fairly consistent as the number of sequences increases. As expected there is an increase in the fraction of alignments under selection as the maximum likelihood estimate of ω_3 for BUSTED increases as well as a general increase along with the number of sites in each alignment.

BUSTED and BUSTED+SRV agree that the majority of the gene alignments (72%) from Selectome experience no selection. According to table 4.1, BUSTED finds evidence of selection occurring on at least one site on at least one branch of the tree for 24% (3,197) data sets out of all alignments analyzed. BUSTED+SRV finds evidence of positive selection in 16% (2,092) of the data sets. Of those

2,092 data sets with evidence of selection according to BUSTED+SRV, 457 or 4% of the total 13,311 data sets did not show evidence of selection according to BUSTED. Of the total data sets analyzed, BUSTED and BUSTED+SRV both find evidence of selection on 12% of them. So of the 3,197 data sets with evidence of selection according to BUSTED, only 1,635 or 12% of the total still exhibit evidence of selection according to BUSTED+SRV. Meaning that, when accounting for positive selection using BUSTED+SRV, half of the data sets in which BUSTED detects evidence of selection no longer show evidence of positive selection. This potentially indicates that half of the cases of positive selection detected by BUSTED may be false positives.

4.6.4 Simulation Finds High False Positives

In figure 4.4, we see the trends of power and the calculated p values according to BUSTED and BUSTED+SRV. For this figure, all replicates have a tree with 31 sequences and a sequence length of 5000 codons. The true ω_3 value for each subsection of the graph is labeled along the right side. We see represented here data sets simulated with $\omega_3 \in \{1, 1.1, 2.077, 6\}$. The x-axis of this figure is the true synonymous coefficient of variation (CV of SRV) which all of the 100 replicates were simulated with. The y-axis goes from 0 to 1 and represents the power according to each method for the lines and the p values for each boxplot simultaneously. Using this figure we are able to see how both the calculated p values and the power for BUSTED and BUSTED+SRV change as the true CV of SRV increases under varying amounts of selective pressures. Similar figures are presented for varying number of sites and sequences in figures 4.5, 4.6, and 4.7.

When we simulate data under neutral selection ($\omega_3 = 1$), we would expect to see higher p values and lower power according to both analyses. The blue boxplots representing the range of p values according to BUSTED+SRV stay compact around $p = 1$, aligning with expectations. BUSTED+SRV's power, represented by the blue line, hovers around 0 and is unaffected by the amount of SRV present when $\omega_3 = 1$. This is what we would expect to see when there is only neutral selection across sites and branches.

This behavior is not what we see for BUSTED. We see that for lower levels of SRV (CV of SRV $\leq$

0.547), the distribution of p values for BUSTED and BUSTED+SRV are similar and indicate there is no evidence of positive selection. As the true CV of SRV increases, the range of p values for BUSTED, represented by the red boxplots, starts to shift downward and grow smaller until all of the p values according to BUSTED are $p < 0.05$ indicating evidence of selection on all replicates. We see a rapid increase of false positives from BUSTED starting at CV of SRV = 0.697 where 46% of the replicates have $p \leq 0.05$. At the same time, we see the red line, representing the power for BUSTED, increase until it is equal to 1 indicating again that BUSTED is detecting evidence of positive selection on all the data sets in each replicate. We see the same trends for the simulations with fewer sites and sequences and $\omega_3 = 1$ in figure 4.5. When we simulate nearly neutral selection ($\omega_3 = 1.1$) we see a similar trend to that of the neutral simulation. The calculated p values for BUSTED+SRV stay above 0.05 no matter the synonymous CV and the BUSTED p values drop to near zero as the true CV of SRV increases. Additionally, we see the same trends for the power of each analysis.

As we increase the amount of selection present to the median estimated $\omega_3 = 2.077$ from the Selectome dataset, we see a change in behavior. When positive selection is present, we see both BUSTED and BUSTED+SRV behave as we would expect. The caluculated p values for both analyses remain lower than 0.05 and the power for both remains close to one. There is some deviation from this trend by BUSTED+SRV in the higher ranges of synonymous CV. It may be that at higher levels of synonymous rate variation and moderate levels of positive selection that the α component of the ω ratio is overestimated and thus outweighs the β component. The values of synonymous rate variation this occurs at are in the extreme compared to our empirical data. While there were alignments with a synonymous CV equal to or greater than 10.579 the majority of the datasets fell between a synonymous CV 0.545 and 0.761 for the Selectome dataset as seen in figure 4.8. In figure 4.6, we also see this loss of power at moderate levels of selection for BUSTED+SRV. Although, with fewer sequences and codons BUSTED's power also decreases but still remains higher than that of BUSTED+SRV.

Finally, when we simulated with a higher amount of positive selection ($\omega_3 = 6$), we see that even at higher levels of synonymous rate variation everything behaved as expected. In figure 4.4 we see

that the range of p remains low as the true CV of SRV increase for both BUSTED and BUSTED+SRV. We see the power of each analysis remains around 1 as well. When we look at figure 4.7, we see this is true of practically all combinations of sequence number and length we looked at when $\omega_3 = 6$. We do see a decrease in power for both analyses with 16 sequences and 100 codons. Also, we see that BUSTED+SRV loses some power around a true CV of SRV = 5.533 for the simulations with fewer sequences and codons. This could be attributed to the conservative nature of the test.

4.7 Discussion

First, we see that the behavior of BUSTED+SRV under the null hypothesis of no selection and no synonymous rate variation behaves similarly to BUSTED. In figure 4.1, we see that BUSTED+SRV exhibits the same conservative nature as BUSTED. The distribution of p values for both analyses are not uniform but they are very similar to each other. This indicates BUSTED+SRV is a valid conservative test of selection.

We find that BUSTED+SRV is a better model fit than BUSTED without SRV which agrees with previous results that including the site-to-site synonymous rate variation improves the fit of a model [Kosakovsky Pond & Muse 2005; Mayrose et al. 2007]. We do see in figure 4.2 that the fraction of alignments under selection for both methods follow similar trends across increasing sequences, sites, ω_3 values and CV of SRV values. However, we also see in the same figure that for the same window of the Selectome data set, BUSTED+SRV finds fewer data sets with evidence of positive selection. This on its own could indicate either a lower power for BUSTED+SRV or a high false positive rate that is consistent across varying parameters for BUSTED. However, the following discussion lists evidence that strongly supports the presence of a high false positive rate.

From our empirical data analysis, we see in table 4.1 that BUSTED and BUSTED+SRV do agree that the majority of the data sets from Selectome show no evidence of positive selection (72%). They also agree that another 12% of the total Selectome data sets show evidence of positive selection. However, when you account for SRV, nearly half of the data sets that BUSTED found to show evidence of selection no longer do. This is similar to the results from Kosakovsky Pond & Muse [2005] and

Mayrose et al. [2007] that show that accounting for SRV changes inferences made on site-wise selection. While they did not look at gene-wide identification of selection, they did see that fewer sites showed evidence of positive selection when a model accounting for SRV was used. This decrease in the number of data sets that show evidence of positive selection when using BUSTED+SRV may indicate that BUSTED has a high false positive rate. However, because this is empirical data we cannot say for certain. Therefore, we simulated data under a variety of parameters to determine if the decrease can be attributed to a high false positive rate.

Our results from the simulated data show that for data sets simulated under neutral or nearly neutral selection ($\omega_3 = 1$ and $\omega_3 = 1.1$ respectively), BUSTED+SRV behaves as expected, finding no evidence of selection. In figure 4.4 we see that the blue line representing the power of BUSTED+SRV stays around zero for all values of the simulated CV of SRV for both $\omega_3 = 1$ and $\omega_3 = 1.1$. We see for the same graphs that the range of the p values represented by the blue boxplots for BUSTED+SRV remains above the significance level of $p \leq 0.05$. The red line in the same plots which represents the power of BUSTED behaves as expected at lower levels of CV of SRV. At a CV of SRV = 0.574, BUSTED shows decreasing estimates of p and increasing power indicating evidence of selection when the data was simulated without positive selection. Note that a synonymous CV of 0.574 is well within the range of the typical CV of SRVs we see in our Selectome data set, as the majority of the data sets have an estimated synonymous CV between 0.545 and 0.761 (see figure 4.8). We see this jump in power for BUSTED for simulations with 16 and 31 sequences as well as varying numbers of codons from 100 to 5000 in figure 4.5. It should be noted that with fewer sequences and sites in the simulated data sets the increase in power for BUSTED (red line) is not as dramatic but it does still occur. Seeing this pattern across simulations with varying numbers of sites and sequences is a strong indication that the behavior is due to unaccounted for SRV and not some other factor. These results taken alongside the decrease in the number of empirical data sets showing evidence of positive selection provides a strong indication that not accounting for SRV results in high false positive rates for BUSTED.

Presented here is strong evidence of a high false positive rate of identifying positive selection using BUSTED due to the constant site-to-site synonymous rate assumption. These false positive

rates occur at levels of SRV that are not atypical of the data sets we have examined here and previously (See Chapter 2). Previous studies have shown that SRV is a widespread phenomenon [Dimitrieva & Anisimova 2014, in prep Wisotsky 2018] suggesting that unless there is *a priori* knowledge that a data set does not experience SRV it should be included in inferences. Furthermore, and perhaps most importantly, the high false positive rate seen here suggests the possibility that other methods of selection detection or any statistical inferences that are reliant on assuming a constant synonymous rate from site to site may also experience similar behavior.

4.8 Acknowledgments

This work was supported in part by grants R01 GM093939 and T32ES007329 (NIH/NIGMS).

4.9 Figures and Tables

Table 4.1 Two-way Table of Selectome Positive Selection. (UNCORRECTED $P \leq 0.05$) Fraction of data sets under selection. The fraction of total data sets categorized by if there is evidence of selection according to BUSTED and according to BUSTED+SRV.

		BUSTED+SRV		
		No Selection	Selection	Total
BUSTED	No Selection	0.72	0.04	0.76
	Selection	0.12	0.12	0.24
	Total	0.84	0.16	

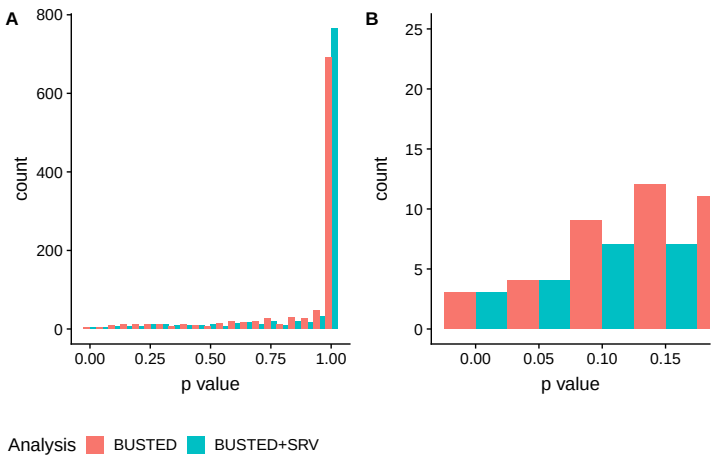


Figure 4.1 Histograms of the P-Values According to BUSTED and BUSTED+SRV. The p-values as calculated by BUSTED (red) and BUSTED+SRV (blue) are plotted with a range of $p = 0$ to $p = 1$ (A) and a range of $p = 0$ to $p = 0.15$ (B). The y-axis for both plots represent the number of data sets that fall within the 0.05 range of each bin. The data sets represented here are simulated with a CV of SRV $= 0$ and an $\omega_3 = 1$ but have a varying number of sites and sequences.

