Analysis of Functional Constraint and Recombination in Gene Sequences of the Cyanobacteria *Prochlorococcus*

by

Rachael Bay

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2010

DALHOUSIE UNIVERSITY

DEPARTMENT OF BIOLOGY

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Analysis of Functional Constraint and Recombination in Gene Sequences of the Cyanobacteria *Prochlorococcus*" by Rachael Bay in partial fulfilment of the requirements for the degree of Master of Science.

Dated: August 17, 2010

Supervisor: ___________________________________

Readers: ___________________________________

____

Departmental Representative: _____________________________

DALHOUSIE UNIVERSITY

DATE: August 17, 2010

AUTHOR: Rachael Bay

TITLE: Analysis of Functional Constraint and Recombination in Gene Sequences
of the Cyanobacteria *Prochlorococcus*

DEPARTMENT OR SCHOOL: Department of Biology

DEGREE: MSc CONVOCATION: October YEAR: 2010

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

____
Signature of Author

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

The author attests that permission has been obtained for the use of any copyrighted material appearing in the thesis (other than the brief excerpts requiring only proper acknowledgement in scholarly writing), and that all such use is clearly acknowledged.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Lineages of the cyanobacteria *Prochlorococcus marinus* have diverged into two genetically distinct 'ecotypes,' high-light adapted (HL) and low-light adapted (LL), which thrive under different environmental conditions. This type of niche differentiation in prokaryotes is often accompanied by genetic and genomic divergence. Differential selection pressure associated with ecotype divergence can be analyzed using models of codon evolution. However, some characteristics of the *Prochlorococcus* genome violate underlying assumptions of these models. For example, high levels of recombination between bacterial strains are known to cause false positives for codon models. Therefore, it is important that statistical methods for detecting recombination be reliable. In Chapter 2, I evaluate a set of recombination detection methods under four different scenarios related to functional divergence: 1) varying tree shape, 2) positive selection, 3) non-stationary evolution, and 4) varying levels of recombination and divergence. I find that some methods yield inflated false positive rates under an asymmetric topology and slightly inflated false positive rates under some non-stationary conditions. Users of these methods should therefore be aware of their performance under conditions relevant to a given dataset.

Another limit posed to codon models by *Prochlorococcus* is the presence of non-stationary nucleotide composition. Current models that infer shifts in selection pressure do not account for this type of heterogeneity. For this reason, in Chapter 3, I conduct a simulation to determine the effects of non-stationary evolution on the inference of selection pressure. I find that non-stationary evolution can cause a false signal for a shift in selective pressure, making analysis of genes from organisms such as *Prochlorococcus* unreliable.

I use *Prochlorococcus* as an empirical supplement to my simulations in Chapters 2 and 3, analyzing a subset of the core genome for functional divergence and within-gene recombination. While my results are preliminary due to the limitations of current analytical methods, they do suggest that both of these processes may play a substantial role in the evolution of the *Prochlorococcus* core genome. In Chapter 4, I take a more focused approach, reanalyzing the *cpeB* gene, which was previously claimed to have signal for positive selection based on analysis with codon models. Using simulations based on parameter values estimated from this specific gene, I conclude that, due to the degree of heterogeneity present in the gene, this result could be a false positive. This study provides an example of analysis under conditions that violate an assumption of the underlying model.

LIST OF SYMBOLS AND ABBREVIATIONS USED

AIC	Akaike information criterion
AT	adenine and thymine
bp	base pairs
BIC	Bayesian information criterion
BLAST	basic local alignment search tool
d_{N}	nonsynonymous substitution rate
d_{S}	synonymous substitution rate
ENC	effective number of codons
GARD	genetic algorithm for recombination detection
GC	guanine and cytosine
GC3	guanine and cytosine at the third codon position
GTR	general time reversible model
HKY85	Hasegawa-Kishino-Yano substituion model
HL	high-light
HR	homologous recombination
KA	Karlin and Altschul method
KH	Kishino-Hasegawa test
LGT	lateral gene transfer
LL	low-light
LRT	likelihood ratio test
M0	one-ratio sites model
M1a	nearly neutral sites model
M2a	positive selection sites model
M3	discrete sites model
M7	beta sites model
M8	beta & ω sites model
MBP	multiple breakpoint
MCMC	Markov chain Monte Carlo
p	beta function shape parameter
q	beta function shape parameter
r_{i}	evolutionary rate at site i
RDP	recombination detection program
RNA	ribonucleic acid
SBP	single breakpoint
UPGMA	unweighted pair group method with arithmetic mean
α	significance level
η	codon bias parameter
κ	transition/transversion ratio
π_{j}	equilibrium frequency of codon j
χ^2	chi-squared distribution
ω	nonsynonymous/synonymous rate ratio ($d_{\mathrm{N}}/d_{\mathrm{S}}$)

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to Dr. Joseph Bielawski for his constant guidance and discussion throughout this entire study and the preparation of this manuscript. I would also like to thank Dr. Katherine Dunn for her continual suggestions and willingness to assist with both theoretical and technical issues. Finally, I would like to acknowledge Dr. Andrew Roger for serving on my committee as well as for his advice on analytical techniques.

CHAPTER 1: INTRODUCTION

1.1 *PROCHLOROCOCCUS* ECOLOGY AND EVOLUTION

The cyanobacteria *Prochlorococcus marinus* (Chisholm et al. 1988) is a globally significant prokaryote that is well studied, particularly with respect to its ability to inhabit a broad range of habitats and its unique genomic features. Although the smallest strain, MED4, has a genome of only 1,657,990 bp (Kettler et al. 2007) and is the smallest known phototroph (Strehl et al. 1999), members of the *Prochlorococcus* genus are among the most abundant photosynthetic organisms in the open ocean, responsible for up to 80% of primary production in oligotrophic surface waters (Goericke and Welschmeyer 1993; Campbell et al. 1994; McManus and Dawson 1994; Vaulot et al. 1995; Suzuki et al. 1995; Liu et al. 1997). The extreme abundance and productivity of this organism make it important to both ecosystem dynamics and global carbon cycling.

The success of *Prochlorococcus* is partly attributed to its ability to inhabit a wide range of environmental conditions. This broad distribution is accomplished through the presence of different ecotypes, or strains, that thrive in different habitats. Although the ribosomal RNA of all known *Prochlorococcus* strains is over 97% identical (Moore et al. 1998), enough genomic variation exists to cause varied relative fitness under different environmental conditions. In most studies, *Prochlorococcus* strains are divided into two ecotypes, a high-light adapted (HL) ecotype and a low-light adapted (LL) ecotype, based on optimal light intensity (Moore et al. 1995; Partensky et al. 1997; West and Scanlan 1999; Rocap et al. 2002). Chlorophyll b/a ratios differ greatly between these two basic ecotypes (Moore et al. 1995). In addition, several other factors can affect ecotype abundance, such as temperature, nutrient concentration, trace metal concentration, and predator abundance (Mann et al. 2002; Tolonen et al. 2006; West et al. 2001; Moore et al. 2002; Ahlgren et al. 2006). Recently, several studies argue that the two ecotype model is oversimplified and that as many as six distinct ecotypes might exist (Coleman et al. 2006; Johnson et al. 2006; Kettler et al. 2007). For the purpose of this thesis, however, I choose to focus on genetic divergence associated with the divergence of HL and LL metabolism in *Prochlorococcus*, so the two-ecotype model is used (Figure 1.1).

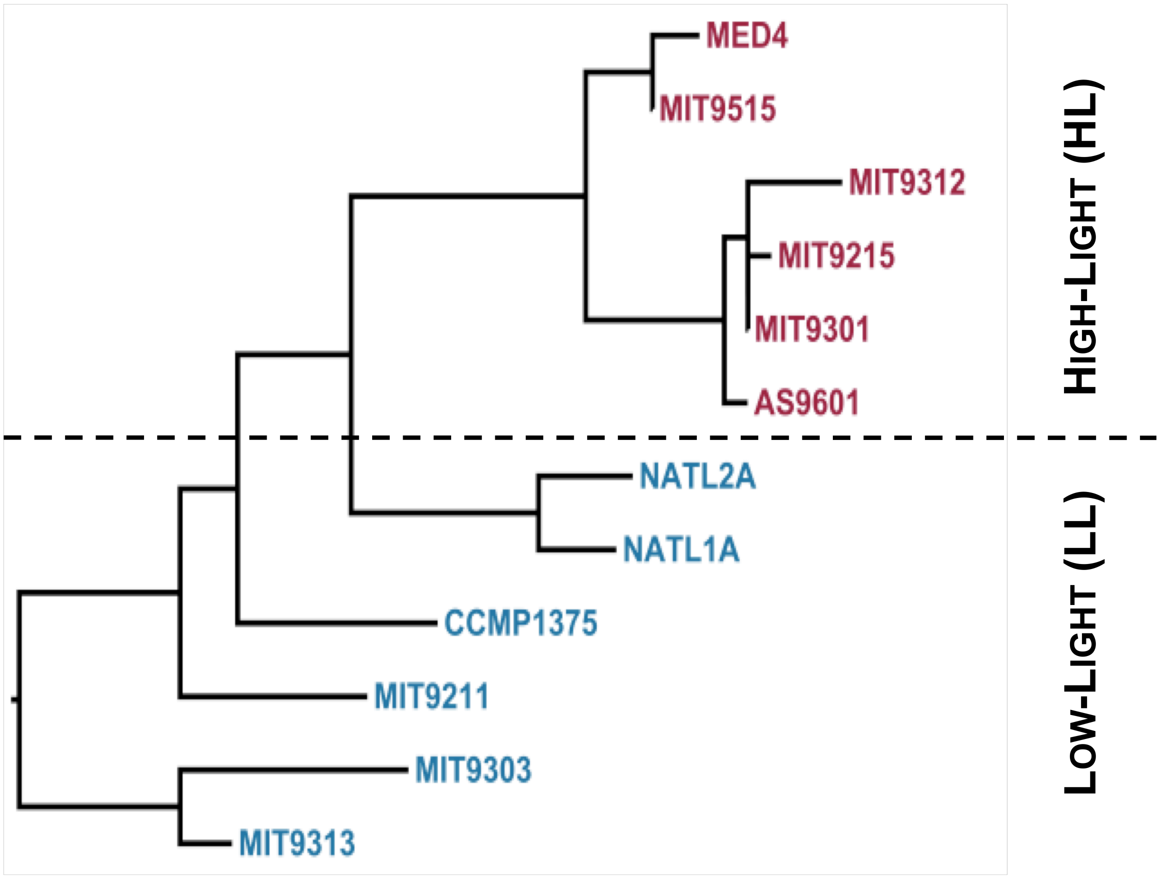


Figure 1.1. Phylogeny of *Prochlorococcus* 16S ribosomal DNA. High-light and low-light adapted ecotypes are separated by a dotted line. This tree was generated for this thesis by using maximum likelihood under and HKY85 model with a gamma distribution for among-site rate variation.

The distinction between HL and LL ecotypes is supported by the *Prochlorococcus* phylogeny. The more recently diverged HL strains form a monophyletic clade while the LL strains form a more ancestral grade (Figure 1.1). Several interesting patterns of evolution are observed in the *Prochlorococcus* genome. Since its divergence from its closest relative, *Synechococcus*, strains of *Prochlorococcus* have experienced an overall reduction in genome size (Hess et al. 2001; Dufresne et al. 2005). While genome reduction is well documented in endosymbionts (*e.g.,* Moran 2003; Lane 2007), *Prochlorococcus* was the first example of this pattern in free-living bacteria (Rocap et al. 2003). The more recently diverged strains, specifically those that make up the HL clade, are much smaller (1686 genes in HL strain MED4) than their LL counterparts (2200 genes in the most deeply branching MIT9313) (Kettler et al. 2007). In endosymbionts, the phenomenon of genome reduction is usually attributed to genetic drift due to very small effective population sizes (Moran 2003; Kuo et al. 2009). However, since *Prochlorococcus* is presumed to have one of the largest effective population sizes on earth, several studies have proposed alternative explanations for the reduction in number of genes, including selection for “streamlining” in order to conserve resources in a nutrient poor environment (*e.g.,* Rocap et al. 2003; Dufresne et al. 2005).

Regardless of the cause, genome reduction is often associated with a shift in nucleotide composition. This pattern is observed many times in endosymbionts (*e.g.,* Moran 2003; McCutcheon et al. 2009; Rispe et al. 2004; Kneip et al. 2008) and is also apparent in *Prochlorococcus* genomes (Hess et al. 2001; Dufresne et al. 2005). The most deeply branching LL strain, MIT9313, has an overall GC content of 50.7% while the most recently diverged HL strain has a GC content of 30.8% (Kettler et al. 2007; Hess et al. 2001). This difference in composition may be due to the loss of the *mutY* gene in HL lineages, which prevents certain mutations caused by damaged guanine residues (Kettler *et al.* 2007). However, codon bias may also arise from other situations, such as preferential use of AT-rich codons for energetic purposes or selection for certain codons based on tRNA concentrations (Hershberg and Petrov 2008).

Another mechanism that can shape the evolution of genomes is recombination. Recombination, in the form of homologous recombination (HR) or lateral gene transfer (LGT), can be an important source of genetic variation by allowing access to an extended gene pool (*e.g.,* Koonin et al. 2001; Boucher et al. 2003; Narra and Ochman 2006). In *Prochlorococcus*, evidence of recombination via phage intermediates is observed in several photosynthesis-related proteins, including photosystem II core proteins (Mann et al. 2003; Zeidner et al. 2005; Sullivan et al. 2006) and high-light inducible proteins (Lindell et al. 2004; Mann et al. 2005). Since these proteins are associated with light-based metabolism in *Prochlorococcus*, LGT and HR may have played a role in the divergence of HL and LL ecotypes.

The motivation of this thesis is to explore genetic and genomic changes associated with the ecological divergence of *Prochlorococcus* ecotypes. To accomplish this, I examine the two main evolutionary forces acting upon *Prochlorococcus* genomes: functional divergence via substitution and homologous recombination. However, the complex evolutionary history of *Prochlorococcus*, including non-stationary nucleotide composition and evolutionary rate variation, may negatively impact methods commonly used for evolutionary analysis. In the next two sections of the introduction, I discuss the two major evolutionary forces as well as current analytical methods.

1.2 FUNCTIONAL DIVERGENCE

Niche differentiation is often accompanied by changes in gene function. In order to better understand the process of functional divergence, several statistical methods have been developed to model the process (for reviews see Gaucher et al. 2002; Yang 2002). Most of these models are based on the concept that evolutionary rate is associated with functional constraint and therefore a shift in rate indicates divergence of gene function. Specifically, the rate of substitution at a given site is assumed to be dependent on the site's functional importance (Kimura 1983). A beneficial mutation will get fixed in a population through natural selection, while a deleterious mutation will be removed from the population. Because most mutations at functionally important sites are deleterious, they are removed from the population by natural selection, so these sites tend to change slowly over evolutionary time. Therefore, the rate of evolution at a site can be used as an indicator for the degree of functional constraint (*e.g.,* Gu 1999; Pupko et al. 2002; Blouin et al. 2003). Methods designed to detect functional divergence utilize this association between functional constraint and evolutionary rate, testing for sites with a shift in evolutionary rate across a phylogeny. However, there are several different approaches to modeling functional divergence, each with unique advantages and limitations.

Several methods are designed to detect shifts in evolutionary rate at both the nucleotide level (*e.g.,* Lockhart et al. 1998; Dorman 2007) and the amino acid level (*e.g.,* Gu 1999; Knudsen and Miyamoto 2001; Susko et al. 2002). However, there are limitations to both of these approaches. At the protein level, many changes in the nucleotide sequence that do not code for a change in amino acid are not utilized. While these changes are apparent at the nucleotide level, models of nucleotide evolution do not take into account the dependency of substitutions at different sites within a codon.

Unlike nucleotide models, codon models take into account the non-independent nature of nucleotide sites. In addition, codon models are able to utilize more information than amino acid models because they do not ignore synonymous substitutions. However, codon models are only effective when divergence is low enough that synonymous changes are not saturated (*i.e.*, new substitutions are replacing previous ones, resulting in a loss of information) and high enough that sufficient information is available within the data to estimate the nonsynonymous substitution rate (Anisimova et al. 2001). Hence, codon models are applicable only within a window of optimal sequence divergence. In addition, because codon models rely on information from synonymous sites, they may be more sensitive than amino acid models to shifts in synonymous codon bias over evolutionary time. Because codon and amino acid models have different advantages and limitations, it may be most effective to use them in conjunction with one another.

1.2.1 Codon Models

Goldman and Yang (1994) and Muse and Gaut (1994) independently developed similar models of codon evolution for the purpose of measuring the intensity of natural selection pressure acting on protein coding sequences. These models use synonymous (silent) substitutions as a measure of the rate of evolution of a coding sequence before the effect of natural selection on its protein product. Nonsynonymous substitutions, which cause a change in the amino acid sequence, are used to measure the rate of evolution after selection acts on the protein. Rates of nonsynonymous (d_N) and synonymous (d_S)

substitutions can then be calculated across a phylogeny. The parameter ω, equal to the ratio of nonsynonymous to synonymous substitutions (d_N/d_S), is employed as a measure of the strength and direction of selection pressure. When ω =1, the rate of nonsynonymous substitution is equal to the rate of synonymous substitution. This indicates that natural selection is not affecting the substitution rate and thus evolution at such a site is consistent with neutrality. When ω is greater than one, there are more nonsynonymous than synonymous mutations being fixed, suggesting that the nonsynonymous substitutions have been beneficial to the organism. Therefore, ω>1 indicates positive, or diversifying selection. Likewise, an ω value of less than one is an indication of negative, or purifying, selection (Goldman and Yang 1994).

Originally, measures of d_N and d_S were computed only in pairwise fashion using counting methods (*e.g.,* Miyata and Yasunaga 1980; Li et al. 1985; Nei and Gojobori 1986; Pamilo and Bianchi 1993). These methods simply count the number of nonsynonymous and synonymous changes between two sequences and correct for multiple substitutions at a site. However, counting methods often over-simplify the evolutionary process. For example, several methods assume equal rates of transitions and transversions (Miyata and Yasunaga 1980; Nei and Gojobori 1986). In addition, the majority of counting methods provide no adjustment for bias in codon usage, which can negatively impact estimates of synonymous and nonsynonymous substitution rates (Yang and Nielsen 1998; Bielawski et al. 2000; Dunn et al. 2001; Aris-Brosou and Bielawski 2006).

Goldman and Yang (1994) and Muse and Gaut (1994) implemented a maximum likelihood model for codon evolution in which ω is an explicit parameter and is computed as a function of all the sequences in a given phylogeny. These models allow for several other parameters to be estimated from the data, including the transition/transversion ratio and codon frequencies. In addition, because the model is implemented in a maximum likelihood framework, it deals with instantaneous rates of substitution and therefore can more accurately estimate substitution rates when multiple changes have occurred at a site. Thus, the main disadvantages of the pairwise counting methods are avoided by maximum likelihood estimation of ω from a phylogeny.

These maximum likelihood models describe the evolution from one codon to another using a Markov process where the states are the 61 sense codons (64 codons minus 3 stop codons). The probability of change from codon *i* to codon *j* is described in matrix Q:

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ by 2 or 3 codon positions} \\ \pi_{\mathrm{j}} & \text{if } i \text{ and } j \text{ differ by one synonymous transversion} \\ \kappa\pi_{\mathrm{j}} & \text{if } i \text{ and } j \text{ differ by one synonymous transition} \\ \omega\pi_{\mathrm{j}} & \text{if } i \text{ and } j \text{ differ by one nonsynonymous transversion} \\ \kappa\omega\pi_{\mathrm{j}} & \text{if } i \text{ and } j \text{ differ by one nonsynonymous transition} \end{cases}$$

where κ is the transition/transversion ratio and π_j is the equilibrium frequency of codon *j*. The probability of changing from codon *i* to codon *j* over time *t* is $P(t)=p_{ij}(t)=e^{Qt}$ (Goldman and Yang 1994); This probability is then used to calculate a log-likelihood score for a given tree topology (Felsenstein 1981).

Modifications of the original models of Goldman and Yang (1994) allow ω to change among sites. These models, known as "sites models," allow variation in ω among codon sites within a gene, but the site categories are constant across the phylogeny. For example, model M1a, or the "nearly neutral model," fits the data to a mixture of two different site categories. The first category contains sites under purifying selection (ω

constrained to be less than one) while the second contains sites that are evolving neutrally (ω fixed to one) (Nielsen and Yang 1998; Wong et al. 2004). Sites model M3, on the other hand, simply fits the data to an unconstrained mixture of three discrete categories with no limits on the value of ω for each category (Yang et al. 2000). In total, 14 different mixture models have been implemented (Appendix A), but only a small subset (M0, M1a, M2a, M3, M7, and M8) are recommended for real data analysis (Bielawski and Yang 2005)

Codon models also have been modified to allow for variation in evolutionary rate among lineages. These "branch models" average the value of ω over all sites within a gene, but allow variation in different branches (Yang and Nielsen 1998; Yang 1998). Branch models may be used to test if selection pressure differs across a phylogeny. While branch models are an improvement over pairwise comparisons because they can compare multiple lineages in a phylogeny, power to detect positive selection is low because ω is averaged over all sites within the gene (Yang and Nielsen 2002).

A combination of sites models and branch models, "branch-sites models" allow ω to vary both among sites and between branches. Background and foreground branches, which are expected to have different rates of codon substitution, are specified *a priori.* For example, branch-sites Model A fits the data to a mixture of four site categories. The first contains sites under purifying selection (ω<1) and the second contains sites evolving neutrally (ω=1). These two categories remain constant throughout the phylogeny. Site categories three and four contain sites that switch from one of the first two categories to positive selection (ω>1) in a specific branch, or branches, of a tree. A branch having ω from the first two categories is called a background branch, whereas any branch having a switch to ω>1 is called a foreground branch (Yang and Nielsen 2002; Zhang et al. 2005). In total, five such branch-sites models have been implemented (Appendix B).

In order to determine the most likely evolutionary scheme for a given dataset, we must determine which model best fits the data without employing unnecessary parameters. Methods such as the Akaike information criterion (AIC), Bayesian information criterion (BIC), and likelihood ratio tests (LRT) are often used for codon models (*e.g.,* Yang et al. 2000; Kosakovsky Pond and Muse 2005a). The likelihood ratio test statistic, used to compare two nested hypotheses, is equal to two times the difference in log-likelihoods between the two models ($2*\Delta lnL$) and is compared to a χ^2 distribution in order to find the probability that the alternate model is a better fit to the data. For example, an explicit test for positive selection in a fraction of sites in foreground branches compares sites model M1a with branch-sites Model A (Yang and Nielsen 2002). When the test statistic is compared to the χ^2 distribution having two degrees of freedom, a significant *p*-value suggests an increased fit, due to the presence of positive selection in the foreground branch (as implemented in Model A), is greater than expected by chance simply due to the inclusion of extra parameters in the model (Nielsen and Yang 1998).

While codon models are useful tools for measuring selection pressure acting on a gene product, they do have some limitations. In order to reduce the number of parameters involved, codon frequencies are measured empirically from the data and fixed across the entire phylogeny. However, in cases where a shift in codon bias occurs in one or more branches of the phylogeny, this is not an accurate representation of the data. We know that other types of heterogeneity, such as large shifts in codon usage among sites

(Bao et al. 2008), can negatively impact biological conclusions drawn using codon models. In addition, time-heterogeneity in codon usage may result in phylogenetic artifacts (Inagaki and Roger 2006). However, the impact of non-stationary codon usage on conclusions drawn using codon models is not yet known.

1.2.2 Amino Acid Models

Heterogeneous rates of amino acid substitution over evolutionary time were first observed in the 1970's. At this point, Fitch and Markowitz (1970) developed the "covarion" (concomitantly variable codon) model, which suggests a proportion of sites will shift between variable and conserved states over time. This model was later expanded upon to allow amino acid sites to shift between variable rate categories (Galtier 2001; Wang et al. 2007) and to allow for different proportions of variable sites (Lopez et al. 2002). This more general form of rate variation is referred to as heterotachy. Heterotachy accounts for shifts in evolutionary rates over time, and these shifts may occur at any point in the phylogeny.

Often, however, we want to test for a shift in evolutionary rate at a given point in a phylogeny (*e.g.,* at a gene duplication event, a speciation event, or an LGT event). "Rate-shift" models accomplish this by testing for shifts in evolutionary rates between subtrees that are specified *a priori* (*e.g.,* Gu 1999; Knudsen and Miyamoto 2001; Susko et al. 2002). These models are employed to measure the rates of evolution in two separate subtrees. These rates are then used to identify sites most likely to have experienced a shift in functional constraint. For some rate-shift models, it is assumed a known divergence in gene function exists between the two subtrees (*e.g.,* Knudsen and

Miyamoto 2001). However, some methods make no such assumption, using information from analysis at amino acid sites to test hypotheses about functional divergence at the gene level (*e.g.,* Gu 1999; Susko et al. 2002). Using these methods, sequence alignments can be simultaneously tested for functional divergence at the gene level while identifying specific sites associated with that divergence in gene function.

While functional divergence is traditionally associated with a shift in evolutionary rate, this is not always the case. Gu (2001) identifies two separate types of functional divergence at the amino acid level: Type I and Type II. Type I functional divergence occurs when there is a shift in evolutionary rate between two subtrees. This type of divergence is detected by rate-shift models, which are discussed above. However, a shift in amino acid composition without a shift in evolutionary rate may also indicate functional divergence. This is known as Type II divergence. Type II divergence may occur if a historical relaxation of functional constraint has occurred at a site, but the site has since gained a new, but equally important function. For example, a site that is completely conserved in two subtrees may be composed of different amino acids in each subtree that have different physico-chemical properties. Therefore, while there is no shift in evolutionary rate at this site, there may still be divergence of gene function. However, only a few methods currently detect Type II functional divergence (*e.g.,* Gu 2006; Gaston unpublished method).