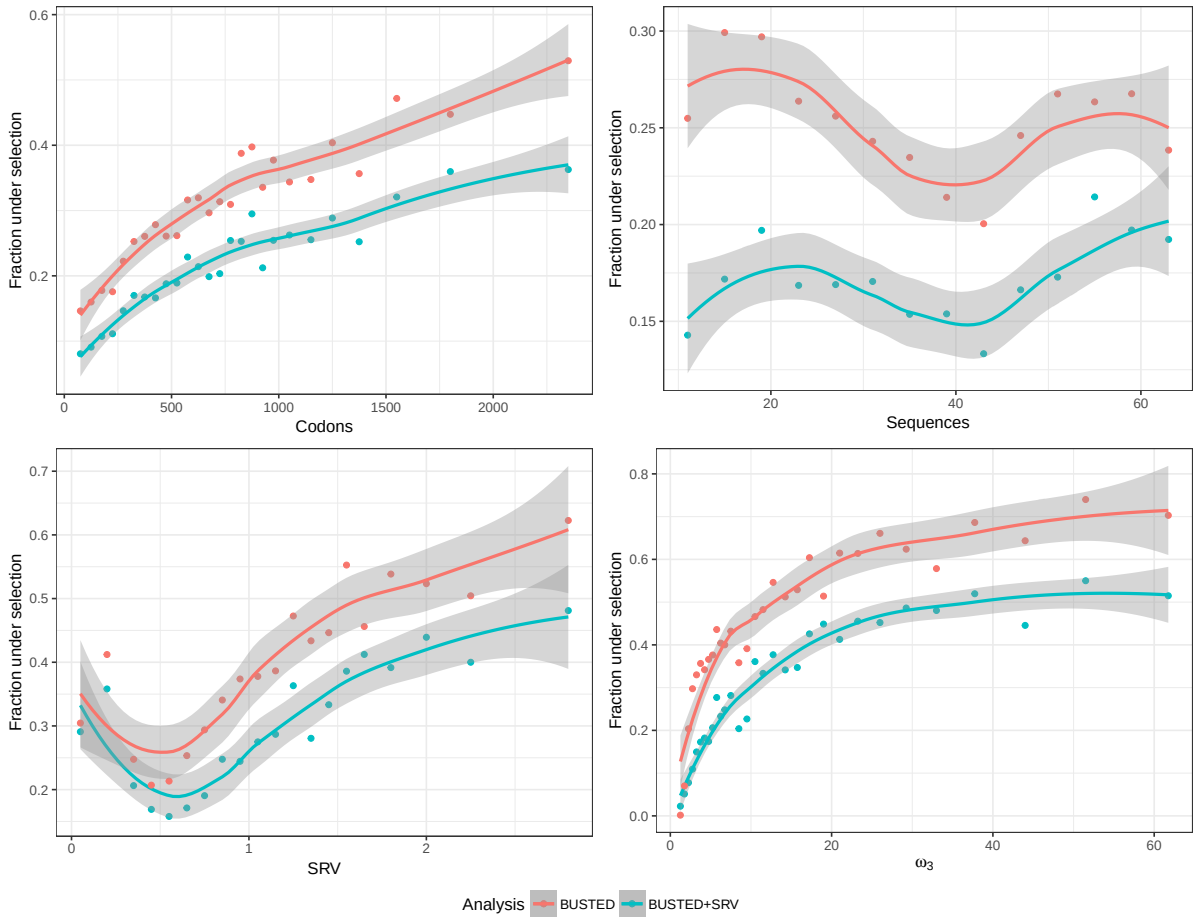


Figure 4.2 Fraction of Alignments Under Selection for the Selectome Data Set. Fraction of alignments under selection ($P \leq 5.6e^{-6}$) versus the median of a sliding window for A) Sequences B) Coefficient of variation of synonymous substitution rates C) Number of Codons D) the ω_3 maximum likelihood estimate according to BUSTED. Lines are Loess fit lines for for BUSTED p-values (red) and BUSTED+SRV p-values (blue).

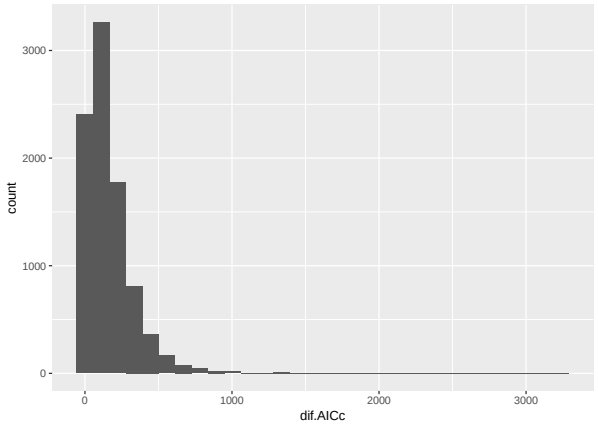


Figure 4.3 Difference of AICc. Histogram of the difference between BUSTED's AICc and BUSTED+SRV's AICc for each data set.

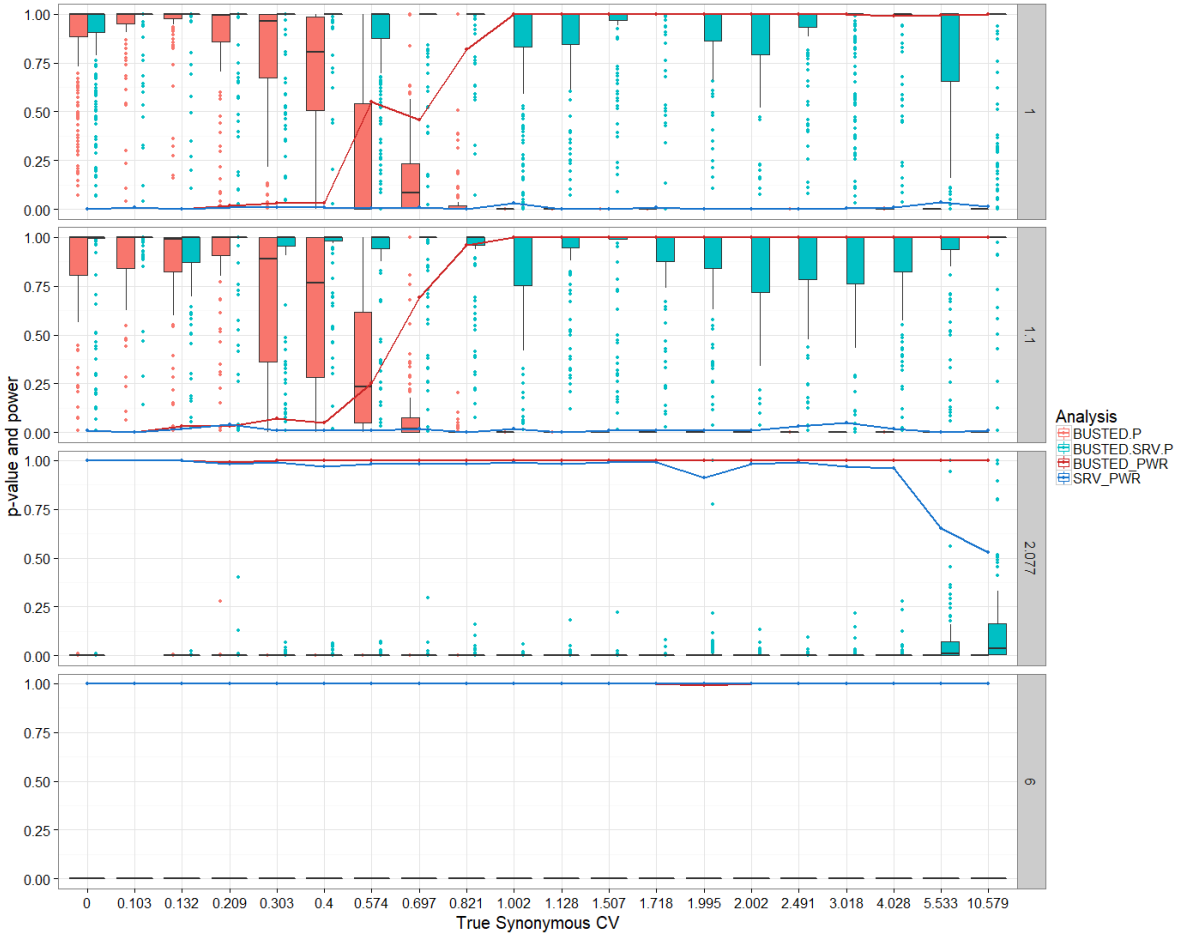


Figure 4.4 Power Curves and P-Value Boxplots for Simulation of a Tree with 33 Sequences and with 5000 Codons Faceted by the True ω_3 Value. The caluculated p-values for each simulation of 100 replicates is represented by the boxplots as the simulated CV value increases. The blue boxplots are the caluculated p-values according to BUSTED+SRV and the red are the caluculated p-values according to BUSTED. The power of each test for each value of synonymous CV is also plotted here for both BUSTED (Red line) and BUSTED+SRV (blue line). The graphs are split by the simulated ω_3.

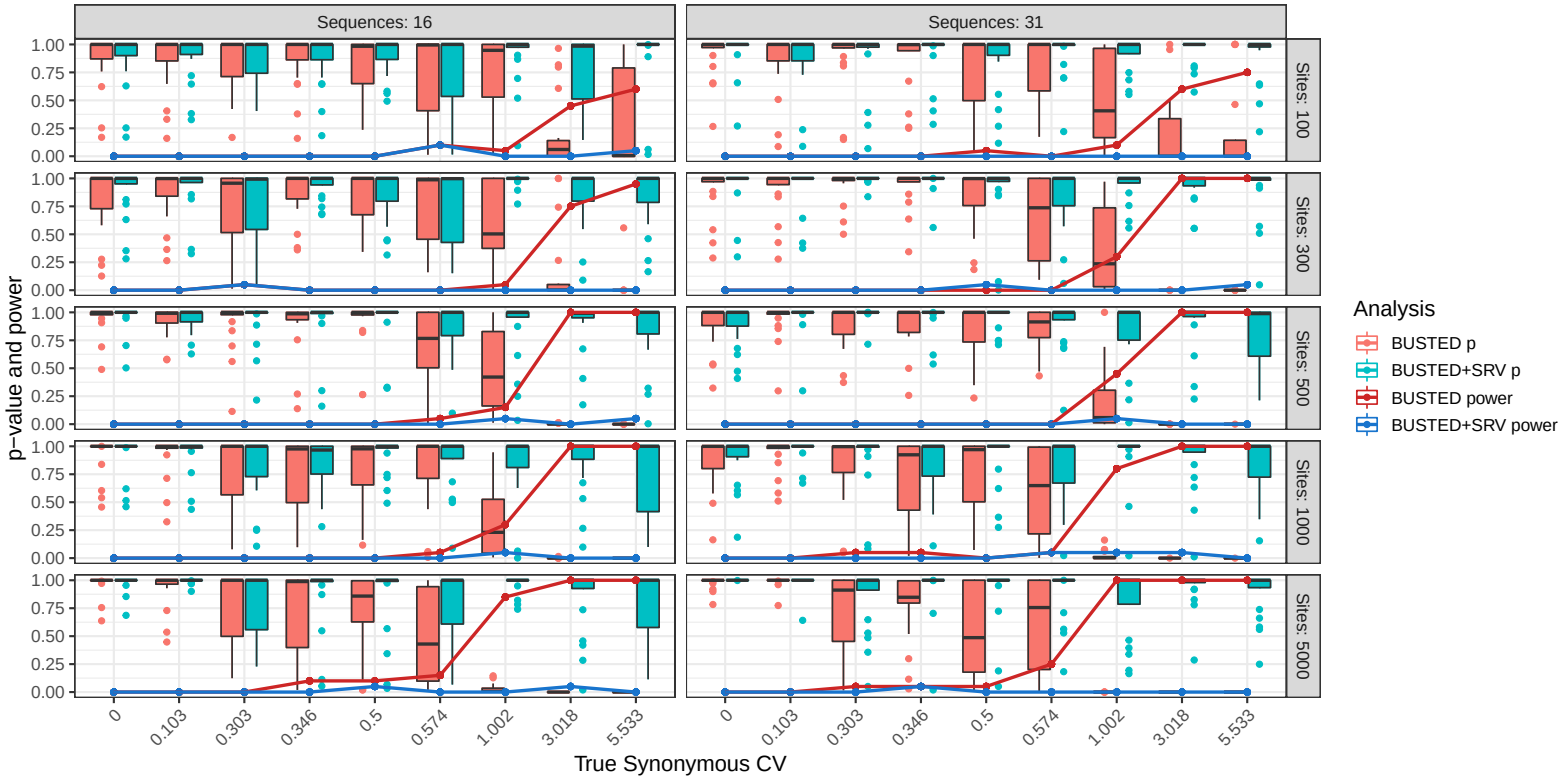


Figure 4.5 Power Curves and P-Value Boxplots for Simulation with $\omega_3 = 1$ for Trees with 16 and 31 Sequences and 100, 300, 500, 1000, 5000 Codons Respectively. The caluculated p-values for each simulation of 100 replicates is represented by the boxplots as the simulated CV value increases. The blue boxplots are the estimates according to BUSTED+SRV and the red are according to BUSTED. The power of each test for each value of synonymous CV is also plotted here for both BUSTED (Red line) and BUSTED+SRV (blue line). The graphs are split horizontally by the number of sequences and vertically by the number of sites.

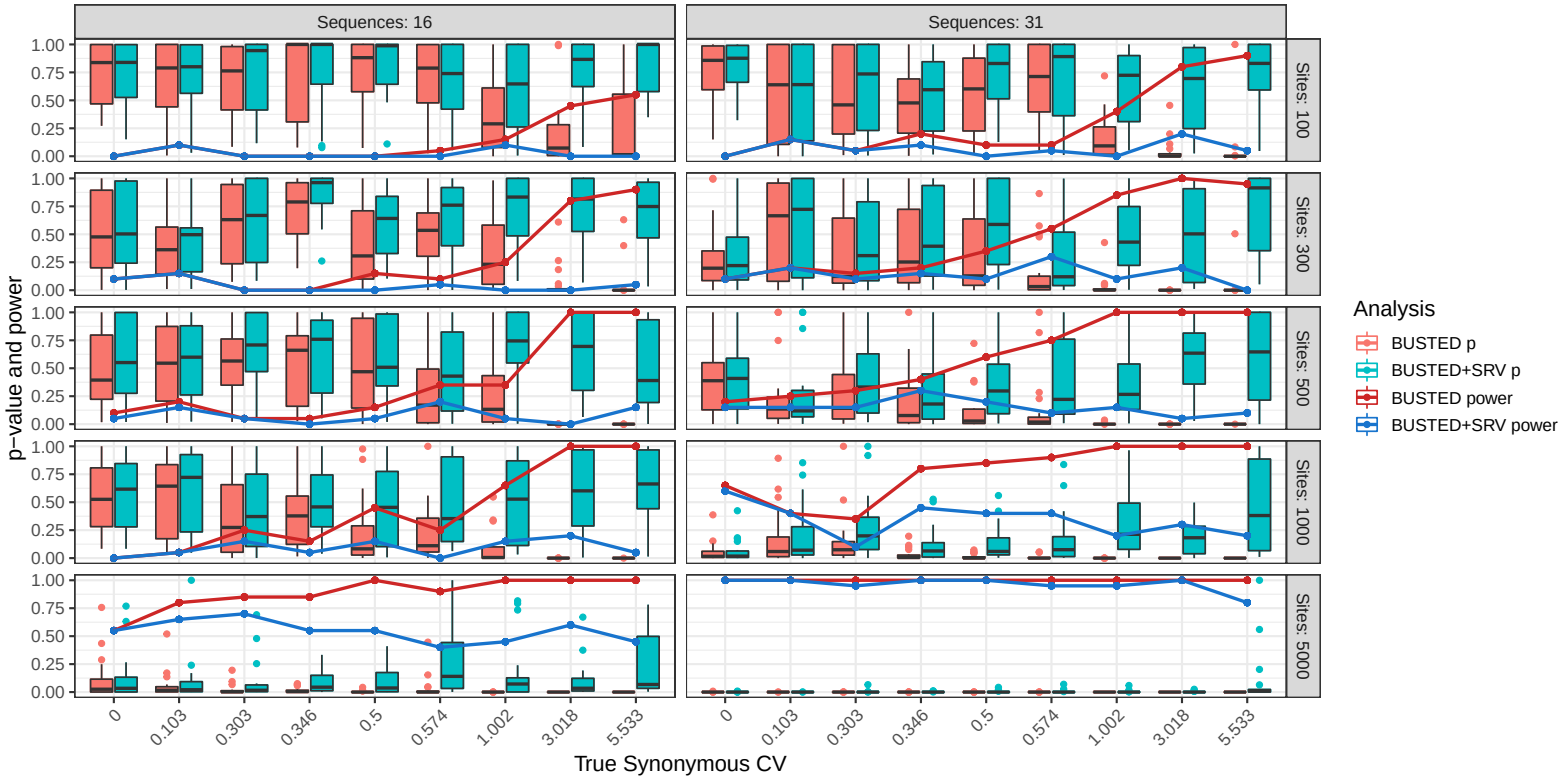


Figure 4.6 Power Curves and P-Value Boxplots for Simulation with $\omega_3 = 2.077$ for Trees with 16 and 31 Sequences and 100, 300, 500, 1000, 5000 Codons Respectively. The caluculated p-values for each simulation of 100 replicates is represented by the boxplots as the simulated CV value increases. The blue boxplots are the estimates according to BUSTED+SRV and the red are according to BUSTED. The power of each test for each value of synonymous CV is also plotted here for both BUSTED (Red line) and BUSTED+SRV (blue line). The graphs are split horizontally by the number of sequences and vertically by the number of sites.

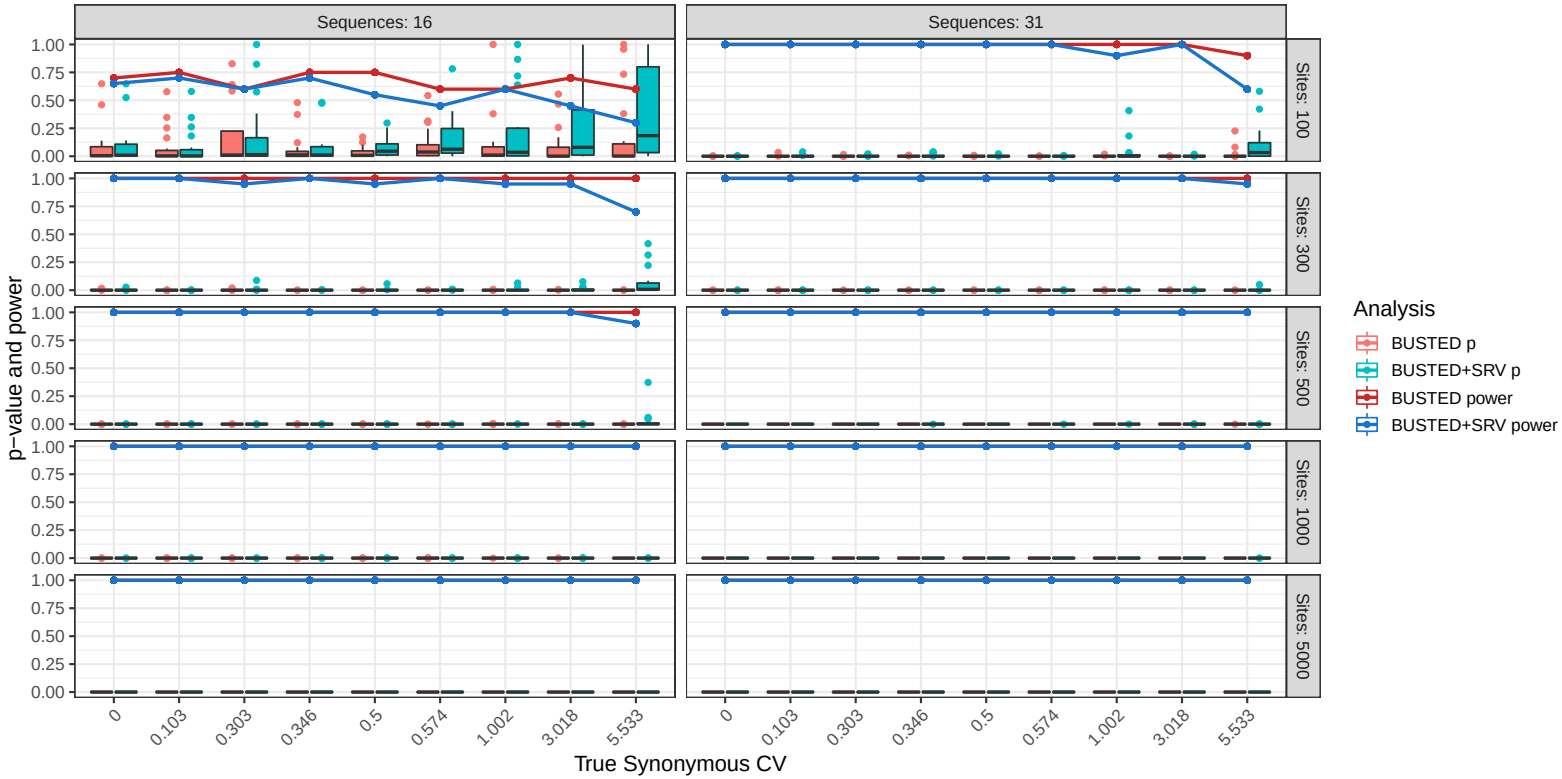


Figure 4.7 Power Curves and P-Value Boxplots for Simulation with $\omega_3 = 6$ for Trees with 16 and 31 Sequences and 100, 300, 500, 1000, 5000 Codons Respectively. The caluculated p-values for each simulation of 100 replicates is represented by the boxplots as the simulated CV value increases. The blue boxplots are the estimates according to BUSTED+SRV and the red are according to BUSTED. The power of each test for each value of synonymous CV is also plotted here for both BUSTED (Red line) and BUSTED+SRV (blue line). The graphs are split horizontally by the number of sequences and vertically by the number of sites.

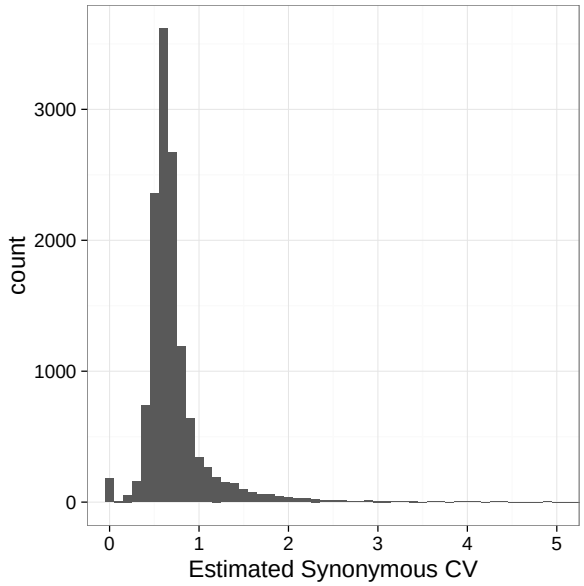


Figure 4.8 Histogram Showing the Range of the Estimated Synonymous CV for the Selectome Data Set. Note that the true maximum synonymous CV for the Selectome data sets was 21.29.