There are many different methods designed to detect functional divergence from amino acid data. They can be divided into categories based on how they identify specific sites that have experienced a shift in functional constraint. One of the most common means of identifying these sites is by using posterior probabilities. These methods (*e.g.,* Gu 1999; Gu 2006; Gaston unpublished method) use empirical Bayesian techniques to identify the sites with the greatest posterior probability of contributing to functional divergence, based on a given model. Another common means of identifying site specific rate shifts is by searching for a shift in the substitution distribution at a given position, using some measure of the difference in rate distribution between the two subtrees (*e.g.,* Gaucher et al. 2001; Lopez et al. 2002; Susko et al. 2002; Pupko and Galtier 2002). Another alternative is to use the likelihood ratio test on a site-by-site basis (Knudsen and Miyamoto 2001) in order to determine whether a rate shift has occurred at a given site. All of these methods have successfully been applied to real data (*e.g.,* Gu 1999; Susko et al. 2002). A list of amino acid level methods useful for investigating functional divergence is presented in Appendix C.

There are both benefits and limitations to modeling evolution at the amino acid level. Unlike most codon models, amino acid models take into account variable substitution rates between different amino acid states by employing empirical exchangeability matrices. Furthermore, because synonymous substitution rates are not measured, amino acid models may be less sensitive than codon models to shifts in nucleotide composition. However, since not all of the sequence information is utilized, amino acid models may have low power to detect functional divergence in some cases. The performance of these models under some complex evolutionary scenarios, such as those seen in *Prochlorococcus*, has yet to be evaluated.

1.3 RECOMBINATION

Another mechanism that impacts gene evolution is recombination. Recombination is the exchange of genetic material between two sequences. This exchange may occur between genomes of organisms or between different chromosomal locations within the same organism. Recombination can extend across species boundaries, permitting organisms to access much larger pools of genetic variation (*e.g.,* Dunn et al. 2009). By allowing access to an extended gene pool, recombination may increase genetic diversity in a population, assisting in adaptation. Although the relative contributions of recombination and mutation vary greatly, both play key roles in the evolution and adaptation of genes and organisms (*e.g.,* Hanage et al. 2005; Fraser et al. 2007; Didelot and Maiden 2010).

Homologous recombination (HR) and lateral gene transfer (LGT) are two mechanisms of recombination that are well documented among prokaryotes (*e.g.,* Konstantinos and DeLong 2008; Barker et al. 2000; Lodders et al. 2005). In HR, a sequence fragment from a donor organism is exchanged with a homologous sequence in the genome of a recipient organism. Because the machinery for HR is limited by the degree of sequence divergence, HR occurs only among highly similar gene sequences. In lateral gene transfer, on the other hand, homologous or non-homologous genetic material can be transferred between organisms, sometimes via phage intermediates (*e.g.,* Zeidner et al. 2005; Sullivan et al. 2005; Lindell et al. 2004; Dammeyer et al. 2008). Viruses can mediate the transfer of gene fragments or whole genes, sometimes even between distantly related organisms (*e.g.,* Beiko et al. 2005; Kunin et al. 2005). HR and LGT extend the pool of genes which an organism can access, increasing diversity in a population and giving the organism an extended source of potentially adaptive sequence variants from outside the population.

The detection of within-gene recombination is important to understanding bacterial evolution, but presents many challenges. Fully understanding the effect of recombination on the evolution of a gene sequence requires successful inference at several levels. First, a qualitative determination of whether or not recombination has occurred must be performed. This is the least difficult task, and there are several methods that are reasonably powerful under certain circumstances and adequately manage the type-I error rate (Posada and Crandall 2001; Chan et al. 2006). Second, the number of recombination breakpoints and their locations within a gene is often important to determine. This task requires that more information be extracted from a given set of sequences and thus is more difficult than the first. The third task is to identify the donor and recipient sequences. This is the most difficult inference, especially because the exact donor sequence is most likely not included in a given sequence alignment (hence, you are often working with relatives of the donor sequence). Because the problem of modeling recombination is so complex, reliable methods for detecting recombination are difficult to implement.

1.3.1 Methods of Recombination Detection

Many different tools are available for the detection and analysis of recombination events. "Basic branch-pattern" methods search for adjacent fragments of genes with different branching orders (*e.g.,* Hein 1990; Jakobsen and Easteal 1996; Martin and Rybicki 2000; Lole et al. 1999). Substitution-distribution methods search for sequence fragments with greater than expected similarity (*e.g.,* Sawyer 1989; Maynard Smith 1992; Maynard Smith and Smith 1998; Worobey 2001; Posada and Crandall 2001). More recently developed methods are characterized by increasing complexity in modeling the evolutionary process (*e.g.,* Suchard et al. 2002; Husmeier and McGuire 2003; Kosakovsky Pond et al. 2006). Because each type of analysis has its benefits and limitations, one must carefully consider the data in hand before choosing a method.

Basic branch-pattern methods search for adjacent gene fragments with discordant phylogenies. These methods are often referred to in the literature as "phylogeny-based methods." However, this is a somewhat confusing term because new methods are available that also make use of a phylogenetic framework, but are not included in this group because they employ a much more complex computational and statistical framework. Therefore, to avoid confusion, I use the term "basic branch-pattern methods" to refer to the group of early methods that use simple means to search for discordant phylogenies. To reduce computational costs some of these methods sample, and work with, triplets of gene sequences. One example is the Recombination Detection Program (RDP) (Martin and Rybicki 2000). This method analyzes combinations of three sequences (A, B, and C) where A and B are more closely related to one another than C. It then searches for sequence fragments in which either AC or BC has a higher similarity score than AB. While this approach allows for quick data analysis, there are analytical costs. First, only a subset of the data is considered at any one time, so power may be lower than methods that simultaneously utilize all data. Also, the method is reliant on the UPGMA algorithm for phylogeny generation, so there is no built-in method for handling among-lineage rate variation. Finally, similarity scoring is done on a match or mis-match basis rather than using a substitution matrix, so compositional variation is not taken into account. Nonetheless, RDP is a popular method and appears to perform well under certain conditions.

Substitution-distribution methods search for segments of genes with significantly high similarity, but they do not take into account phylogenetic relatedness. Here I will focus on three such methods (GENECONV, MaxChi, and Chimaera), which a previous simulation study found to have relatively high power (Posada and Crandall 2001). Using either pairs or triplets of sequences, MaxChi (Maynard Smith 1992) searches for regions of a gene with different proportions of variable to invariable sites than adjacent regions. Monomorphic sites are first discarded and then a sliding window is used to compare the difference in proportion of variable vs. invariable sites using a χ^2 distribution. When the χ^2 values are plotted by alignment location, peaks in the distribution indicate potential recombination breakpoints. The Chimaera method is a modification of MaxChi that has a more conservative method for discarding monomorphic sites and uses only triplets of sequences (Posada and Crandall 2001). Like MaxChi and Chimaera, GENECONV (Sawyer 1989) also discards monomorphic sites in the initial step. Pairs of sequences are then compared for segments that are either identical or have high similarity scores. Highly similar segments are scored and assigned a significance value based on the Karlin and Altschul (KA) method, which is similar to a BLAST search.

Some limitations are inherent in substitution-distribution methods. In MaxChi and Chimaera, the entire alignment is used to determine the proportion of variable sites, so if an alignment contains both closely related and very divergent sequences, some recombination may be overlooked. Likewise, a stretch of conserved sites in two very closely related sequences will have high similarity scores, which may result in a false positive in GENECONV. In addition, a single highly diverged sequence may introduce a large number of polymorphic sites, resulting in false negatives for some recombination events with any of the substitution-distribution methods.

Another group of recombination detection methods make use of a Bayesian framework (*e.g.,* DualBrothers, BARCE). These methods use posterior probabilities to determine the number and location of breakpoints in a given alignment (*e.g.,* Suchard et al. 2002; Suchard et al. 2003; Husmeier and McGuire 2003; Minin et al. 2005; Marttinen et al. 2008; Webb et al. 2009). The majority of these methods employ Markov Chain Monte Carlo (MCMC) analysis, where the state is the tree topology and a transition between states indicates a recombination break point. These methods are often very computationally intensive, and some are only designed to analyze groups of four taxa (*e.g.,* Husmeier and McGuire 2003). In addition, these methods use more complex models than substitution-distribution or phylogenetic methods and thus may be more susceptible to model misspecification.

Besides Bayesian methods, one alternative to the more simplistic approaches is the Genetic Algorithm for Recombination Detection (GARD). GARD employs a heuristic population-based genetic algorithm to search for recombination. Using a maximum likelihood framework, it selects the pattern of recombination, including number and location of breakpoints, that best fits a given alignment (Kosakovsky Pond et al. 2006). While GARD is attractive because it employs realistic substitution parameters, a complicated dataset may easily fall subject to model misspecification errors when characteristics of the data do not match the assumptions of the underlying model, possibly resulting in false biological conclusions.

Because each recombination detection method has different advantages and limitations, it is important to know which methods are appropriate for a given dataset. Past simulation studies have evaluated methods for recombination detection under a variety of evolutionary scenarios, including various levels of divergence and recombination (Wiuf et al. 2001; Brown CJ et al. 2001; Posada and Crandall 2001; Chan et al. 2006). However, these methods have not been evaluated for their performance under some complex evolutionary scenarios, such as those relevant to the process of functional divergence.

1.4 MOTIVATION FOR THESIS

With the increasing availability of genetic and genomic data, there is an increasing desire to carry out large-scale investigations of molecular adaptation (*e.g.,* Chen et al. 2006; Dunn et al. 2009). Codon models are among the more attractive means of exploring varying selection pressure on a gene because they permit estimation of an easily interpretable measure (ω) of selection pressure. However, with increasing data, there is an ever-increasing push to apply codon models beyond originally intended limits. For example, codon models do not take into account non-stationary codon usage, such as that observed in *Prochlorococcus* genes. Indeed, other types of heterogeneity within a dataset may cause serious errors in codon models, leading to false biological conclusions in other settings (*e.g.,* Bao et al. 2008; Inagaki and Roger 2006).

Recombination can yield heterogeneous patterns of evolution within gene sequences and thus in some cases can lead to a very high rate of false positives for positive selection under codon models (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006). Therefore, analysis of adaptive evolution in any system where a history of recombination is plausible should be accompanied by recombination analysis. While several simulation studies have evaluated the performance of recombination detection at various levels of mutation and recombination (Wiuf et al. 2001; Brown CJ et al. 2001; Posada and Crandall 2001; Chan et al. 2006), those studies explore performance under scenarios that are less complex than many real datasets. In Chapter 2 of this thesis, I employ simulation to evaluate several methods of recombination detection under four basic scenarios: 1) varying tree shape, 2) positive selection, 3) non-stationary evolution, and 4) varying levels of recombination and divergence.

In addition to recombination, non-stationary evolution may also cause problems for codon models. Despite our lack of knowledge about the impact of non-stationary codon usage on inferences about natural selection pressure, codon models have been, and continue to be, used to draw conclusions about the biology of *Prochlorococcus*. Without such knowledge, we simply do not know how to interpret accounts of molecular adaptation in *Prochlorococcus*. For instance, how reliable is the Zhao and Qin (2007) claim that the *cpeB* gene, which is related to a light harvesting pigment, is subject to positive Darwinian selection? Chapter 3 of this thesis presents an extensive simulation study to explore the impact of non-stationary evolution on the inference of selection pressure.

In Chapter 4, I conduct an extensive reanalysis of the *cpeB* gene that I interpret in light of the information acquired in Chapter 3. Zhao and Qin (2007) originally analyzed this gene using branch-site codon models, ignoring the fact that non-stationary codon bias present in this gene violates an assumption of the underlying model. Based on their analysis, they conclude that HL *Prochlorococcus* has experienced positive selection in this particular gene. To determine whether this conclusion can be supported or could be a false positive due to model misspecification, I conduct a thorough reanalysis of the *cpeB* gene. Here, using a more focused approach than employed in Chapter 3, I conduct simulations based on parameter values derived from the *cpeB* gene. In addition to providing insight into the selective pressures experienced by the *cpeB* gene, this study serves as an example of evolutionary analysis under non-stationary conditions.

Because of *Prochlorococcus'* divergence into two well-separated ecotypes, it is a prime candidate for a large-scale survey of genomic adaptation. Although several large-scale genomic surveys have explored the evolution of *Prochlorococcus* genomes, most have sampled only a few lineages and employ simple methods (*e.g.,* Hess et al. 2001; Rocap et al. 2003). In addition, most studies focus on the role of the flexible genome in *Prochlorococcus* divergence, ignoring the possible role of core genes in adaptation. For example, Kettler and colleagues (2007) analyzed all 12 published genomes, but focus on the impact of gene gain/loss in *Prochlorococcus* ecotype divergence. Another recently published study examined the role of selection on the core genome in the divergence of HL and LL ecotypes, but conducts only a pairwise analysis of three different concatenated genomes (Paul et al. 2010). Initially, an objective of this research was to utilize all 12 genomes to investigate the role of the core genome in the divergence of HL and LL ecotypes of *Prochlorococcus*. However, preliminary analyses of the real data, combined with simulation studies, revealed that this is not an easy task given the limitations of existing methodologies. While a large-scale genomic analysis is not feasible, the simulation studies of Chapters 2 and 3 are followed up with an exploratory application of the evaluated models to a subset of the *Prochlorococcus* core genome. These studies provide the first steps towards a comprehensive view of core genomic changes associated with the divergence of light-based metabolism in *Prochlorococcus*.

CHAPTER 2: RECOMBINATION DETECTION UNDER SCENARIOS RELATED TO FUNCTIONAL DIVERGENCE

2.1 INTRODUCTION

Genetic recombination can facilitate the process of functional divergence by allowing organisms to access an "extended gene pool." Through the exchange of genetic material between organisms, or even between species, recombination increases the genetic diversity in a population, which can help the population evolve. Examples of recombination-assisted evolution are well documented in bacteria (*e.g.,* Koonin et al. 2001; Boucher et al. 2003; Narra and Ochman 2006). For example, homologous recombination (HR) and lateral gene transfer (LGT) have impacted the evolution of several photosynthesis-related genes in cyanobacteria (*e.g.,* Mann et al. 2003; Zeidner et al. 2005; Sullivan et al. 2006; Lindell et al. 2004; Mann et al. 2005). However, detecting the signatures of adaptive substitutions within recombinant gene sequences is a challenge, as methods for modeling the substitution process typically assume that sequences are non-recombinant.

A clear picture of recombination events within a gene not only provides an understanding of gene evolution, but also helps avoid error in phylogeny-based analysis. Recombination events may cause individual gene or gene fragment phylogenies to be incongruent with the evolutionary history of the organism (*e.g.,* Ochman et al. 2000; Ragan 2001). While this phylogenetic disagreement may be useful for understanding the recombination process, it may also cause errors in analyses that depend on an accurate phylogeny. For example, methods that detect positive selection at the codon level have been shown to yield false positives when recombinant segments are present within gene sequences (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006). For this reason, studies of functional divergence among organisms where recombination is plausible should be accompanied by an analysis of recombination, as it may lead the identification of phylogenetic variation in different segments of a gene.

Several types of methods are available for the detection of recombination. Substitution-distribution methods (*e.g.,* GENECONV, MaxChi, Chimaera) detect regions that have significant sequence similarity based on a scoring method. Basic branch-pattern methods (*e.g.,* RDP) detect recombination by searching for variability in tree topologies of adjacent sequence fragments. In addition, more computationally complex methods are available. Examples include Bayesian methods (*e.g.,* DualBrothers, BARCE), which use posterior probabilities to search for adjacent regions with discordant phylogenies, and the Genetic Algorithm for Recombination Detection (GARD), which uses a likelihood-based heuristic algorithm to find the best-fit number and location of recombination breakpoints. A more detailed description of these methods can be found in Section 1.3.1.

Previous simulation studies evaluated methods of recombination under various conditions. Posada and Crandall (2001) simulated varying levels of diversity, recombination, and rate variation, finding that in general recombination detection methods are not powerful, although power does increase with diversity. In addition, few false positives occur in their simulations. Wiuf and colleagues (2001) found that certain combinations of branch lengths (*e.g.,* short internal branches and long tips) might cause detection methods to have increased power. Chan and colleagues (2006) found, as expected, that post-recombination substitutions decrease the ability of methods to detect breakpoints. While these simulations evaluate recombination detection methods under a wide variety of evolutionary patterns, there are still several conditions relevant to cases of functional divergence under which we know little about their performance.

To further explore the performance of recombination detection methods, I evaluate several of them using four simulation studies designed to cover parameters that are especially relevant to functional divergence analysis: 1) tree shape, 2) positive selection, 3) non-stationary evolution (including functional divergence), and 4) varied recombination and diversity. Simulations 1-3 contain no recombination and so are used to evaluate the false positive rates of selected methods under a range of conditions. In Simulation 4, I use recombinant sequences previously simulated by Kosakovsky Pond and colleagues (2006) to evaluate the power of the same methods used in Simulations 1-3. I then carry out an exploratory analysis of recombination in the core genome of the cyanobacteria *Prochlorococcus*, which is known to have complex evolutionary patterns.

2.2 Simulations and Analyses

For this study, I choose to apply three substitution-distribution methods (GENECONV, MaxChi, and Chimaera) and a basic branch-pattern method (RDP), because all were found to be relatively powerful by Posada and Crandall (2001). In addition, I also evaluate the more recently published GARD method which, based on the authors' simulations, has high accuracy (Kosakovsky Pond et al. 2006). Because the substitution-distribution and basic branch-pattern methods are based on simple summary statistics measured empirically from the data rather than complex models of evolution, they might be less subject to the negative effects of model misspecification. Using the same simulations to compare those four methods with one or more of the more recent methods may be very informative. Both the Bayesian methods and GARD are based on explicit models of evolution, which should make them more powerful, but may also make them prone to errors when model assumptions are violated. Of these more complex methods, I choose to apply only the GARD method because it is less computationally intensive than the Bayesian methods and able to deal with large numbers of taxa.

All five selected methods (Table 2.1) are applied to a series of four simulation studies. MaxChi, Chimaera, RDP, and GENECONV are included in the RDP3 software package (Martin 2009) and GARD is part of the HyPhy software package (Kosakovsky Pond and Muse 2005b). The authors of GARD suggest a single breakpoint analysis (hereafter called "GARD-SBP") for a qualitative determination of the presence of recombination (Kosakovsky Pond et al. 2006). Therefore I apply GARD-SBP, which is part of HyPhy, to Simulations 1-3. This method employs a maximum likelihood framework to conduct rapid screening for a single breakpoint with discordant phylogenies on either side. The genetic algorithm for multiple breakpoint analysis that is also implemented under GARD will hereafter be referred to as GARD-MBP. When recombination is detected under GARD-MBP, a Kishino-Hasegawa (KH) test (Kishino and Hasegawa 1989) is employed with a Bonferroni correction for multiple testing. The purpose of the KH test is to determine whether phylogenies before and after a putative breakpoint are significantly different. Although the correction for multiple tests makes GARD-MBP more conservative, resulting in lower power, it also helps to control the rate of false positives (Kosakovsky Pond et al. 2006) and so will be applied in the simulation studies.

Table 2.1. Recombination detection methods evaluated in Chapter 2.

Method	Type	Reference
GENECONV	Substitution	Sawyer (1989)
MaxChi	Substitution	Maynard Smith (1992)
Chimaera	Substitution	Posada and Crandall (2001)
RDP	Basic Branch-Pattern	Martin and Rybicki (2000)
GARD-SBP	Likelihood Phylogeny	Kosakovsky Pond et al. (2006)
GARD-MBP	Genetic Algorithm	Kosakovsky Pond et al. (2006)

Note: Because the single breakpoint method is used as a supplement for GARD-MBP, I refer to it throughout this thesis as GARD-SBP. However, GARD-SBP does not use a genetic algorithm, instead it employs a maximum likelihood framework to search for discordant phylogenies.

Three separate simulation studies are used to measure how the number of false positives yielded by each method might depend on 1) tree shape, 2) positive selection, or 3) non-stationary evolution. Because no recombination is simulated in these sequences, I am concerned only with the number of replicates that contain a false signal for recombination rather than the number of breakpoints detected. Replicates are considered to have significant evidence for recombination if, for a given method, at least one breakpoint is detected ($p \leq 0.05$). In addition, for methods that employ phylogenies (RDP and GARD) a recombination event is only considered significant if there is phylogenetic incongruence on either side of the breakpoint. In the fourth simulation study, I examine power for each method when multiple breakpoints are present. In that study, the number of inferred breakpoints is counted for each method. These counts are then compared to the number of simulated breakpoints.

2.2.1 Simulation 1: Tree Shape

Methods for recombination detection may be impacted by tree topology. Most simulations in the literature employ either symmetric trees or the empirical estimate of a tree for a gene of interest (*e.g.,* Posada and Crandall 2001; Chan et al. 2006). However, because most methods measure parameters from the entire alignment, tree shape (specifically asymmetry in tree topology) may impact recombination detection by increasing the number of informative sites and skewing the parameter estimates. The only reference that alludes to this problem is the RDP manual (Martin 2009), which suggests that having both closely related and very divergent sequences in an alignment may result in errors for some methods. Some recombination detection methods have been shown to have different performance when branch lengths within a tree are varied (Wiuf et al. 2001), but no studies have explored the impact of tree shape. Simulation 1 is designed to explore the effects of tree shape on false positive rates for recombination detection. In addition, because divergence is known to impact recombination detection, each tree shape is simulated under a range of lengths.

To explore the effects of tree shape, I use INDELible (Fletcher and Yang 2009) to simulate a range of tree lengths for two different topologies. Sequences are simulated based on either a 10 taxa asymmetric topology or a 16 taxa symmetric topology (Figure 2.1). The evolutionary model is homogeneous throughout the phylogeny, with no positive selection. All non-stop codons have equal frequencies and the transition/transversion ratio (κ) is set to 2. Alternative levels of sequence divergence are achieved by altering the internal branch lengths as follows: for the asymmetric topology, all internal branch lengths, as well as the most recently diverged tip, are equal to either

0.3, 0.2, 0.1, or 0.05 substitutions per codon site and the all terminal branches are adjusted so the tree is consistent with a molecular clock. For the symmetric topology, all internal branch lengths are set to either 0.3 or 0.05. For each level of sequence divergence, 50 replicate datasets with an alignment length of 200 codons are simulated.

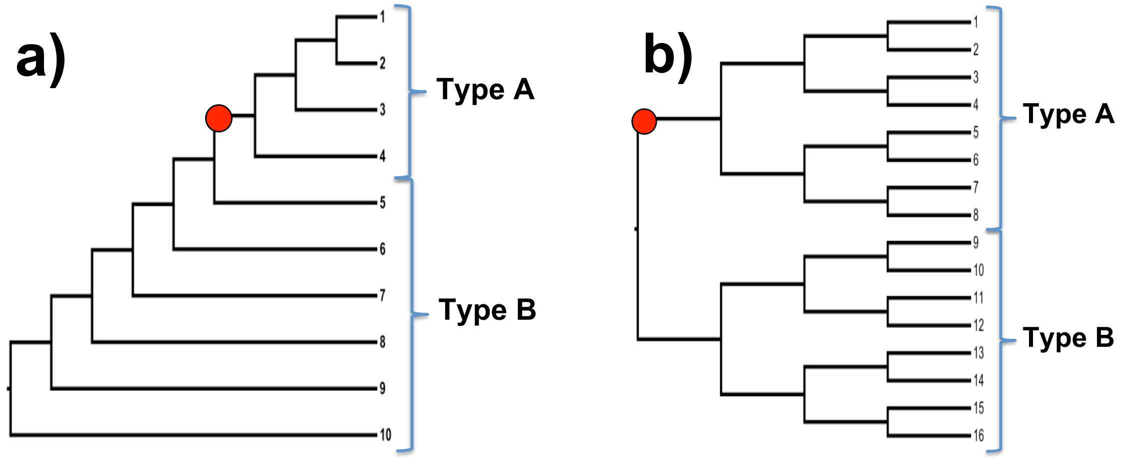


Figure 2.1. a) Asymmetric and b) Symmetric trees used for simulation studies. In Simulations 2 and 3, shifts in selection pressure and codon bias occur at the point represented by the red circle, which separates Type A and Type B evolution. Type A evolution differs from Type B due to a shift in the evolutionary process at this point.

All recombination detection methods shown in Table 2.1 are applied to the simulated data. Results indicate that tree topology can have a large impact on the false positive rate for some recombination detection methods (Table 2.2). All methods yield low levels of false positives under a symmetric topology. Because the type-1 error rate is expected to be equal to the level of the test (α=0.05), false positive rates below 5% are not considered significant. These low false positive rates are consistent with those found by Posada and Crandall (2001) and Kosakovsky Pond and colleagues (2006). However, GARD-SBP, MaxChi, and Chimaera yield much higher false positive rates under an asymmetric topology. When topology is asymmetric, the alignment contains sequences that are both closely and distantly related. Because MaxChi and Chimaera search for changes in the proportion of variable and invariable sites, which are estimated from the entire alignment, conserved sites in closely related sequences may give a false signal for recombination. Likewise, GARD-SBP measures substitution parameters from the entire alignment, which may cause false positives when divergence is more heterogeneous among sites, such as in an asymmetric tree.