BIBLIOGRAPHY

Agashe, D. et al. (2016). “Large-Effect Beneficial Synonymous Mutations Mediate Rapid and Parallel Adaptation in a Bacterium”. *Molecular Biology and Evolution* **33**.6, pp. 1542–1553.

Bhardwaj, A. (2014). “Investigating the role of site specific synonymous variation in disease association studies.” *Mitochondrion* **16**, pp. 83–8.

Brandis, G. & Hughes, D. (2016). “The Selective Advantage of Synonymous Codon Usage Bias in Salmonella”. *PLoS Genetics* **12**.3. Ed. by Ibba, M., e1005926.

Chamary, J. V. & Hurst, L. D. (2005). “Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals”. en. *Genome Biology* **6**.9, R75.

Chamary, J. V. et al. (2006). “Hearing silence: non-neutral evolution at synonymous sites in mammals”. en. *Nature Reviews Genetics* **7**.2, pp. 98–108.

Dimitrieva, S. & Anisimova, M. (2014). “Unraveling Patterns of Site-to-Site Synonymous Rates Variation and Associated Gene Properties of Protein Domains and Families”. *PLoS ONE* **9**.6. Ed. by Tuller, T., e95034.

Duan, J. et al. (2003). “Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor”. en. *Human Molecular Genetics* **12**.3, pp. 205–216.

Eyre-Walker, A. (1996). “Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy?” en. *Molecular Biology and Evolution* **13**.6, pp. 864–872.

Hurst, L. D. & Pál, C. (2001). “Evidence for purifying selection acting on silent sites in BRCA1”. *Trends in Genetics* **17**.2, pp. 62–65.

Kosakovsky Pond, S. L. & Muse, S. V. (2005). “Site-to-site variation of synonymous substitution rates.” en. *Molecular Biology and Evolution* **22**.12, pp. 2375–2385.

Kosakovsky Pond, S. L. et al. (2005). “HyPhy: Hypothesis testing using phylogenies”. *Bioinformatics* **21**.5, pp. 676–679.

Kubatko, L. et al. (2016). “A codon model of nucleotide substitution with selection on synonymous codon usage”. *Molecular Phylogenetics and Evolution* **94**, pp. 290–297.

Lawrie, D. S. et al. (2013). “Strong Purifying Selection at Synonymous Sites in D. melanogaster”. *PLoS Genetics* **9**.5. Ed. by Plotkin, J. B., e1003527.

Mayrose, I. et al. (2007). “Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates”. *Bioinformatics* **23**.13, pp. i319–i327.

Moretti, S. et al. (2014). “Selectome update: quality control and computational improvements to a database of positive selection.” en. *Nucleic Acids Research* **42**.Database issue, pp. 917–21.

Murrell, B. et al. (2015). “Gene-Wide Identification of Episodic Selection”. en. *Molecular Biology and Evolution* **32**.5, pp. 1365–1371.

Muse, S. V. & Gaut, B. S. (1994). “A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome.” *Molecular Biology and Evolution* **11**.5, pp. 715–724.

Shields, D. C. et al. (1988). “"Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons.” *Molecular biology and evolution* **5**.6, pp. 704–716.

Supek, F. et al. (2014). “Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers”. *Cell* **156**.6, pp. 1324–1335.

Takata, A. et al. (2016). “De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia”. *Neuron* **89**.5, pp. 940–947.

Tavaré, S (1986). *Some probabilistic and statistical problems in the analysis of DNA sequences*.

Yang, Z. (2007). “PAML 4: Phylogenetic Analysis by Maximum Likelihood”. *Molecular Biology and Evolution* **24**.8, pp. 1586–1591.

Zhang, J. et al. (2005). “Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level.” en. *Molecular biology and evolution* **22**.12, pp. 2472–9.

APPENDICES

APPENDIX

A

NCBI ACCESSION NUMBERS

A.1 Description

Table of NCBI accession number for the mitochondrial genome data sets used in chapter 2.

A.2 Table

Table A.1 NCBI Accession numbers. Table listing the NCBI accession number for the mitochondrial data sets as well as the order and species used.

Species	Order	Accession Number
Acipenser_dabryanus	ACIPENSERIFORMES	NC_005451.1

Table A.1 continued

Acipenser_gueldenstaedtii	ACIPENSERIFORMES	NC_012576.1
Acipenser_sinensis	ACIPENSERIFORMES	NC_012646.1
Acipenser_stellatus	ACIPENSERIFORMES	NC_005795.1
Acipenser_transmontanus	ACIPENSERIFORMES	NC_004743.1
Huso_huso	ACIPENSERIFORMES	NC_005252.1
Scaphirhynchus	ACIPENSERIFORMES	NC_030326.1
Polyodon_spathula	ACIPENSERIFORMES	NC_004419.1
Psephurus_gladius	ACIPENSERIFORMES	NC_005834.1
Anas_formosa	ANSERIFORMES	NC_015482.1
Anas_platyrhynchos	ANSERIFORMES	NC_009684.1
Anser_albifrons	ANSERIFORMES	NC_004539.1
Anser_anser	ANSERIFORMES	NC_011196.1
Aythya_americana	ANSERIFORMES	NC_000877.1
Branta_canadensis	ANSERIFORMES	NC_007011.1
Cairina_moschata	ANSERIFORMES	NC_010965.1
Cygnus_atratus	ANSERIFORMES	NC_012843.1
Cygnus_columbianus	ANSERIFORMES	NC_007691.1
Dendrocygna_javanica	ANSERIFORMES	NC_012844.1
Anseranas_semipalmata	ANSERIFORMES	NC_005933.1
Bombina_orientalis	ANURA	NC_006689.1
Discoglossus_galganoi	ANURA	NC_006690.1
Leiopelma_archeyi	ANURA	NC_014691.1
Pelobates_cultripes	ANURA	NC_008144.1
Pipa_carvalhoi	ANURA	NC_015617.1

Table A.1 continued

Pseudhymenochirus_merlini	ANURA	NC_015618.1
Xenopus_laevis	ANURA	NC_001573.1
Rhinophrynus_dorsalis	ANURA	NC_015620.1
Bufo_gargarizans	ANURA	NC_008410.1
Hyla_chinensis	ANURA	NC_006403.1
Microhyla_ornata	ANURA	NC_009422.1
Mantella_madagascariensis	ANURA	NC_007888.1
Euphlyctis_hexadactylus	ANURA	NC_014584.1
Fejervarya_limnocharis	ANURA	NC_005055.1
Limnonectes_fujianensis	ANURA	NC_007440.1
Occidozyga_martensii	ANURA	NC_014685.1
Rana_nigromaculata	ANURA	NC_002805.1
Buergeria_buergeri	ANURA	NC_008975.1
Polypedates_megacephalus	ANURA	NC_032344.1
Rhacophorus_schlegelii	ANURA	NC_007178.1
Habronattus_oregonensis	ARANEAE	NC_005942.1
Hypochilus_thorelli	ARANEAE	NC_010777.1
Nephila_clavata	ARANEAE	NC_008063.1
Heptathela_hangzhouensis	ARANEAE	NC_005924.1
Calisoga_longitarsis	ARANEAE	NC_010780.1
Ornithoctonus_huwena	ARANEAE	NC_005925.1
Anisakis_simplex	ASCARIDIDA	NC_007934.1
Contracaecum_rudolphii	ASCARIDIDA	NC_014870.1
Ascaris_suum	ASCARIDIDA	NC_001327.1

Table A.1 continued

Toxocara_canis	ASCARIDIDA	NC_010690.1
Toxocara_cati	ASCARIDIDA	NC_010773.1
Toxocara_malaysiensis	ASCARIDIDA	NC_010527.1
Xenopoecilus_sarasinorum	BELONIFORMES	NC_011172.1
Oryzias_dancena	BELONIFORMES	NC_012976.1
Oryzias_javanicus	BELONIFORMES	NC_012981.1
Oryzias_latipes	BELONIFORMES	NC_004387.1
Oryzias_luzonensis	BELONIFORMES	NC_012979.1
Oryzias_minutillus	BELONIFORMES	NC_012975.1
Oryzias_sinensis	BELONIFORMES	NC_013434.1
Cypselurus_hiraii	BELONIFORMES	NC_007403.1
Exocoetus_volitans	BELONIFORMES	NC_003184.1
Hyporhamphus_sajori	BELONIFORMES	NC_011173.1
Ablennes_hians	BELONIFORMES	NC_011180.1
Cololabis_saira	BELONIFORMES	NC_003183.1
Beryx_decadactylus	BERYCIFORMES	NC_004393.1
Beryx_mollis	BERYCIFORMES	NC_013845.1
Beryx_splendens	BERYCIFORMES	NC_003188.1
Myripristis_berndti	BERYCIFORMES	NC_003189.1
Ostichthys_japonicus	BERYCIFORMES	NC_004394.1
Sargocentron_rubrum	BERYCIFORMES	NC_004395.1
Anomalops_katoptron	BERYCIFORMES	NC_008128.1
Anoplogaster_cornuta	BERYCIFORMES	NC_004391.1
Diretmoides_veriginae	BERYCIFORMES	NC_008126.1

Table A.1 continued

Diretmus_argenteus	BERYCIFORMES	NC_008127.1
Hoplostethus_japonicus	BERYCIFORMES	NC_003187.1
Monocentris_japonicus	BERYCIFORMES	NC_004392.1
Arctocephalus_forsteri	CARNIVORES	NC_004023.1
Canis_lupus	CARNIVORES	NA
Eumetopias_jubatus	CARNIVORES	NC_004030.2
Felis_catus	CARNIVORES	NC_001700.1
Halichoerus_grypus	CARNIVORES	NC_001602.1
Odobenus_rosmarus_rosmarus	CARNIVORES	NC_004029.2
Phoca_vitulina	CARNIVORES	NC_001325.1
Ursus_americanus	CARNIVORES	NC_003426.1
Ursus_arctos	CARNIVORES	NC_003427.1
Ursus_maritimus	CARNIVORES	NC_015924.1
Eubalaena_australis	CETACEA	NC_006930.1
Balaenoptera_borealis	CETACEA	NC_006929.1
Balaenoptera_edeni	CETACEA	NC_007938.1
Balaenoptera_physalus	CETACEA	NC_001321.1
Megaptera_novaeangliae	CETACEA	NC_006927.1
Eschrichtius_robustus	CETACEA	NC_005270.1
Caperea_marginata	CETACEA	NC_005269.1
Grampus_griseus	CETACEA	NC_012062.1
Orcinus_orca	CETACEA	NC_023889.1
Stenella_attenuata	CETACEA	NC_012051.1
Tursiops_aduncus	CETACEA	NC_012058.1

Table A.1 continued

Inia_geoffrensis	CETACEA	NC_005276.1
Lipotes_vexillifer	CETACEA	NC_007629.1
Monodon_monoceros	CETACEA	NC_005279.1
Phocoena_phocoena	CETACEA	NC_005280.1
Kogia_breviceps	CETACEA	NC_005272.1
Platanista_minor	CETACEA	NC_005275.1
Pontoporia_blainvillei	CETACEA	NC_005277.1
Berardius_bairdii	CETACEA	NC_005274.1
Acestrorhynchus_sp_NM_2010	CHARACIFORMES	NC_015755.1
Micralestes_sp_NM_2010	CHARACIFORMES	NC_015753.1
Phenacogrammus_interruptus	CHARACIFORMES	NC_004699.1
Chalceus_macrolepidotus	CHARACIFORMES	NC_004700.1
Hydrolycus_scomberoides	CHARACIFORMES	NC_015813.1
Myleus_sp_NM_2010	CHARACIFORMES	NC_015751.1
Pygocentrus_nattereri	CHARACIFORMES	NC_015840.1
Chilodus_punctatus	CHARACIFORMES	NC_015801.1
Citharinus_congicus	CHARACIFORMES	NC_015805.1
Ichthyborus_sp_NM_2010	CHARACIFORMES	NC_015752.1
Hemiodopsis_gracilis	CHARACIFORMES	NC_015816.1
Apareiodon_affinis	CHARACIFORMES	NC_015834.1
Hepsetus_odoe	CHARACIFORMES	NC_015819.1
Lebiasina_astrigata	CHARACIFORMES	NC_015750.1
Distichodus_sexfasciatus	CHARACIFORMES	NC_015836.1
Pteropus_dasymallus	CHIROPTERA	NC_002612.1