Table 2.2. Percent false positives (*n*=50) for each recombination detection method under a) symmetric and b) asymmetric topologies with different tree lengths.

a)

	GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	Chimaera
Short (0.20 subst./codon)	0	3	2	2	8	6
Long (1.20 subst./codon)	0	2	4	0	4	8

b)

	GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	Chimaera
Short (0.45 subst./codon)	2	16	4	2	22	20
Long (2.70 subst./codon)	0	28	8	4	46	48

Note: Tree length in parentheses indicates root-to-tip length.

Although the simulated tree length has little effect on false positive rate under a symmetric tree, increased tree length has a large impact on the rate of false positives under an asymmetric tree for some methods (Table 2.2; Figure 2.2). When the phylogeny is asymmetric, an increase in tree length leads to an increase in false positives for Chimaera, MaxChi and GARD-SBP. Chimaera yields the highest level of false positives, finding significant evidence for recombination in 48% of replicates simulated under the longest tree (0.3 substitutions per codon site for internal branches) compared to 20% under the shortest (0.05 substitutions per codon site for internal branches). MaxChi has similar performance, with 46% false positives under the longest tree and 22% in the shortest. Although GARD-SBP detects fewer false positives than Chimaera or MaxChi, the false positive rate still increases with tree length, with false positive rates ranging from 28% in the longest tree to 16% in the shortest. Previous studies, carried out under symmetric tree topologies, have shown that for many methods power to detect recombination events increases with sequence divergence, likely because the increased variability provides additional information for recombination detection (Posada and Crandall 2001). For similar reasons, the increase in variable sites may result in more opportunities for the false detection of recombination events. For some methods (GARD-SBP, MaxChi, and Chimaera), this effect appears to be heightened under an asymmetric tree, likely due to an increasing disparity between conserved and quickly evolving sequence segments.

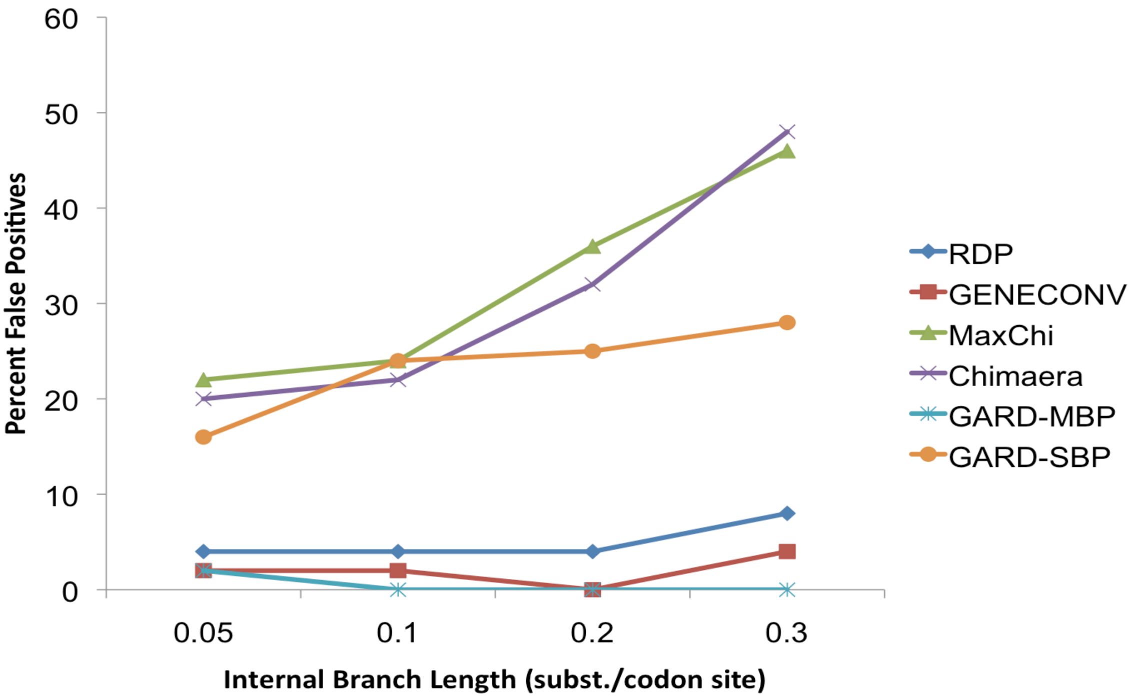


Figure 2.2. Percent false positives (*n*=50) for Simulation 1 under asymmetric tree with increasing tree length.

The other methods, GENECONV, RDP, and GARD-MBP, yield consistently low levels of false positives across the simulated levels of tree length and shape (Table 2.2; Figure 2.2). GARD-MBP and GENECONV are most noteworthy, in that levels of false positives are insignificant, ranging from 0-4%, and do not increase with tree length or asymmetry. RDP has only slightly higher false positive levels, with a very minor increase in false positives with both tree length and asymmetry, reaching a maximum of 8%. These results are comparable to those from previous work (Posada and Crandall 2001), which also found RDP and GENECONV to have low rates of false positives. My results further indicate that these methods are more robust to differing tree shapes than are MaxChi, Chimaera, or GARD-SBP.

2.2.2 Simulation 2: Positive Selection

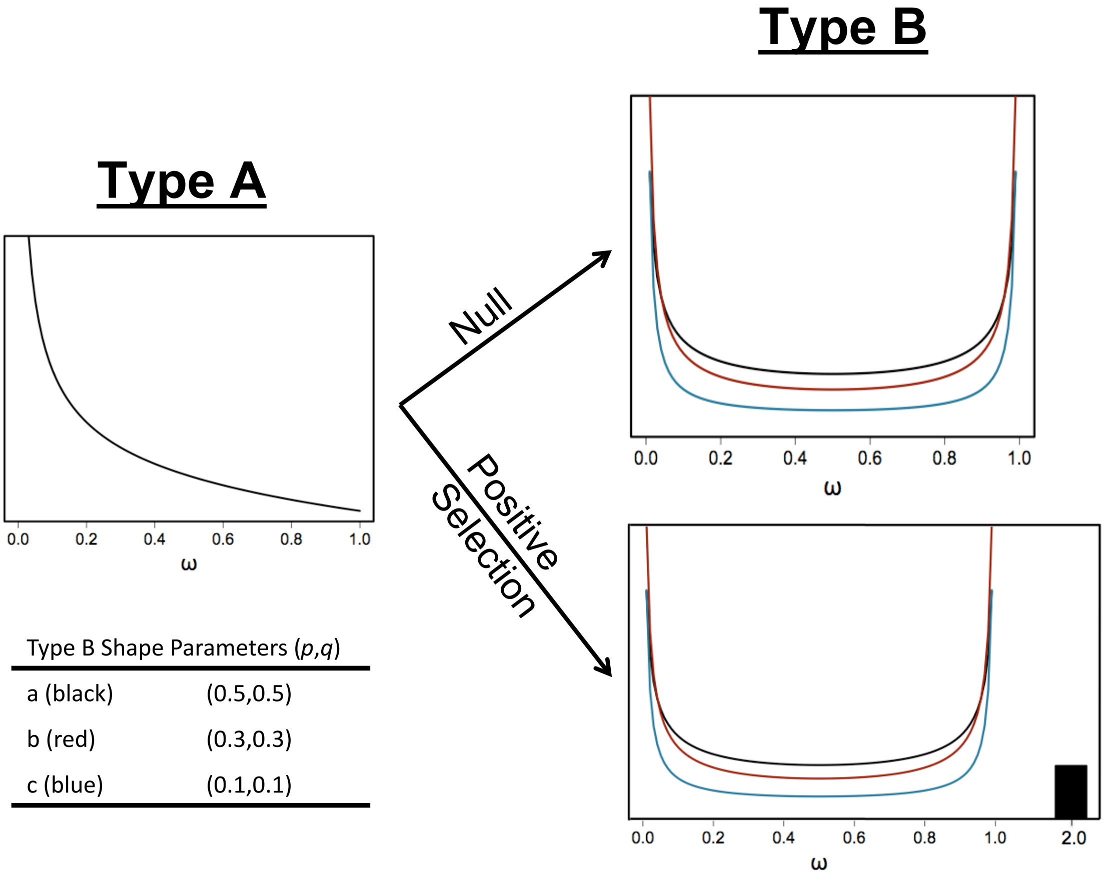
Recombination is known to negatively impact statistical tests for detecting positive selection at the codon level (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006). Genetic variability generated by recombination may resemble patterns of molecular adaptation because substitution rates of recombinant gene fragments may differ from the rest of the alignment. However, because the patterns of variability created by recombination and adaptation are similar, the actual presence of sites in a dataset subject to positive selection may likewise affect statistical tests for recombination. For this reason, Simulation 2 evaluates the performance of recombination detection methods when some of the sequences in an alignment are under positive selection.

For this study, I simulate datasets in which positive selection is present in part of the phylogeny. Simulated datasets, each 200 codons in length, are based on either the 10 taxa asymmetric phylogeny or the 16 taxa symmetric phylogeny in Figure 2.1. All internal branch lengths in the phylogeny are set to 0.3 substitutions per codon site and all other branch lengths are adjusted to simulate rate constancy. When a shift to positive selection is simulated, it occurs at the point shown in red in Figure 2.1 and is present in all branches that evolve after that point. This effectively splits the tree into two types, "Type A" and "Type B," with different evolutionary models (Figure 2.1).

The strength and direction of selection pressure is simulated at the codon level by specifying a distribution for ω ($\omega = d_N/d_S$) separately for each part of the phylogeny. Omega distributions are modeled using a beta function, which is convenient for this purpose because its range from zero to one is ideal for modeling an omega distribution with no positive selection ($0<\omega<1$) while employing only two shape parameters (p,q). To simulate positive selection in a fraction of sites, a single discrete category is added in which ω=2. Type A evolution remains constant throughout all simulation conditions while Type B evolution changes (Figure 2.3). For the Type A evolutionary model, most codon sites are under strong purifying selection (ω<<1) and very few sites are evolving close to neutrality (p=0.5, q=2). Type B evolution is simulated with 10% of sites under positive selection. The remainder of sites are distributed according to a U-shaped beta function, with a large proportion of sites evolving nearly neutrally. Sites having ω between 0 and 1 are divided into nine discrete site categories. I simulate three different sets of shape parameters (hereafter referred to as a, b, and c) for the omega distribution, each with a tenth site category under positive selection (Figure 2.3). I also simulate three "null" cases (*i.e.* no positive selection) under the same shape parameters (a, b, and c). For each condition, 50 replicate datasets are simulated.

As observed in Simulation 1, data having an asymmetric topology in Simulation 2 yield a consistently higher rate of false positives for some methods (Table 2.3). GARD-MBP and GENECONV yield insignificant rates of false positives (0-6%). Surprisingly, the presence of positive selection in some sequences does not affect the false positive rate for most of the methods (Table 2.3). GARD-SBP is the one exception, yielding increased levels of false positives when positive selection is present, but this effect only occurs when tree topology is asymmetric. Overall, methods are robust to the presence of positive selection in part of the phylogeny.

While positive selection in a portion of the phylogeny may not affect recombination detection, the presence of positive selection throughout the entire phylogeny might differently impact these methods. To test this hypothesis, I perform a small simulation study in which positive selection is present throughout the entire phylogeny (See Appendix D for detailed methods and results). None of the methods applied show increased rates of false positives under positive selection. Collectively, these results suggest that recombination detection methods might be largely robust to the presence of sites subject to positive selection.



Type B Shape Parameters (*p*,*q*)	
a (black)	(0.5,0.5)
b (red)	(0.3,0.3)
c (blue)	(0.1,0.1)

Figure 2.3. Omega distributions used in Simulation 2. Distributions are modeled by a beta function and three different sets of shape parameters (*p*,*q*) are represented by different colored curves: a) black, b) red, and c) blue.

Although studies have shown that recombination events can lead to false inferences in positive selection analysis (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006), my results show that the reverse is not true. One possible explanation is that positively selected sites, while often localized in 3D space of the folded protein product of the gene, are typically spread out along a gene sequence. Recall that recombination detection methods search for local variability in adjacent gene fragments; *i.e.*, they search for spatial organization along the gene sequence. However, this pattern is unlikely to result from selection acting on the mature and folded protein product. In addition, while positive selection results in a shift in substitution parameters, it does not change phylogenetic relationships, so methods that require phylogenetic variability among sites would not likely yield false positives for recombination under such conditions.

Table 2.3. Percent false positives (*n*=50) yielded in Simulation 2.

			GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	Chimaera
Symmetric	Null	a	0	2	2	2	14	16
		b	0	0	0	2	10	8
		c	0	2	4	2	8	12
	Positive Selection	a	0	2	6	0	8	10
		b	0	2	8	0	14	16
		c	0	2	2	2	6	6
Asymmetric	Null	a	0	78	18	6	46	48
		b	0	66	12	0	40	44
		c	0	64	18	2	40	46
	Positive Selection	a	0	82	6	4	36	42
		b	0	70	6	2	28	36
		c	0	90	14	2	38	44

Note: For both trees, the internal branch length is equal to 0.3 subst./codon site. Root-to-tip lengths are therefore 1.2 subst./codon (symmetric) and 2.7 subst./codon (asymmetric).

2.2.3 Simulation 3: Non-Stationary Evolution

Current recombination detection methods often assume that evolutionary processes are homogeneous over time. Parameters such as nucleotide composition and substitution parameters are often averaged over an entire phylogeny. However, these assumptions are often violated in real data. For example, sequences that have experienced a divergence in gene function may have evolutionary rates that change across a phylogeny. Simulation 3 is designed to explore the impact of non-stationary codon bias and selective constraints on methods for recombination detection.

For this simulation, I use the same basic pattern as Simulation 2, expanding on the evolutionary schemes. Sequences are simulated under the same two phylogenies (symmetric and asymmetric) and each alignment is 200 codons in length. When a shift in the evolutionary process occurs, it takes place in the same phylogenetic location as in Simulation 2, again dividing the tree into two "types" of evolution (Figure 2.1). However, for Simulation 3, I simulate additional shifts in both codon bias and selection pressure to investigate the effects of more complex shifts in selection pressure, but without the presence of positively selected sites.

Codon bias is modeled using the method of Aris-Brosou and Bielawski (2006). This method employs a single parameter, "η," to specify codon frequencies for changing GC3 content. The values of η range from $0 \leq \eta \leq 1$, where a value of $\eta=0.5$ indicates a GC3 content of 50%; all codons that do not code for a stop codon have equal frequencies. As η approaches zero, GC3 content increases (Figure 2.4). Using this system, I can easily calculate and specify separate codon biases for different parts of the phylogeny.

SECOND CODON POSITION

FIRST CODON POSITION	T	C	A	G	THIRD CODON POSITION
T	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	T
	η/Σ	η/Σ	η/Σ	η/Σ	C
	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	0	0	A
	η/Σ	η/Σ	0	η/Σ	G
C	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	T
	η/Σ	η/Σ	η/Σ	η/Σ	C
	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	A
	η/Σ	η/Σ	η/Σ	η/Σ	G
A	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	T
	η/Σ	η/Σ	η/Σ	η/Σ	C
	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	A
	η/Σ	η/Σ	η/Σ	η/Σ	G
G	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	T
	η/Σ	η/Σ	η/Σ	η/Σ	C
	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	$(1-\eta)/\Sigma$	A
	η/Σ	η/Σ	η/Σ	η/Σ	G

Figure 2.4. Scheme used for simulation of GC3 bias. The single parameter η ($0<\eta<1$) is used to simulate codon frequencies based on the desired GC3 content. The denominator Σ is a scaling factor to ensure frequencies sum to 1.

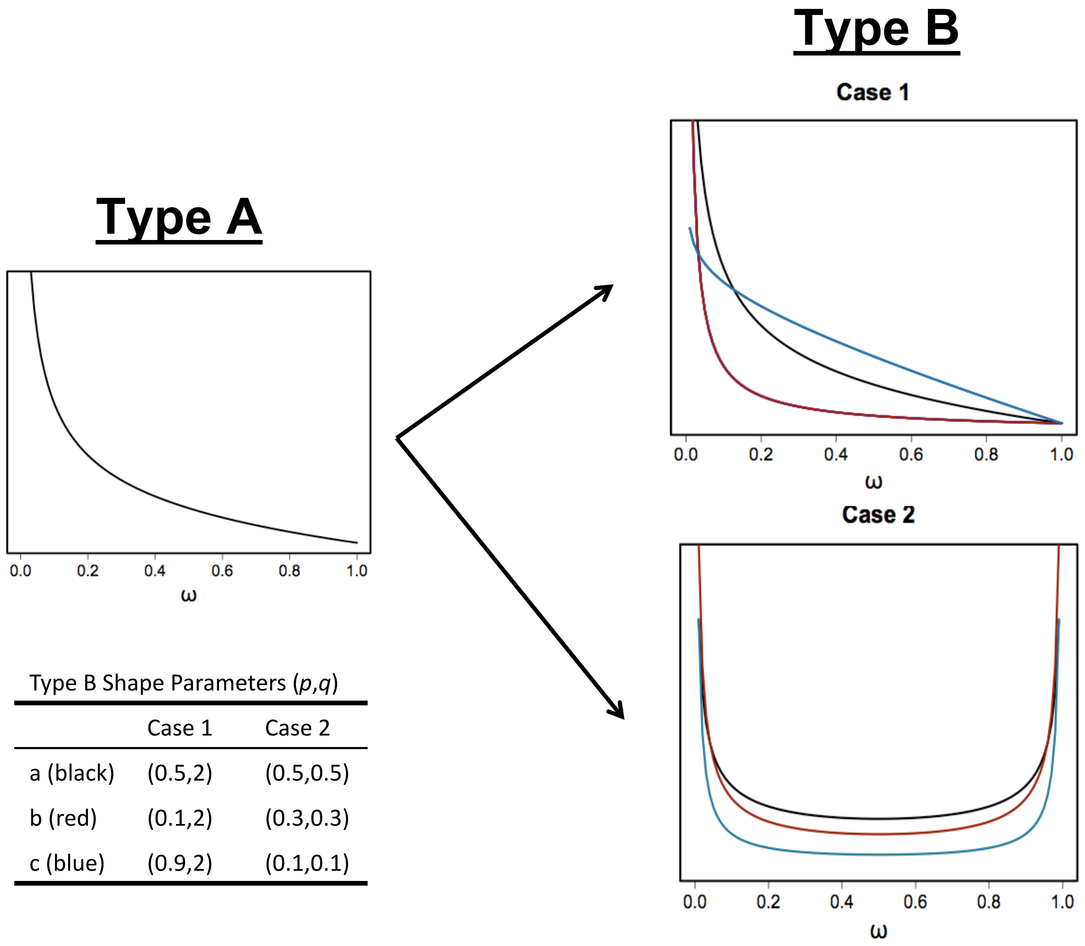
As before, the two parts of the tree (A and B) are simulated under different models of evolution. Type A evolution is simulated under the same model as in Simulation 2, with most sites under purifying selection (p=0.5, q=2). The codon bias parameter for Type A evolution is set to η_1=0.5 (GC3≈50%). Type B evolution, however, varies for each simulation condition. There are two basic cases for Type B evolution (Figure 2.5):

Case 1: Most sites under strong purifying selection, similar to Type A evolution. This is modeled by an L-shaped beta function.

Case 2: A larger proportion of sites evolving close to neutrality. This is modeled with a U-shaped beta function.

For each general case, I simulate 3 different ω distributions (designated sub-cases "a," "b," and "c") (Figure 2.5). Case 1a is the "null"; Type B evolution is simulated under the same conditions as Type A evolution, so selection pressure is homogeneous across the phylogeny. In all other cases (Cases 1b, 1c, and 2a-c) there is a shift in ω distribution. In addition, each case is simulated both with a shift in codon bias (η_2=0.1: GC3≈10%) and with no shift in codon bias (η_2=0.5) across the phylogeny. The result is a total of 12 unique evolutionary conditions (Figure 2.5). For each condition, 50 replicate datasets are simulated.

Table 2.4 shows the results under the null case (Case 1a), with no shift in selection pressure. Here, I examine the impact of non-stationary codon bias, while all other parameters remain homogeneous. While a shift in codon bias alone does not affect the false positive rates under a symmetric tree, for some methods (RDP, MaxChi, and Chimaera) non-stationary codon bias does have a small effect on false positive rates under an asymmetric tree (Table 2.4). However, even under an asymmetric tree, this effect is not large, with a maximum of 8% increase in false positives due to non-stationary codon bias. Because these methods estimate parameters from the entire alignment, a shift in codon bias may result in a false signal for recombination.



Type B Shape Parameters (*p*,*q*)

	Case 1	Case 2
a (black)	(0.5,2)	(0.5,0.5)
b (red)	(0.1,2)	(0.3,0.3)
c (blue)	(0.9,2)	(0.1,0.1)

Figure 2.5. Omega distributions for Simulation 3. Sub-cases are designated by different colored curves a) black, b) red, and c) blue. Each sub-case is simulated with both stationary and non-stationary codon bias.

Using the remaining cases (Case 1b, 1c, and 2a-c), I examine the effects of a combined shift in the distribution of selection pressure and codon bias. The full data are presented in Appendix E, but to aid presentation, I present averaged false positive rates over all selection pressure schemes (Table 2.5). Taken over the different types of shifts in selection pressure, differences due to non-stationary codon bias are, again, small. Although some methods do show a slight increase in false positives under non-stationary codon bias, this increase does not exceed an average of 5% for any method. As expected, false positive rates are generally higher for most methods under an asymmetric tree, but these values still do not substantially increase when codon bias is non-stationary. Taken together with previous results, this suggests that methods for recombination detection are generally robust to shifts in codon bias, with some methods experiencing a minor effect.

Table 2.4. Percent false positives (*n*=50) for recombination detection methods under homogeneous selection pressure (Case 1a) for both asymmetric and symmetric tree topologies with both stationary and non-stationary codon bias.

		GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	Chimaera
Symmetric	Stationary	0	4	10	0	12	14
	Non-stationary	0	6	6	0	10	14
Asymmetric	Stationary	0	58	10	0	32	36
	Non-stationary	2	58	18	2	38	44

Note: For both symmetric and asymmetric trees, the internal branch length is equal to 0.3 subst./codon site. This results in root-to-tip lengths of 1.2 subst./codon (symmetric) and 2.7 subst./codon (asymmetric).

Table 2.5. Percent false positives when non-stationary codon bias is combined with a shift in selection pressure.

		GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	Chimaera
Symmetric	Stationary	0	2	4	1	11	14
	Non-stationary	0	6	6	1	11	16
Asymmetric	Stationary	0	75	10	1	31	38
	Non-stationary	1	69	14	2	34	43

Note: Values averaged across different shifts in selection pressure (Cases 1b, 1c, 2a-c). For both symmetric and asymmetric trees, the internal branch length is equal to 0.3 subst./codon site. This results in root-to-tip lengths of 1.2 subst./codon (symmetric) and 2.7 subst./codon (asymmetric).

This simulation design also permits an investigation of the effect of shifts in the distribution of selection pressure characteristic of cases of functional divergence. Here, I calculate average false positive rates separately for Cases 1 and 2 (Table 2.6). Because Case 1a is the null, where selection pressure is stationary, it is not combined with Cases 1b and 1c in Table 2.6. Note that there is much variability in false positive rates among sub-cases a, b, and c, but no consistent pattern (Appendix E). Under a symmetric tree, a shift in selection pressure does not cause a consistent increase in false positive rates as compared to those under the null condition (Table 2.6). The asymmetric tree is similar in that there is no systematic increase in rates of false positives. However, the one exception is the GARD-SBP method, which yields substantially higher false positives for both Cases 1 and 2 under an asymmetric tree. GARD-SBP estimates substitution parameters from the entire alignment and uses these in maximum likelihood estimations of phylogenies for each gene fragment. Therefore, this method may be especially prone to false positives under non-stationary selection pressures where the effect is variable among sites. GARD-SBP does not explicitly model anything other than rate variability among sites; when other types of variability exist among sites (such as in the simulations), the only way it can be accommodated is via rate variation among sites. The estimated rate variation among sites under GARD-SBP appears to be "soaking up" other aspects of the evolutionary process and this signal is mistaken as a signal for recombination.

Table 2.6. Percent false positives under different shifts in selection pressure for both symmetric and asymmetric tree topologies.

		GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	Chimaera
Symmetric	Null (Case 1a)	0	5	8	0	11	14
	Case 1 (b,c)	0	2	3	1	11	17
	Case 2 (a,b,c)	0	5	6	2	11	16
Asymmetric	Null (Case 1a)	1	58	14	1	35	40
	Case 1 (b,c)	1	84	8	2	24	38
	Case 2	0	76	14	3	40	44

Note: Values averaged over stationary (η_2=0.5) and non-stationary (η_2=0.1) codon bias. For both symmetric and asymmetric trees, the internal branch length is equal to 0.3 subst./codon site. This results in root-to-tip lengths of 1.2 subst./codon (symmetric) and 2.7 subst./codon (asymmetric).

2.2.4 Simulation 4: Varied recombination and diversity

Because recombination analysis is a key step in phylogeny-based inference, it is important that detection methods be reliable when recombination has truly impacted the evolution of a set of gene sequences. Methods designed to detect recombination events must not only detect whether or not recombination is present, but also estimate the number and location of breakpoints. Several simulation studies have evaluated the power of these methods under different levels of recombination and divergence (Posada and Crandall 2001; Wiuf et al. 2001; Chan et al. 2006; Kosakovsky Pond et al. 2006). However, except for GARD-MBP (Kosakovsky Pond et al. 2006), none of the methods have been evaluated for their ability to determine the correct number of breakpoints when multiple are present. In Simulation 4, I analyze the power of recombination detection methods to qualitatively detect the presence of recombination and to accurately detect the number of breakpoints.