Table A.1 continued

Pteropus_scapulatus	CHIROPTERA	NC_002619.1
Rousettus_aegyptiacus	CHIROPTERA	NC_007393.1
Artibeus_jamaicensis	CHIROPTERA	NC_002009.1
Chalinolobus_tuberculatus	CHIROPTERA	NC_002626.1
Mystacina_tuberculata	CHIROPTERA	NC_006925.1
Pipistrellus_abramus	CHIROPTERA	NC_005436.1
Plecotus_auritus	CHIROPTERA	NC_015484.1
Rhinolophus_formosae	CHIROPTERA	NC_011304.1
Rhinolophus_monoceros	CHIROPTERA	NC_005433.1
Rhinolophus_pumilus	CHIROPTERA	NC_005434.1
Ardea_novaehollandiae	CICONIIFORMES	NC_008551.1
Egretta_eulophotes	CICONIIFORMES	NC_009736.1
Ixobrychus_cinnamomeus	CICONIIFORMES	NC_015077.1
Cathartes_aura	CICONIIFORMES	NC_007628.1
Ciconia_boyciana	CICONIIFORMES	NC_002196.1
Ciconia_ciconia	CICONIIFORMES	NC_002197.1
Nipponia_nippon	CICONIIFORMES	NC_008132.1
Platalea_leucorodia	CICONIIFORMES	NC_012772.1
Platalea_minor	CICONIIFORMES	NC_010962.1
Threskiornis_aethiopicus	CICONIIFORMES	NC_013146.1
Chirocentrus_dorab	CLUPEIFORMES	NC_006913.1
Alosa_pseudoharengus	CLUPEIFORMES	NC_009576.1
Brevoortia_tyrannus	CLUPEIFORMES	NC_014266.1
Clupeonella_cultriventris	CLUPEIFORMES	NC_015109.1

Table A.1 continued

Dorosoma_cepedianum	CLUPEIFORMES	NC_008107.1
Ethmalosa_fimbriata	CLUPEIFORMES	NC_009582.1
Jenkinsia_lamprotaenia	CLUPEIFORMES	NC_006917.1
Nematalosa_japonica	CLUPEIFORMES	NC_009586.1
Odaxothrissa_vittata	CLUPEIFORMES	NC_009590.1
Pellonula_leonensis	CLUPEIFORMES	NC_009591.1
Sardinella_maderensis	CLUPEIFORMES	NC_009587.1
Sardinops_melanostictus	CLUPEIFORMES	NC_002616.1
Spratelloides_gracilis	CLUPEIFORMES	NC_009589.1
Sprattus_sprattus	CLUPEIFORMES	NC_009593.1
Denticeps_clupeoides	CLUPEIFORMES	NC_007889.1
Amazonsprattus_scintilla	CLUPEIFORMES	NC_014265.1
Anchoviella	CLUPEIFORMES	NC_014269.1
Coilia_reynaldi	CLUPEIFORMES	NC_014276.1
Engraulis_encrasicolus	CLUPEIFORMES	NC_009581.1
Lycengraulis_grossidens	CLUPEIFORMES	NC_014279.1
Lycothrissa_crocodilus	CLUPEIFORMES	NC_014277.1
Thryssa_baelama	CLUPEIFORMES	NC_014264.1
Ilisha_elongata	CLUPEIFORMES	NC_009585.1
Pellona_flavipinnis	CLUPEIFORMES	NC_014268.1
Sundasalanx_mekongensis	CLUPEIFORMES	NC_006919.1
Cryptopygus_antarcticus	COLLEMBOLA	NC_010533.1
Orchesella_villosa	COLLEMBOLA	NC_010534.1
Bilobella_aurantiaca	COLLEMBOLA	NC_011195.1

Table A.1 continued

Friesea_grisea	COLLEMBOLA	NC_010535.1
Gomphiocephalus_hodgsoni	COLLEMBOLA	NC_005438.1
Onychiurus_orientalis	COLLEMBOLA	NC_006074.1
Podura_aquatica	COLLEMBOLA	NC_006075.1
Tetrodontophora_bielanensis	COLLEMBOLA	NC_002735.1
Sminthurus_viridis	COLLEMBOLA	NC_010536.1
Aplocheilus_panchax	CYPRINODONTIFORMES	NC_011176.1
Kryptolebias_marmoratus	CYPRINODONTIFORMES	NC_003290.1
Nothobranchius_furzeri	CYPRINODONTIFORMES	NC_011814.1
Cyprinodon_rubrofluviatilis	CYPRINODONTIFORMES	NC_009125.2
Fundulus_diaphanus	CYPRINODONTIFORMES	NC_012361.1
Fundulus_grandis	CYPRINODONTIFORMES	NC_012377.1
Fundulus_heteroclitus	CYPRINODONTIFORMES	NC_012312.1
Fundulus_olivaceus	CYPRINODONTIFORMES	NC_011380.1
Gambusia_affinis	CYPRINODONTIFORMES	NC_004388.1
Jordanella_floridae	CYPRINODONTIFORMES	NC_011387.1
Xenotoca_eiseni	CYPRINODONTIFORMES	NC_011381.1
Xiphophorus_hellerii	CYPRINODONTIFORMES	NC_013089.1
Xiphophorus_maculatus	CYPRINODONTIFORMES	NC_011379.1
Dasyurus_hallucatus	DASYUROMORPHIA	NC_007630.1
Phascogale_tapoatafa	DASYUROMORPHIA	NC_006523.1
Sminthopsis_crassicaudata	DASYUROMORPHIA	NC_007631.1
Sminthopsis_douglasi	DASYUROMORPHIA	NC_006517.1
Myrmecobius_fasciatus	DASYUROMORPHIA	NC_011949.1

Table A.1 continued

Thylacinus_cynocephalus	DASYUROMORPHIA	NC_011944.1
Farfantepenaeus_californiensis	DECAPODA	NC_012738.1
Fenneropenaeus_chinensis	DECAPODA	NC_009679.1
Litopenaeus_stylirostris	DECAPODA	NC_012060.1
Marsupenaeus_japonicus	DECAPODA	NC_007010.1
Penaeus_monodon	DECAPODA	NC_002184.1
Panulirus_stimpsoni	DECAPODA	NC_014339.1
Shinkaia_crosnieri	DECAPODA	NC_011013.1
Pagurus_longicarpus	DECAPODA	NC_003058.1
Homarus_americanus	DECAPODA	NC_015607.1
Cherax_destructor	DECAPODA	NC_011243.1
Gandalfus_yunohana	DECAPODA	NC_013713.1
Callinectes_sapidus	DECAPODA	NC_006281.1
Charybdis_japonica	DECAPODA	NC_013246.1
Portunus_trituberculatus	DECAPODA	NC_005037.1
Scylla_serrata	DECAPODA	NC_012565.1
Scylla_tranquebarica	DECAPODA	NC_012567.1
Geothelphusa_dehaani	DECAPODA	NC_007379.1
Pseudocarcinus_gigas	DECAPODA	NC_006891.1
Eriocheir_sinensis	DECAPODA	NC_006992.1
Xenograpsus_testudinatus	DECAPODA	NC_013480.1
Alpheus_distinguendus	DECAPODA	NC_014883.1
Halocaridina_rubra	DECAPODA	NC_008413.1
Exopalaemon_carinicauda	DECAPODA	NC_012566.1

Table A.1 continued

Macrobrachium_rosenbergii	DECAPODA	NC_006880.1
Distoechurus_pennatus	DIPROTODONTIA	NC_008145.1
Lagorchestes_hirsutus	DIPROTODONTIA	NC_008136.1
Lagostrophus_fasciatus	DIPROTODONTIA	NC_008447.1
Macropus_robustus	DIPROTODONTIA	NC_001794.1
Dactylopsila_trivirgata	DIPROTODONTIA	NC_008134.1
Petaurus_breviceps	DIPROTODONTIA	NC_008135.1
Phalanger_interpositus	DIPROTODONTIA	NC_008137.1
Trichosurus_vulpecula	DIPROTODONTIA	NC_003039.1
Phascolarctos_cinereus	DIPROTODONTIA	NC_008133.1
Potorous_tridactylus	DIPROTODONTIA	NC_006524.1
Pseudocheirus_peregrinus	DIPROTODONTIA	NC_006519.1
Tarsipes_rostratus	DIPROTODONTIA	NC_006518.1
Vombatus_ursinus	DIPROTODONTIA	NC_003322.1
Bactrocera_carambolae	DIPTERA	NC_009772.1
Bactrocera_minax	DIPTERA	NC_014402.1
Bactrocera_tryoni	DIPTERA	NC_014611.1
Ceratitis_capitata	DIPTERA	NC_000857.1
Chrysomya_putoria	DIPTERA	NC_002697.1
Cochliomyia_hominivorax	DIPTERA	NC_002660.1
Cydistomyia_duplonotata	DIPTERA	NC_008756.1
Dermatobia_hominis	DIPTERA	NC_006378.1
Drosophila_mauritiana	DIPTERA	NC_005779.1
Drosophila_melanogaster	DIPTERA	NA

Table A.1 continued

Drosophila_simulans	DIPTERA	NC_005781.1
Exorista_sorbillans	DIPTERA	NC_014704.1
Haematobia_irritans_irritans	DIPTERA	NC_007102.1
Hypoderma_lineatum	DIPTERA	NC_013932.1
Liriomyza_trifolii	DIPTERA	NC_014283.1
Lucilia_sericata	DIPTERA	NC_009733.1
Simosyrphus_grandicornis	DIPTERA	NC_008754.1
Trichophthalma_punctata	DIPTERA	NC_008755.1
Aedes_aegypti	DIPTERA	NC_010241.1
Anopheles_darlingi	DIPTERA	NC_014275.1
Anopheles_quadrimaculatus	DIPTERA	NC_000875.1
Culex_quinquefasciatus	DIPTERA	NC_014574.1
Culicoides_arakawae	DIPTERA	NC_009809.1
Mayetiola_destructor	DIPTERA	NC_013066.1
Rhopalomyia_pomum	DIPTERA	NC_013063.1
Accipiter_gentilis	FALCONIFORMES	NC_011818.1
Buteo_buteo	FALCONIFORMES	NC_003128.3
Nisaetus_alboniger	FALCONIFORMES	NC_007599.1
Nisaetus_nipalensis	FALCONIFORMES	NC_007598.1
Pandion_haliaetus	FALCONIFORMES	NC_008550.1
Falco_peregrinus	FALCONIFORMES	NC_000878.1
Falco_sparverius	FALCONIFORMES	NC_008547.1
Falco_tinnunculus	FALCONIFORMES	NC_011307.1
Micrastur_gilvicollis	FALCONIFORMES	NC_008548.1

Table A.1 continued

Bregmaceros_nectabanus	GADIFORMES	NC_008124.1
Arctogadus_glacialis	GADIFORMES	NC_010122.1
Boreogadus_saida	GADIFORMES	NC_010121.1
Gadus_chalcogrammus	GADIFORMES	NC_004449.1
Gadus_morhua	GADIFORMES	NC_002081.1
Gadus_ogac	GADIFORMES	NC_012323.1
Melanogrammus_aeglefinus	GADIFORMES	NC_007396.1
Merlangius_merlangus	GADIFORMES	NC_007395.1
Micromesistius_poutassou	GADIFORMES	NC_015102.1
Pollachius_pollachius	GADIFORMES	NC_015097.1
Pollachius_virens	GADIFORMES	NC_015094.1
Lota_lota	GADIFORMES	NC_004379.1
Bathygadus_antrodes	GADIFORMES	NC_008222.1
Coelorinchus_kishinouyei	GADIFORMES	NC_003169.1
Ventrifossa_garmani	GADIFORMES	NC_008225.1
Squalogadus_modificatus	GADIFORMES	NC_008223.1
Trachyrincus_murrayi	GADIFORMES	NC_008224.1
Merluccius_merluccius	GADIFORMES	NC_015120.1
Physiculus_japonicus	GADIFORMES	NC_004377.1
Alectura_lathami	GALLIFORMES	NC_007227.1
Acryllium_vulturinum	GALLIFORMES	NC_014180.1
Numida_meleagris	GALLIFORMES	NC_006382.1
Meleagris_gallopavo	GALLIFORMES	NC_010195.2
Arborophila_rufipectus	GALLIFORMES	NC_012453.1

Table A.1 continued

Bambusicola_thoracica	GALLIFORMES	NC_011816.1
Coturnix_japonica	GALLIFORMES	NC_003408.1
Francolinus_pintadeanus	GALLIFORMES	NC_011817.1
Chrysolophus_pictus	GALLIFORMES	NC_014576.1
Gallus_gallus_bankiva	GALLIFORMES	NA
Gallus_varius	GALLIFORMES	NC_007238.1
Lophophorus_lhuysii	GALLIFORMES	NC_013979.1
Lophura_ignita	GALLIFORMES	NC_010781.1
Pavo_muticus	GALLIFORMES	NC_012897.1
Phasianus_colchicus	GALLIFORMES	NC_015526.1
Polyplectron_bicalcaratum	GALLIFORMES	NC_012900.1
Syrmaticus_humiae	GALLIFORMES	NC_010774.1
Syrmaticus_soemmerringi_ijimae	GALLIFORMES	NC_010767.1
Tragopan_caboti	GALLIFORMES	NC_013619.1
Aulichthys_japonicus	GASTEROSTEIFORMES	NC_011569.1
Aulorhynchus_flavidus	GASTEROSTEIFORMES	NC_010268.1
Aulostomus_chinensis	GASTEROSTEIFORMES	NC_010269.1
Apeltes_quadracus	GASTEROSTEIFORMES	NC_011580.1
Culaea_inconstans	GASTEROSTEIFORMES	NC_011577.1
Gasterosteus_aculeatus	GASTEROSTEIFORMES	NA
Gasterosteus_wheatlandi	GASTEROSTEIFORMES	NC_011570.1
Pungitius_kaibarae	GASTEROSTEIFORMES	NC_014893.1
Pungitius_pungitius	GASTEROSTEIFORMES	NC_011571.1
Pungitius sinensis	GASTEROSTEIFORMES	NA

Table A.1 continued

Spinachia_spinachia	GASTEROSTEIFORMES	NC_011582.1
Hypoptychus_dybowskii	GASTEROSTEIFORMES	NC_004400.1
Otis_tarda	GRUIFORMES	NC_014046.1
Coturnicops_exquisitus	GRUIFORMES	NC_012143.1
Gallinula_chloropus	GRUIFORMES	NC_015236.1
Gallirallus_okinawae	GRUIFORMES	NC_012140.1
Porphyrio_hochstetteri	GRUIFORMES	NC_010092.1
Rallina_eurizonoides_sepiaria	GRUIFORMES	NC_012142.1
Rhynochetos_jubatus	GRUIFORMES	NC_010091.1
Gegeneophis_ramaswamii	GYMNOPHIONA	NC_006301.1
Siphonops_annulatus	GYMNOPHIONA	NC_007911.1
Ichthyophis_bannanicus	GYMNOPHIONA	NC_006404.1
Ichthyophis_glutinosus	GYMNOPHIONA	NC_006302.1
Rhinatrema_bivittatum	GYMNOPHIONA	NC_006303.1
Scolecomorphus_vittatus	GYMNOPHIONA	NC_006304.1
Typhlonectes_natans	GYMNOPHIONA	NC_002471.1
Uraeotyphlus_cf_oxyurus_MW_212	GYMNOPHIONA	NC_006305.1
Abidama_producta	HEMIPTERA	NC_015799.1
Homalodisca_vitripennis	HEMIPTERA	NC_006899.1
Geisha_distinctissima	HEMIPTERA	NC_012617.1
Hydrometra_sp_NKMT020	HEMIPTERA	NC_012842.1
Orius_niger	HEMIPTERA	NC_012429.1
Triatoma_dimidiata	HEMIPTERA	NC_002609.1
Saldula_arsenjevi	HEMIPTERA	NC_012463.1

Table A.1 continued

Laccotrephes_robustus	HEMIPTERA	NC_012817.1
Ochterus_marginatus	HEMIPTERA	NC_012820.1
Neuroctenus_parus	HEMIPTERA	NC_012459.1
Aeschyntelus_notatus	HEMIPTERA	NC_012446.1
Yemmalysus_parallelus	HEMIPTERA	NC_012464.1
Nezara_viridula	HEMIPTERA	NC_011755.1
Coptosoma_bifaria	HEMIPTERA	NC_012449.1
Physopelta_gutta	HEMIPTERA	NC_012432.1
Aleurochiton_aceris	HEMIPTERA	NC_006160.1
Trialeurodes_vaporariorum	HEMIPTERA	NC_006280.1
Schizaphis_graminum	HEMIPTERA	NC_006158.1
Pachypsylla_venusta	HEMIPTERA	NC_006157.1
Abispa_ephippium	HYMENOPTERA	NC_011520.1
Apis_cerana	HYMENOPTERA	NC_014295.1
Apis mellifera ligustica	HYMENOPTERA	NC_001566.1
Bombus_hypocrita_sapporoensis	HYMENOPTERA	NC_011923.1
Bombus_ignitus	HYMENOPTERA	NC_010967.1
Cotesia_vestalis	HYMENOPTERA	NC_014272.1
Diadegma_semiclausum	HYMENOPTERA	NC_012708.1
Evania_appendigaster	HYMENOPTERA	NC_013238.1
Melipona_bicolor	HYMENOPTERA	NC_004529.1
Pristomyrmex_punctatus	HYMENOPTERA	NC_015075.1
Radoszkowskius_oculata	HYMENOPTERA	NC_014485.1
Solenopsis_geminata	HYMENOPTERA	NC_014669.1

Table A.1 continued

Solenopsis_invicta	HYMENOPTERA	NC_014672.1
Solenopsis_richteri	HYMENOPTERA	NC_014677.1
Spathius_agrili	HYMENOPTERA	NC_014278.1
Vanhornia eucnemidaru	HYMENOPTERA	NC_008323.1
Cephus_cinctus	HYMENOPTERA	NC_012688.1
Orussus_occidentalis	HYMENOPTERA	NC_012689.1
Erinaceus_europaeus	INSECTIVORA	NC_002080.2
Hemiechinus_auritus	INSECTIVORA	NC_005033.1
Echinosorex_gymnura	INSECTIVORA	NC_002808.1
Hylomys_suillus	INSECTIVORA	NC_010298.1
Crocidura_russula	INSECTIVORA	NC_006893.1
Episoriculus_fumidus	INSECTIVORA	NC_003040.1
Sorex_unguiculatus	INSECTIVORA	NC_005435.1
Galemys_pyrenaicus	INSECTIVORA	NC_008156.1
Mogera_wogura	INSECTIVORA	NC_005035.1
Talpa_europaea	INSECTIVORA	NC_002391.1
Urotrichus_talpoides	INSECTIVORA	NC_005034.1
Carios_capensis	IXODIDA	NC_005291.1
Ornithodoros_moubata	IXODIDA	NC_004357.1
Ornithodoros_porcinus	IXODIDA	NC_005820.1
Amblyomma_triguttatum	IXODIDA	NC_005963.1
Haemaphysalis_flava	IXODIDA	NC_005292.1
Ixodes_hexagonus	IXODIDA	NC_002010.1
Ixodes_holocyclus	IXODIDA	NC_005293.1

Table A.1 continued

Ixodes_persulcatus	IXODIDA	NC_004370.1
Ixodes_uriae	IXODIDA	NC_006078.1
Rhipicephalus_sanguineus	IXODIDA	NC_002074.1
Lepus_capensis	LAGOMORPHA	NC_015841.1
Lepus_europaeus	LAGOMORPHA	NC_004028.1
Oryctolagus_cuniculus	LAGOMORPHA	NC_001913.1
Ochotona_collaris	LAGOMORPHA	NC_003033.1
Ochotona_curzoniae	LAGOMORPHA	NC_011029.1
Ochotona_princeps	LAGOMORPHA	NC_005358.1
Adoxophyes_honmai	LEPIDOPTERA	NC_008141.1
Grapholita_molesta	LEPIDOPTERA	NC_014806.1
Spilonota_lechriaspis	LEPIDOPTERA	NC_014294.1
Acraea_issoria	LEPIDOPTERA	NC_013604.1
Apatura_metis	LEPIDOPTERA	NC_015537.1
Calinaga_davidis	LEPIDOPTERA	NC_015480.1
Chilo_suppressalis	LEPIDOPTERA	NC_015612.1
Coreana_raphaelis	LEPIDOPTERA	NC_007976.1
Diatraea_saccharalis	LEPIDOPTERA	NC_013274.1
Eriogyna_pyretorum	LEPIDOPTERA	NC_012727.1
Helicoverpa_armigera	LEPIDOPTERA	NC_014668.1
Hipparchia_autonoe	LEPIDOPTERA	NC_014587.1
Hyphantria_cunea	LEPIDOPTERA	NC_014058.1
Lymantria_dispar	LEPIDOPTERA	NC_012893.1
Manduca_sexta	LEPIDOPTERA	NC_010266.1

Table A.1 continued

Ochrogaster_lunifer	LEPIDOPTERA	NC_011128.1
Ostrinia_nubilalis	LEPIDOPTERA	NC_003367.1
Papilio_maraho	LEPIDOPTERA	NC_014055.1
Parnassius_bremeri	LEPIDOPTERA	NC_014053.1
Phthonandria_atrilineata	LEPIDOPTERA	NC_010522.1
Pieris_melete	LEPIDOPTERA	NC_010568.1
Saturnia_boisduvalii	LEPIDOPTERA	NC_010613.1
Teinopalpus_aureus	LEPIDOPTERA	NC_014398.1
Antheraea_yamamai	LEPIDOPTERA	NC_012739.1
Tetrabrachium_ocellatum	LOPHIIFORMES	NC_013879.1
Caulophryne_pelagica	LOPHIIFORMES	NC_016020.1
Ceratias_uranoscopus	LOPHIIFORMES	NC_013882.1
Cryptopsaras_couesii	LOPHIIFORMES	NC_013880.1
Bufoceratias_thele	LOPHIIFORMES	NC_013869.1
Diceratias_pileatus	LOPHIIFORMES	NC_013870.1
Gigantactis_vanhoeffeni	LOPHIIFORMES	NC_013885.1
Rhynchactis_macrothrix	LOPHIIFORMES	NC_013863.1
Himantolophus_albinares	LOPHIIFORMES	NC_013867.1
Himantolophus_groenlandicus	LOPHIIFORMES	NC_013868.1
Haplophryne_mollis	LOPHIIFORMES	NC_013865.1
Melanocetus_johnsoni	LOPHIIFORMES	NC_013866.1
Melanocetus_murrayi	LOPHIIFORMES	NC_004384.1
Neoceratias_spinifer	LOPHIIFORMES	NC_013864.1
Oneirodes_thompsoni	LOPHIIFORMES	NC_013871.1