For this simulation, I use datasets from a previous study that have been analyzed only with GARD-MBP (Kosakovsky Pond et al. 2006). These simulated datasets consist of 8 taxa alignments with different levels of recombination and diversity. Each alignment is 3,000 bp long and has 0, 1, 2, 4, or 8 recombination breakpoints. In addition, for each number of breakpoints, there are datasets with both low (5%) and high (25%) genetic diversity for a total of 10 simulation conditions, each with 100 replicate datasets.

Consistent with previous simulation studies, this analysis shows that recombination detection methods are not typically powerful (*e.g.,* Posada and Crandall 2001). Power for detecting a single recombination event is low for all methods (Table 2.7). When just one recombination event is simulated and diversity is low, RDP, GENECONV, MaxChi, and Chimaera have similar performance; detecting just 12-19% of replicates as having been subject to recombination. GARD-MBP has substantially lower power, only detecting recombination in 8% of replicates. This is much lower than previously reported for the same set of simulations (56%) (Kosakovsky Pond et al. 2006) because I apply the KH test for phylogenetic incongruence, which has the desirable effect of controlling the number of false positives (see Simulations 1-3).

As the number of simulated recombination events increases, so does the number of replicates in which recombination is detected (Table 2.7). When 8 breakpoints are simulated at low diversity, MaxChi and Chimaera detect recombination in 74-75% of replicates. RDP and GENECONV have slightly lower power, detecting recombination in 67-68% of replicates. These results are consistent with previous studies (Posada and Crandall 2001; Kosakovsky Pond et al. 2006), which show that recombination detection methods have higher power when levels of recombination are higher.

GARD-MBP has much lower power, detecting recombination in less than half of replicates even when multiple breakpoints are simulated. This is not consistent with previous results (Kosakovsky Pond et al. 2006) due to my use of the KH test for phylogenetic incongruence. Kosakovsky Pond and colleagues (2006) endorse using GARD-MBP without requiring phylogenetic incongruence on either side of a breakpoint, as this yields very good power (recombination is detected in as many as 98% of replicates when both recombination and diversity levels are high). However, it also increases the number of false positives in their simulations (10% in their simulations, as compared to 1% in my simulations where the KH test is applied).

Table 2.7. Capacity of five recombination detection methods to correctly infer 0 to 8 breakpoints.

		Low Diversity (5%)					High Diversity (25%)				
		0	1	2	4	8	0	1	2	4	8
RDP	0	99	83	73	49	32	98	62	53	19	11
	1	1	16	26	39	35	2	37	38	44	30
	2		1	1	11	25		1	8	25	30
	3				1	3			1	6	16
	4					5				4	7
	5+									2	6
	T	1	17	27	51	68	2	38	47	81	89
GENECONV	0	99	88	75	59	33	98	76	68	41	29
	1	1	12	25	32	47	2	23	27	43	43
	2				8	15		1	5	14	17
	3				1	2				2	8
	4					2					3
	5+										
	T	1	12	25	41	67	2	24	32	59	71
MaxChi	0	97	81	60	37	25	94	69	48	18	10
	1	3	19	36	42	38	6	28	40	40	28
	2			4	19	27		3	11	25	29
	3				2	6			1	14	20
	4					4				2	7
	5+									2	6
	T	3	19	40	63	75	6	31	52	83	90
Chimaera	0	97	84	61	41	26	94	69	48	18	11
	1	3	16	36	43	37	6	29	39	44	25
	2			3	14	26		2	12	25	31
	3				2	6			1	10	19
	4					5				3	9
	5+										5
	T	3	16	39	59	74	6	31	52	82	89
GARD-MBP	0	99	92	72	74	51	94	70	70	57	47
	1	1	8	28	24	43	6	30	25	38	34
	2				2	6			5	4	16
	3									1	3
	4										
	5+										
	T	1	8	28	26	49	6	30	30	43	53

Notes: For each method evaluated, columns indicate the number of simulated breakpoints while rows indicate the number of breakpoints inferred by a given method. Values indicate the number of replicates (*n*=100) that infer the number of breakpoints specified by the row label. The rows marked "T" indicate the total number of replicates with a signal for recombination for a given condition.

Although most methods have increased ability to qualitatively assess recombination with increasing number of breakpoints, accuracy in inferring the correct number of breakpoints when multiple are present is extremely low (<4% for low diversity and <12% for high diversity). In fact, even when 8 breakpoints are simulated, no method infers greater than 5 breakpoints when diversity is low. As I apply it, GARD-MBP is particularly conservative, never inferring more than two breakpoints at low diversity. In general all methods underestimate the number of recombination events. This decreasing ability to correctly identify multiple breakpoints may be due to decreased amount of information from which to make inferences. Small fragments simply do not provide enough information for accurate estimation of some parameters, including substitution parameters and phylogenies. Because recombination detection methods require parameter estimates to be compared with those in adjacent fragments, decreased fragment size may result in false negatives.

For all methods, increasing diversity increases power to detect recombination (Table 2.7). For example, in sequences with one simulated breakpoint, RDP detects recombination in 38% of alignments with high diversity, compared to just 16% of low diversity replicates. This general pattern of increasing recombination detection at high diversity is present throughout all levels of recombination. In addition, when multiple breakpoints are present, a larger number of replicates have a signal for more than one breakpoint when diversity is high. These results are consistent with previous findings, which suggest the increased information available when genetic variance is high leads to increased power for detection (Posada and Crandall 2001; Kosakovsky Pond et al. 2006).

However, even under high diversity, methods still do not have the power to accurately detect the number of breakpoints.

2.3 *PROCHLOROCOCCUS* GENOMIC DATA

Recombination has had a significant impact on the evolution of the cyanobacteria *Prochlorococcus* (*e.g.,* Mann et al. 2003; Zeidner et al. 2005; Sullivan et al. 2006; Zhaxybayeva et al. 2009). While several studies have explored the role of lateral gene transfer (LGT) in the genomic evolution of *Prochlorococcus* (*e.g.,* Zhaxybayeva et al. 2006; Zhaxybayeva et al. 2009), none have attempted to quantify within-gene recombination. The evolutionary history of *Prochlorococcus* is complex, including non-stationary nucleotide composition and evolutionary rate along with an asymmetric organismal phylogeny. According to my simulation studies, tree topology, and to a lesser extent codon bias, are conditions that can impact the performance of recombination detection methods. Here, I use the knowledge acquired from my simulation studies to evaluate a subset of the core genome of *Prochlorococcus* for recombination.

For the analysis of genomic data, I use a subset of 585 genes from the *Prochlorococcus* core genome. For the purposes of this study, the "core genome" contains all genes present in all 12 *Prochlorococcus* genomes. The subset I chose to use for recombination analysis consists of those genes with the same fundamental evolutionary history as the accepted phylogeny for the organism (See the 16S phylogeny in Figure 1.1). I start with amino acid alignments from a full set of 1812 genes, previously clustered into orthologous groups containing both *Prochlorococcus* and its closest relative, *Synechococcus,* by Zhaxybayeva and colleagues (2009). Nucleotide sequences from genomic data downloaded from GenBank (Table 2.8) are aligned using the amino acid alignments as templates. For each gene, RAxML (Stamatakis 2006) is used to create a phylogeny. Because the algorithm is start-point dependent, 10 maximum likelihood phylogenies are generated using a GTR model with a gamma distribution for among-site rate variation. The phylogeny with the best likelihood score is taken as the best estimate for that gene. A bipartition analysis is then conducted to select those genes with phylogenies that separate 1) *Prochlorococcus* from *Synechococcus*, 2) HL and LL *Prochlorococcus*, and 3) the two clades within HL *Prochlorococcus*. Exclusion of genes having other topologies effectively filters out many having experienced whole-gene LGT events (Zhaxybayeva et al. 2009). Thus, the remaining genes have phylogenies “closer” to the accepted organismal phylogeny, but nonetheless may have experienced one or more within-gene recombination events. A total of 585 genes from the core genome meet this criteria and thus are used for recombination analysis.

Table 2.8. *Prochlorococcus* strains used for the genomic analysis.

Strain	Ecotype	%GC	Accession	Reference
CCMP1986 (MED4)	HL	30.8	BX548174	Moore 1995; Rocap et al. 2003
MIT9515	HL	30.8	CP000552	Rocap et al. 2002
MIT9301	HL	31.3	CP000576	Rocap et al. 2002
AS9691	HL	31.3	CP000551	Shalapyonok et al. 1998
MIT9215	HL	31.1	CP000825	Moore and Chisholm 1999
MIT9312	HL	31.2	CP000111	Moore et al. 1998; Coleman et al. 2006
NATL1A	LL	35.0	CP000553	Partensky et al. 1993
NATL2A	LL	35.1	AE017126	Scanlan et al. 1996
CCMP1375 (SS120)	LL	36.4	CP000878	Dufresne et al. 2003; Chisholm et al. 1992
MIT9211	LL	38	CP000554	Moore and Chisholm 1999
MIT9303	LL	50.0	BX548175	Moore et al. 1998
MIT9313	LL	50.7	CP000435	Moore et al. 1998

Each gene in the subset is analyzed for within-gene recombination by using all recombination detection methods evaluated in the simulation studies (GENECONV, RDP, MaxChi, Chimaera, GARD-MBP, and GARD-SBP). The number of genes in which recombination is detected is extremely variable, from just 9 genes (1.5% of those analyzed) using GARD-MBP to 534 genes (91.3%) using Chimaera (Table 2.9).

Table 2.9. Results from recombination analysis of 585 *Prochlorococcus* genes.

Method	Number of Genes	Percent of Total
GARD-MBP	9	1.5
GARD-SBP	487	83.2
RDP	209	35.7
GENECONV	50	8.5
MaxChi	476	81.4
Chimaera	534	91.3

MaxChi is similar to Chimaera, detecting recombination in a large percentage of the genes. Based on the simulations, asymmetric topologies combined with longer branch lengths can cause MaxChi and Chimaera to detect false signals for recombination. While most branches within the *Prochlorococcus* topology are generally short, the branch between ecotypes as well as some of the branches within the LL grade are comparable to the longer branches used in the simulated asymmetric topology. In addition, the phylogenies of all *Prochlorococcus* genes are largely asymmetric. Because both of these scenarios are exactly those associated with false positives for Chimaera and MaxChi in my simulations, the values yielded by these methods for real *Prochlorococcus* genes should be considered inflated due to potentially large numbers of false positives for genes with recombination.

Interestingly, GARD-based methods infer among the highest and lowest numbers of genes having a history of recombination. GARD-MBP seems to be extremely conservative, detecting recombination in only 9 genes. This is consistent with results from the simulation studies, where GARD-MBP is shown to have fairly low power even when multiple breakpoints are present. However, the GARD-SBP method detects recombination in 83.2% of genes, with only Chimaera detecting higher levels. Given the high false positive rate of GARD-SBP under an asymmetric tree in Simulation 1, as well as under non-stationary evolution in Simulation 3, these results are best interpreted as negatively impacted by false positives.

RDP and GENECONV fall between these two extremes and, based on simulations, they are expected to more reliably identify actual cases of recombination in *Prochlorococcus*. In the simulations, these two methods show reasonable power without yielding a large number of false positives when branch length is increased. In addition, RDP and GENECONV perform well under an asymmetric topology and non-stationary evolution, which are characteristic of *Prochlorococcus*. However, the difference between levels of recombination detected with these two methods is still quite large; RDP detects recombination in 209 genes while GENECONV detects recombination in only 50. With such vastly different estimates, it is difficult to determine the actual level of recombination present in the *Prochlorococcus* core genome, but it is likely near the values detected by GENECONV and RDP (8.5 and 37.5%), suggesting an important role for within-gene recombination in the evolution of the *Prochlorococcus* core genome.

2.4 CONCLUSIONS

Analysis of recombination is an important part of any phylogeny-based analysis. However, the performances of different methods that are designed to detect recombination are impacted by a variety of evolutionary processes. Based on my results, I suggest that caution be exercised when choosing the method best suited for a given dataset. Different methods may lead to drastically different biological conclusions depending on the evolutionary forces acting on a gene.

Of the scenarios examined, I find tree shape to have the most substantial impact on false positive rate. Although all methods experience at least a slight increase in false positive rates under an asymmetric tree, some methods (GARD-SBP, MaxChi, and Chimaera) yield substantially higher false positive rates. In addition, false positive rates for these methods increase with increasing tree length when topology is asymmetric. It is therefore recommended that these three methods not be used when the gene in question has a topology that is not symmetric.

Somewhat surprisingly, most methods perform consistently under conditions of positive selection as well as under non-stationary codon bias and shifts in the distribution of selection pressure. None of the methods show a significant increase in number of false positives when positive selection is simulated. With the exception of the GARD-SBP method, none of the recombination detection methods yield substantially higher levels of false positives under non-stationary evolution as compared with stationary conditions. Two methods (GENECONV and GARD-MBP) stand out in that they are robust to the influence of both asymmetric tree topology and non-stationary codon bias.