Table A.1 continued

Thaumatichthys_pagidostomus	LOPHIIFORMES	NC_013875.1
Chaunax_abei	LOPHIIFORMES	NC_004381.1
Chaunax_pictus	LOPHIIFORMES	NC_013883.1
Chaunax_tosaensis	LOPHIIFORMES	NC_004382.1
Lophiodes_caulinaris	LOPHIIFORMES	NC_013872.1
Lophiomus_setigerus	LOPHIIFORMES	NC_008125.1
Lophius_americanus	LOPHIIFORMES	NC_004380.1
Sladenia_gardineri	LOPHIIFORMES	NC_013873.1
Coelophrys_brevicaudata	LOPHIIFORMES	NC_013886.1
Agamermis_sp_BH_2006	MERMITHIDA	NC_008231.1
Hexamermis_agrotis	MERMITHIDA	NC_008828.1
Romanomermis_culicivorax	MERMITHIDA	NC_008640.1
Romanomermis_iyengari	MERMITHIDA	NC_008693.1
Romanomermis_nielseni	MERMITHIDA	NC_008692.1
Strelkovimermis_spiculatus	MERMITHIDA	NC_008047.1
Thaumamermis_cosgrovei	MERMITHIDA	NC_008046.1
Ascaloptynx_appendiculatus	NEUROPTERA	NC_011277.1
Libelloides_macaronius	NEUROPTERA	NC_015609.1
Apochrysa_matsumurae	NEUROPTERA	NC_015095.1
Chrysoperla_nipponensis	NEUROPTERA	NC_015093.1
Ditaxis_biseriata	NEUROPTERA	NC_013257.1
Polystoechotes_punctatus	NEUROPTERA	NC_011278.1
Cataetyx_rubrirostris	OPHIDIIFORMES	NC_004375.1
Diplacanthopoma_brachysoma	OPHIDIIFORMES	NC_004376.1

Table A.1 continued

Carapus_bermudensis	OPHIDIIFORMES	NC_004373.1
Bassozetus_zenkevitchi	OPHIDIIFORMES	NC_004374.1
Lamprogrammus_niger	OPHIDIIFORMES	NC_004378.1
Sirembo_imberbis	OPHIDIIFORMES	NC_008123.1
Calliptamus_italicus	ORTHOPTERA	NC_011305.1
Ellipes_minuta	ORTHOPTERA	NC_014488.1
Euchorthippus_fusigeniculatus	ORTHOPTERA	NC_014449.1
Gastrimargus_marmoratus	ORTHOPTERA	NC_011114.1
Gomphocerus_sibiricus_tibetanus	ORTHOPTERA	NC_015478.1
Locusta_migratoria_tibetensis	ORTHOPTERA	NC_015624.1
Mekongiana_xiangchengensis	ORTHOPTERA	NC_014450.1
Oedaleus_decorus_asiaticus	ORTHOPTERA	NC_011115.1
Ognevia_longipennis	ORTHOPTERA	NC_013701.1
Oxya_chinensis	ORTHOPTERA	NC_010219.1
Phlaeoba_albonema	ORTHOPTERA	NC_011827.1
Physemacris_variolosa	ORTHOPTERA	NC_014491.1
Schistocerca_gregaria_gregaria	ORTHOPTERA	NC_013240.1
Thrinchus_schrenkii	ORTHOPTERA	NC_014610.1
Traulia_szetschuanensis	ORTHOPTERA	NC_013826.1
Xyleus_modestus	ORTHOPTERA	NC_014490.1
Anabrus_simplex	ORTHOPTERA	NC_009967.1
Deracantha_onos	ORTHOPTERA	NC_011813.1
Elimaea_cheni	ORTHOPTERA	NC_014289.1
Gampsocleis_gratiosa	ORTHOPTERA	NC_011200.1

Table A.1 continued

Gryllotalpa_orientalis	ORTHOPTERA	NC_006678.1
Myrmecophilus_manni	ORTHOPTERA	NC_011301.1
Ruspolia_dubia	ORTHOPTERA	NC_009876.1
Teleogryllus_emma	ORTHOPTERA	NC_011823.1
Troglophilus_neglectus	ORTHOPTERA	NC_011306.1
Hiodon_alosoides	OSTEOGLOSSIFORMES	NC_005145.1
Hiodon_tergisus	OSTEOGLOSSIFORMES	NC_015082.1
Gymnarchus_niloticus	OSTEOGLOSSIFORMES	NC_012707.1
Brienomyrus_niger	OSTEOGLOSSIFORMES	NC_012705.1
Genyomyrus_donnyi	OSTEOGLOSSIFORMES	NC_015086.1
Gnathonemus_petersii	OSTEOGLOSSIFORMES	NC_012717.2
Hippopotamyrus_wilverthi	OSTEOGLOSSIFORMES	NA
Marcusenius_senegalensis	OSTEOGLOSSIFORMES	NC_015090.1
Myomyrus_sp_CU6182	OSTEOGLOSSIFORMES	NC_015089.1
Paramormyrops_gabonensis	OSTEOGLOSSIFORMES	NC_015107.1
Petrocephalus_microphthalmus	OSTEOGLOSSIFORMES	NC_015098.1
Petrocephalus_soudanensis	OSTEOGLOSSIFORMES	NC_015092.1
Chitala_blanci	OSTEOGLOSSIFORMES	NC_012710.1
Chitala_lopis	OSTEOGLOSSIFORMES	NC_012711.1
Chitala_ornata	OSTEOGLOSSIFORMES	NC_012712.1
Notopterus_notopterus	OSTEOGLOSSIFORMES	NC_012713.1
Papyrocranus_congoensis	OSTEOGLOSSIFORMES	NC_012714.1
Xenomystus_nigri	OSTEOGLOSSIFORMES	NC_012715.1
Arapaima_gigas	OSTEOGLOSSIFORMES	NC_010570.1

Table A.1 continued

Heterotis_niloticus	OSTEOGLOSSIFORMES	NC_015081.1
Osteoglossum_bicirrhosum	OSTEOGLOSSIFORMES	NC_003095.1
Scleropages_formosus	OSTEOGLOSSIFORMES	NC_007012.1
Pantodon_buchholzi	OSTEOGLOSSIFORMES	NC_003096.1
Crassostrea_angulata	OSTREOIDA	NC_012648.1
Crassostrea_ariakensis	OSTREOIDA	NC_012650.1
Crassostrea_gigas	OSTREOIDA	NC_001276.1
Crassostrea_hongkongensis	OSTREOIDA	NC_011518.2
Crassostrea_iredalei	OSTREOIDA	NC_013997.1
Crassostrea_nippona	OSTREOIDA	NC_015248.1
Crassostrea_sikamea	OSTREOIDA	NC_012649.1
Crassostrea_virginica	OSTREOIDA	NC_007175.2
Ostrea_denselamellosa	OSTREOIDA	NC_015231.1
Saccostrea_mordax	OSTREOIDA	NC_013998.1
Pica_pica	PASSERIFORMES	NC_015200.1
Podoces_hendersoni	PASSERIFORMES	NC_014879.1
Pseudopodoces_humilis	PASSERIFORMES	NC_014341.1
Smithornis_sharpei	PASSERIFORMES	NC_000879.1
Menura_novaehollandiae	PASSERIFORMES	NC_007883.1
Cyanoptila_cyanomelana	PASSERIFORMES	NC_015232.1
Carduelis_sinica	PASSERIFORMES	NC_015196.1
Carduelis_spinus	PASSERIFORMES	NC_015198.1
Emberiza_chrysophrys	PASSERIFORMES	NC_015233.1
Emberiza_tristrami	PASSERIFORMES	NC_015234.1

Table A.1 continued

Taeniopygia_guttata	PASSERIFORMES	NC_007897.1
Vidua_chalybeata	PASSERIFORMES	NC_000880.1
Pycnonotus_sinensis	PASSERIFORMES	NC_013838.1
Pycnonotus_taivanus	PASSERIFORMES	NC_013483.2
Acridotheres_cristatellus	PASSERIFORMES	NC_015613.1
Sturnus_cineraceus	PASSERIFORMES	NC_015237.1
Sturnus_sericeus	PASSERIFORMES	NC_014455.1
Sturnus_tristis	PASSERIFORMES	NC_015195.1
Acrocephalus_scirpaceus	PASSERIFORMES	NC_010227.1
Sylvia_atricapilla	PASSERIFORMES	NC_010228.1
Sylvia_crassirostris	PASSERIFORMES	NC_010229.1
Leiothrix_argentauris	PASSERIFORMES	NC_015114.1
Luscinia_calliope	PASSERIFORMES	NC_015074.1
Cnemotriccus_fuscatus	PASSERIFORMES	NC_007975.1
Argopecten_irradians	PECTINOIDA	NC_009687.1
Argopecten_irradians_irradians	PECTINOIDA	NC_012977.1
Chlamys_farreri	PECTINOIDA	NC_012138.1
Mimachlamys_nobilis	PECTINOIDA	NC_011608.1
Mizuhopecten_yessoensis	PECTINOIDA	NC_009081.1
Placopecten_magellanicus	PECTINOIDA	NC_007234.1
Equus_asinus	PERISSODACTYLA	NC_001788.1
Equus_caballus	PERISSODACTYLA	NA
Ceratotherium_simum	PERISSODACTYLA	NC_001808.1
Coelodonta_antiquitatis	PERISSODACTYLA	NC_012681.1

Table A.1 continued

Dicerorhinus_sumatrensis	PERISSODACTYLA	NC_012684.1
Diceros_bicornis	PERISSODACTYLA	NC_012682.1
Rhinoceros_sondaicus	PERISSODACTYLA	NC_012683.1
Rhinoceros_unicornis	PERISSODACTYLA	NC_001779.1
Micadina_phluctainoides	PHASMATODEA	NC_014673.1
Phraortes_illepidus	PHASMATODEA	NC_014695.1
Phraortes_sp_Iriomote_Island	PHASMATODEA	NC_014705.1
Entoria_okinawaensis	PHASMATODEA	NC_014694.1
Ramulus_hainanense	PHASMATODEA	NC_013185.1
Ramulus_irregulariterdentatus	PHASMATODEA	NC_014702.1
Phobaeticus_serratipes	PHASMATODEA	NC_014678.1
Megacrania_alpheus	PHASMATODEA	NC_014688.1
Heteropteryx_dilatata	PHASMATODEA	NC_014680.1
Paralichthys_olivaceus	PLEURONECTIFORMES	NC_002386.1
Hippoglossus_hippoglossus	PLEURONECTIFORMES	NC_009709.1
Hippoglossus_stenolepis	PLEURONECTIFORMES	NC_009710.1
Kareius_bicoloratus	PLEURONECTIFORMES	NC_003176.1
Platichthys_stellatus	PLEURONECTIFORMES	NC_010966.1
Reinhardtius_hippoglossoides	PLEURONECTIFORMES	NC_009711.1
Verasper_moseri	PLEURONECTIFORMES	NC_008461.1
Verasper_variegatus	PLEURONECTIFORMES	NC_007939.1
Psetta_maxima	PLEURONECTIFORMES	NC_013183.1
Cynoglossus_abbreviatus	PLEURONECTIFORMES	NC_014881.1
Cynoglossus_semilaevis	PLEURONECTIFORMES	NC_012825.1

Table A.1 continued

Solea_senegalensis	PLEURONECTIFORMES	NC_008327.1
Cebus_albifrons	PRIMATES	NC_002763.1
Gorilla_gorilla	PRIMATES	NC_001645.1
Homo_sapiens	PRIMATES	NA
Hylobates_lar	PRIMATES	NC_002082.1
Lemur_catta	PRIMATES	NC_004025.1
Macaca_sylvanus	PRIMATES	NC_002764.1
Nycticebus_coucang	PRIMATES	NC_002765.1
Pan_paniscus	PRIMATES	NC_001644.1
Pan_troglodytes	PRIMATES	NC_001643.1
Papio_hamadryas	PRIMATES	NC_001992.1
Pongo_pygmaeus	PRIMATES	NC_001646.1
Pongo_abelii	PRIMATES	NC_002083.1
Tarsius_bancanus	PRIMATES	NC_002811.1
Nymphicus_hollandicus	PSITTACIFORMES	NC_015192.1
Agapornis_roseicollis	PSITTACIFORMES	NC_011708.1
Aratinga_pertinax_chrysogenys	PSITTACIFORMES	NC_015197.1
Brotogeris_cyanoptera	PSITTACIFORMES	NC_015530.1
Melopsittacus_undulatus	PSITTACIFORMES	NC_009134.1
Strigops_habroptilus	PSITTACIFORMES	NC_005931.1
Ancylostoma_caninum	RHABDITIDA	NC_012309.1
Ancylostoma_duodenale	RHABDITIDA	NC_003415.1
Bunostomum_phlebotomum	RHABDITIDA	NC_012308.1
Necator_americanus	RHABDITIDA	NC_003416.2

Table A.1 continued

Angiostrongylus_cantonensis	RHABDITIDA	NC_013065.1
Angiostrongylus_costaricensis	RHABDITIDA	NC_013067.1
Metastrongylus_pudendotectus	RHABDITIDA	NC_013813.1
Metastrongylus_salmi	RHABDITIDA	NC_013815.1
Chabertia_ovina	RHABDITIDA	NC_013831.1
Oesophagostomum_dentatum	RHABDITIDA	NC_013817.1
Oesophagostomum_quadrispinulatum	RHABDITIDA	NC_014181.1
Cylicocyclus_insignis	RHABDITIDA	NC_013808.1
Strongylus_vulgaris	RHABDITIDA	NC_013818.2
Syngamus_trachea	RHABDITIDA	NC_013821.1
Cooperia_oncophora	RHABDITIDA	NC_004806.1
Haemonchus_contortus	RHABDITIDA	NC_010383.2
Mecistocirrus_digitatus	RHABDITIDA	NC_013848.1
Teladorsagia_circumcincta	RHABDITIDA	NC_013827.1
Trichostrongylus_axei	RHABDITIDA	NC_013824.1
Trichostrongylus_vitrinus	RHABDITIDA	NC_013807.1
Steinernema_carpocapsae	RHABDITIDA	NC_005941.1
Heterorhabditis_bacteriophora	RHABDITIDA	NC_008534.1
Caenorhabditis_briggsae	RHABDITIDA	NA
Caenorhabditis_elegans	RHABDITIDA	NC_001328.1
Coregonus_lavaretus	SALMONIFORMES	NA
Oncorhynchus_clarkii_henshawi	SALMONIFORMES	NC_006897.1
Oncorhynchus_gorbuscha	SALMONIFORMES	NC_010959.1
Oncorhynchus_kisutch	SALMONIFORMES	NC_009263.1

Table A.1 continued

Oncorhynchus_masou_Biwa	SALMONIFORMES	NC_009262.1
Oncorhynchus_masou_formosanus	SALMONIFORMES	NC_008745.1
Oncorhynchus_masou_ishikawae	SALMONIFORMES	NC_008746.1
Oncorhynchus_masou_masou	SALMONIFORMES	NC_008747.1
Oncorhynchus_mykiss	SALMONIFORMES	NC_010976.1
Oncorhynchus_nerka	SALMONIFORMES	NC_008615.1
Oncorhynchus_tshawytscha	SALMONIFORMES	NC_002980.1
Salmo_salar	SALMONIFORMES	NC_008815.1
Salmo_trutta_trutta	SALMONIFORMES	NC_010007.1
Salvelinus_alpinus	SALMONIFORMES	NC_000861.1
Salvelinus_fontinalis	SALMONIFORMES	NC_000860.1
Thymallus_arcticus	SALMONIFORMES	NC_027408.1
Thymallus_thymallus	SALMONIFORMES	NC_009682.1
Liobagrus_obesus	SILURIFORMES	NC_008232.1
Leiocassis_longirostris	SILURIFORMES	NC_014586.1
Pelteobagrus_nitidus	SILURIFORMES	NC_014859.1
Pelteobagrus_vachellii	SILURIFORMES	NC_014862.1
Pseudobagrus_brevicorpus	SILURIFORMES	NC_015625.1
Pseudobagrus_tokiensis	SILURIFORMES	NC_004697.1
Corydoras_rabauti	SILURIFORMES	NC_004698.1
Cranoglanis_bouderius	SILURIFORMES	NC_008280.1
Ictalurus_punctatus	SILURIFORMES	NC_003489.1
Pangasianodon_gigas	SILURIFORMES	NC_006381.1
Silurus_glanis	SILURIFORMES	NC_014261.1

Table A.1 continued

Silurus_lanzhouensis	SILURIFORMES	NC_015650.1
Silurus_meridionalis	SILURIFORMES	NC_014866.1
Aeoliscus_strigatus	SYNGNATHIFORMES	NC_010270.1
Macroramphosus_scolopax	SYNGNATHIFORMES	NC_010265.1
Fistularia_commersonii	SYNGNATHIFORMES	NC_010274.1
Eurypegasus_draconis	SYNGNATHIFORMES	NC_010264.1
Pegasus_volitans	SYNGNATHIFORMES	NC_010271.1
Hippocampus_kuda	SYNGNATHIFORMES	NC_010272.1
Indostomus_paradoxus	SYNGNATHIFORMES	NC_004401.1
Microphis_brachyurus	SYNGNATHIFORMES	NC_010273.1
Solenostomus_cyanopterus	SYNGNATHIFORMES	NC_010267.1
Eretmochelys_imbricata	TESTUDINES	NC_012398.1
Macrochelys_temminckii	TESTUDINES	NC_009260.1
Chrysemys_picta	TESTUDINES	NC_002073.3
Chinemys_reevesi	TESTUDINES	NC_006082.1
Cuora_amboinensis	TESTUDINES	NC_014769.1
Cuora_flavomarginata	TESTUDINES	NC_012054.1
Cyclemys_atripons	TESTUDINES	NC_010970.1
Mauremys_megalocephala	TESTUDINES	NA
Sacalia_quadriocellata	TESTUDINES	NC_011819.1
Indotestudo_forstenii	TESTUDINES	NC_007696.1
Malacochersus_tornieri	TESTUDINES	NC_007700.1
Manouria_impressa	TESTUDINES	NC_011815.1
Psammobates_pardalis	TESTUDINES	NC_007694.1

Table A.1 continued

Testudo_graeca	TESTUDINES	NC_007692.1
Testudo_kleinmanni	TESTUDINES	NC_007699.1
Carettochelys_insculpta	TESTUDINES	NC_014048.1
Apalone_ferox	TESTUDINES	NC_014054.1
Palea_steindachneri	TESTUDINES	NC_013841.1
Trionyx_triunguis	TESTUDINES	NC_012833.1
Pelomedusa_subrufa	TESTUDINES	NC_001947.1
Loligo_opalescens	TEUTHIDA	NC_012840.1
Loligo_bleekeri	TEUTHIDA	NC_002507.1
Sepioteuthis_lessoniana	TEUTHIDA	NC_007894.1
Architeuthis_dux	TEUTHIDA	NC_011581.1
Watasenia_scintillans	TEUTHIDA	NC_007893.1
Dosidicus_gigas	TEUTHIDA	NC_009734.1
Sthenoteuthis_oualaniensis	TEUTHIDA	NC_010636.1
Todarodes_pacificus	TEUTHIDA	NC_006354.1
Panonychus_citri	TROMBIDIFORMES	NC_014347.1
Panonychus_ulmi	TROMBIDIFORMES	NC_012571.1
Tetranychus_cinnabarinus	TROMBIDIFORMES	NC_014399.1
Tetranychus_urticae	TROMBIDIFORMES	NC_010526.1
Unionicola_foili	TROMBIDIFORMES	NC_011036.1
Unionicola_parkeri	TROMBIDIFORMES	NC_014683.1
Ascoschoengastia_sp_TATW_1	TROMBIDIFORMES	NC_010596.1
Leptotrombidium_akamushi	TROMBIDIFORMES	NC_007601.1
Leptotrombidium_deliense	TROMBIDIFORMES	NC_007600.1

Table A.1 continued

Leptotrombidium_pallidum	TROMBIDIFORMES	NC_007177.1
Walchia_hayashii	TROMBIDIFORMES	NC_010595.1
Margaritifera_falcata	UNIONOIDA	NC_015476.1
Quadrula_quadrula	UNIONOIDA	NC_013658.1
Cristaria_plicata	UNIONOIDA	NC_012716.1
Lasmigona_compressa	UNIONOIDA	NC_015481.1
Pyganodon_grandis	UNIONOIDA	NC_013661.1
Utterbackia_imbecillis	UNIONOIDA	NC_015479.1
Utterbackia_peninsularis	UNIONOIDA	NC_015477.1
Lampsilis_ornata	UNIONOIDA	NC_005335.1
Toxolasma_parvus	UNIONOIDA	NC_015483.1
Hyriopsis_cumingii	UNIONOIDA	NC_011763.1
Hyriopsis_schlegelii	UNIONOIDA	NC_015110.1
Unio_pictorum	UNIONOIDA	NC_015310.1
Venustaconcha_ellipsiformis	UNIONOIDA	NC_013659.1
Acanthocardia_tuberculata	VENEROIDA	NC_008452.1
Loripes_lacteus	VENEROIDA	NC_013271.1
Lucinella_divaricata	VENEROIDA	NC_013275.1
Sinonovacula_constricta	VENEROIDA	NC_011075.1
Meretrix_lusoria	VENEROIDA	NC_014809.1
Meretrix_meretrix	VENEROIDA	NC_013188.1
Meretrix_petechialis	VENEROIDA	NC_012767.1
Paphia_euglypta	VENEROIDA	NC_014579.1
Venerupis_philippinarum	VENEROIDA	NC_031332.1

Table A.1 continued

Zenion_japonicum	ZEIFORMES	NC_004397.1
Allocyttus_niger	ZEIFORMES	NC_004398.1
Neocyttus_rhomboidalis	ZEIFORMES	NC_004399.1
Parazen_pacificus	ZEIFORMES	NC_004396.1
Zenopsis_nebulosus	ZEIFORMES	NC_003173.1
Zeus_faber	ZEIFORMES	NC_003190.1

APPENDIX

B

SIMULATION PARAMETERS

B.1 Description

Table of parameters used for simulations in chapter 4. Not all parameters shown here were used in figures.

B.2 Table

Table B.1 Simulation Parameters. Table of simulation parameters for the simulations discussed in 4. Not every combination of these parameters were run but they are listed here to give a general idea of the breadth of the simulation.

Sequences	Sites	ω_3	Synonymous CV
16	100	1	0
31	300	1.1	0.103
	500	2.077	0.132
	1000	6	0.209
	5000	13	0.303
			0.346
			0.4
			0.5
			0.574
			0.697
			0.821
			1.002
			1.128
			1.507
			1.718
			1.995
			2.002
			2.491
			3.018
			4.028
			5.533

Table B.1 continued from previous page

			10.579
			15.58