As shown in previous studies, levels of divergence and recombination impact power to detect recombination events. Although power is typically low for all recombination detection methods, ability to detect recombination increases with both diversity and number of breakpoints. However, even when diversity is high, the power of these methods to accurately determine the number of recombination breakpoints in an alignment is very low. It seems that for all methods, there is a tradeoff between power and accuracy. While the more powerful methods (MaxChi, Chimaera) are more likely to accurately detect the presence of recombination when it truly exists, they are also more likely to detect a false signal for recombination when it is absent. On the other hand, the more conservative methods (GARD-MBP, GENECONV, RDP) have low false positive rates, but are also likely to miss recombination when it is present. It appears that a good tradeoff between power and accuracy is difficult to achieve with these statistical methods for detecting recombination.

Although using a combination of methods may be the best way to obtain a clear understanding of recombination present in a given dataset, one must be careful to choose methods that are appropriate and not prone to especially large false positive rates based on the characteristics of the data at hand. For example, the fraction of genes for which recombination is detected in the *Prochlorococcus* genome ranges from 1.5% to 91.3%: fundamentally different conclusions would be derived from each extreme if taken alone; the average does not seem to be a biologically defendable estimate; and, a consensus that includes a method with very low power could yield substantial underestimates. The key to avoiding extremely conservative or extremely liberal estimates of recombination on real data is to choose methods that have reasonable power while still controlling the type-1 error rate.

The problem of recombination detection is obviously a complex one and there is much room for improvement in developing these methods before we have a true understanding of recombination as an evolutionary process. For the time being, users of these methods are encouraged to apply the following guidelines:

- Do not ignore evolutionary characteristics such as non-stationary codon bias and functional divergence. Analysis for such evolutionary patterns should be carried out before conducting recombination analysis.

- When characteristics of a given dataset may violate model assumptions, simulations should be carried out under conditions that are relevant to the data at hand. These simulations should evaluate both power and false positive rates of candidate methods.
- Pay particular attention to tree shape and sequence divergence, as these characteristics may drastically affect performance of recombination detection methods.
- A consensus of several different methods may provide a robust way of detecting recombination within a gene, but methods to use in a consensus should be chosen carefully. For instance, taking into account a strict consensus across a set of methods that includes an extremely conservative one (*i.e.,* very low power) will be unlikely to detect recombination in most cases where it truly exists.
- Overall, my simulations find GENECONV to be both reasonably powerful and generally robust. Therefore, it may be desirable to include GENECONV when evaluating a set of potential methods under a given set of conditions.

CHAPTER 3: ANALYSIS OF DIVERGENCE IN FUNCTIONAL CONSTRAINT IN CASES OF NON-STATIONARY EVOLUTION

3.1 INTRODUCTION

Statistical methods designed to detect functional divergence between homologous genes provide valuable tools for analyzing the increasing amount of available genetic information. Most of these methods are derived from the concept that the substitution rate at a site is inversely related to functional constraint (Kimura 1983). Therefore, a measurable shift in substitution rate in a protein-coding gene can be associated with a divergence in the function of the protein product (*e.g.,* Gu 1999; Knudsen and Miyamoto 2001; Susko et al. 2002). A variety of strategies have been developed for detecting shifts in the rate of nonsynonymous substitution. Here I group them into two broad categories: 1) methods based on a scaled nonsynonymous rate (scaled by the inverse of the synonymous rate), and 2) methods based on the absolute rate of nonsynonymous substitution.

Scaling the nonsynonymous rate (d_N) by the synonymous rate (d_S) yields a useful index, ω (ω=d_N/d_S), of the strength and direction of natural selection pressure on a protein. Here the rate of synonymous change represents the rate of gene evolution prior to the impact of selection on the protein product of the gene. A value of $\omega > 1$ indicates nonsynonymous substitutions have been fixed more often than synonymous substitutions, which suggests the protein product has been subject to positive selection. The ratio ω=d_N/d_S can be formulated as an explicit parameter of a Markov model of codon evolution (Goldman and Yang 1994; Muse and Gaut 1994). See Section 1.2.1 of this thesis for a more comprehensive review of codon models. Among the more recent developments are models called branch-site models, which are designed to detect sites with a shift in ω among lineages in a phylogeny (Yang and Nielsen 2002; Bielawski and Yang 2004; Zhang et al. 2005). These models make it possible to detect divergence in functional constraint by estimating the strength and direction of shifts in natural selection (ω) among entire clades (*e.g.,* Bielawski and Yang 2004)

Alternatively, functional divergence can be detected by differences in the absolute rate of nonsynonymous substitution. Methods using this approach typically formulate this rate as an explicit parameter in a model of amino acid evolution, where it is referred to simply as the "evolutionary rate." In addition to shifts in evolutionary rate within a phylogeny (Type I divergence), some amino acid models also detect historical shifts in functional constraint, which are manifested as shifts in amino acid frequencies while substitution rate is the same (Type II divergence). For this reason, I use the term "functional divergence" to encompass both Type I and Type II divergence. There are several methods for detecting Type I shifts in functional constraint (*e.g.,* Gu 1999; Knudsen and Miyamoto 2001; Susko et al. 2002) and although each method takes a different approach, the overall process is essentially the same. First, two subtrees are specified *a priori* and rates of amino acid substitution at each site are estimated separately for each subtree. Amino acid models are then used to produce some measure of the difference between rates at each site. The methods differ in how rate shifts are quantified, in how they are statistically tested, and in how other model parameters are treated. See Appendix C for a list of methods for detecting functional divergence at the amino acid level. Regardless of the differences in methods, those sites with an inferred difference in evolutionary rate among subtrees are interpreted as having undergone a shift in functional constraint.

Occasionally, other aspects of the evolutionary process, such as codon usage bias or amino acid composition also shift within a phylogeny. However, current methods for functional divergence analysis do not account for this type of non-stationary evolution. Indeed, shifts in codon bias or amino acid composition are known causes of phylogenetic artifacts (Inagaki and Roger 2006; Lake 1994; Mooers and Holmes 2000). Other types of heterogeneity, such as a shift in compositional variation among sites, are known to cause false positives in the inference of positive selection using codon models (Bao et al. 2008). Because statistical methods for detecting functional divergence can provide valuable insight into the genetic basis of adaptation, they are popular and used in a wide variety of settings. However, we do not know their limits; *i.e.,* when violating the assumptions of an underlying model will lead to analytical artifacts and, ultimately, false biological conclusions. In this chapter, I use a series of simulations to explore the effects of non-stationary evolution on the inference of functional divergence using both codon and amino acid models. I specifically address the following three issues: 1) impacts of non-stationary codon bias alone on codon models, 2) combined effects of non-stationary codon bias and changing selection pressure on codon models, 3) power and reliability of amino acid models of functional divergence under non-stationary evolution. I then analyze a subset of the core genome of the cyanobacteria *Prochlorococcus*, which is known to have non-stationary evolution, as an empirical supplement to my simulation studies.

3.2 METHODS

3.2.1 Simulation Studies

Simulating gene sequence evolution

Using the INDELible simulation software (Fletcher and Yang 2009), I simulate a series of datasets under 18 different evolutionary schemes. Simulated datasets, each 200 codons in length, are based on the 10 taxa, asymmetric phylogeny shown in Figure 2.1a. All internal branch lengths in the phylogeny are set to 0.3 substitutions per codon site and all other branch lengths are adjusted to achieve rate constancy. When a shift in either codon bias or selective constraints is simulated, it occurs at the point shown in red in Figure 2.1a and is stable in all branches that evolve after that point. This effectively splits the tree into two types, "Type A" and "Type B," with different evolutionary models (Figure 2.1a).

Codon bias is modeled using the method of Aris-Brosou and Bielawski (2006) (Figure 2.4). This method employs a single parameter, "η," to specify codon frequencies for changing GC3 content (%G+C at the third codon position). The values of η range from $0 \leq \eta \leq 1$, where a value of $\eta = 0.5$ indicates a GC3 content of 50%; *i.e.*, all non-stop codons have equal frequencies. As η approaches zero, GC3 content increases. Using this system, I can easily specify separate codon biases for different parts of the phylogeny by setting different η values for Type A and Type B models.

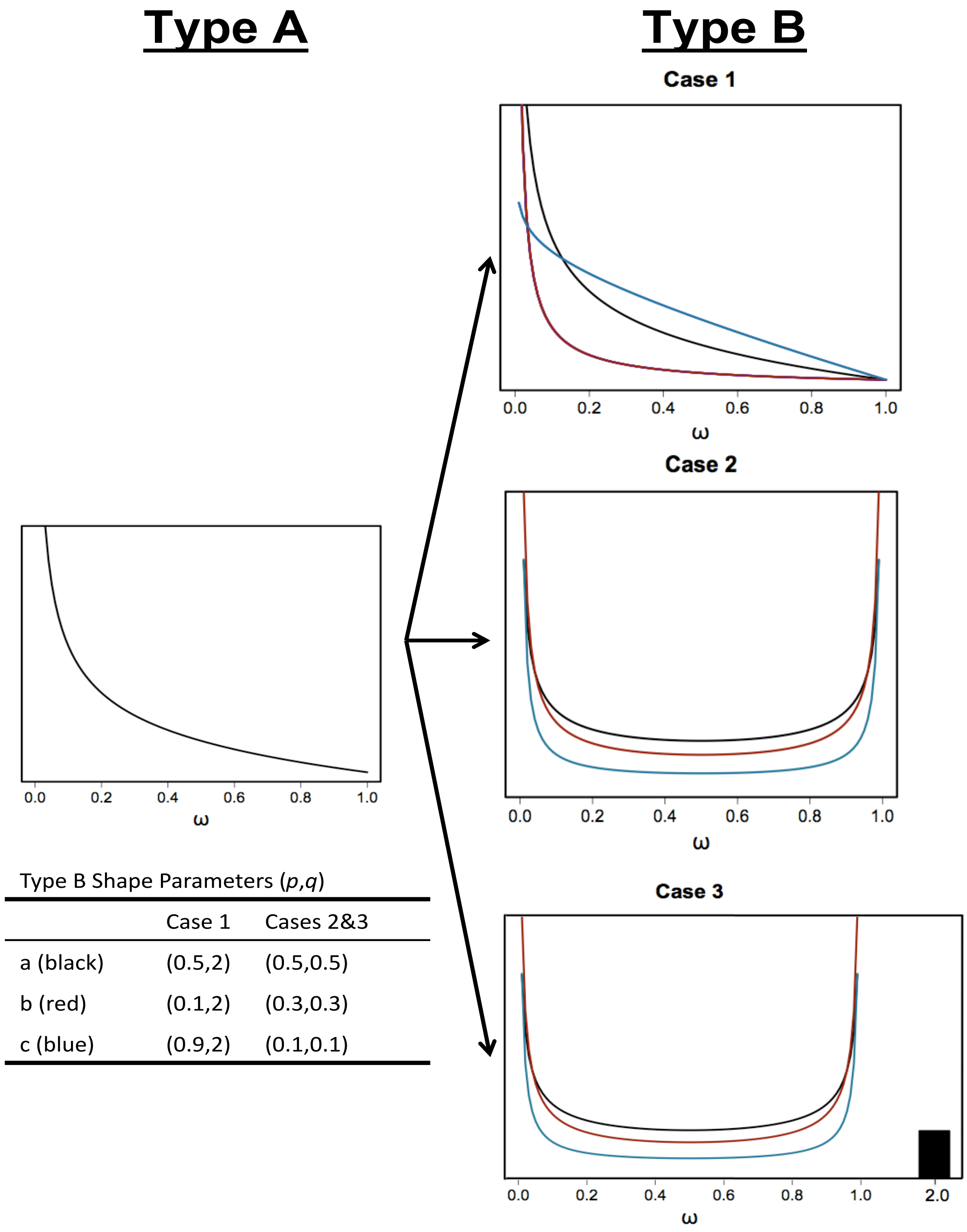
A shift in the distribution of selection pressures (*i.e.*, a shift in functional constraints) is simulated on the codon level by separately specifying a distribution for ω for Type A and Type B models. Omega distributions are modeled using a beta function. The beta function is convenient for this purpose because its range from zero to one is ideal for modeling an ω distribution with no positive selection ($0<\omega<1$). In addition, a variety of shapes can be specified using only two shape parameters (p,q). The beta distribution is split into 10 discrete categories of equal probability for the purpose of generating datasets without positive selection. When positive selection is simulated, the beta distribution is divided into nine categories and an extra site category with ω=2 is added (10% of sites).

The six most deeply branching taxa in the phylogeny are simulated under Type A evolution. For this evolutionary model, most codon sites are under strong purifying selection (ω<<1) and very few sites are evolving close to neutrally (p=0.5, q=2) (Figure 3.1). The codon bias parameter for Type A evolution is set to η_1=0.5 (GC3≈50%). Type A evolution remains constant throughout all simulation conditions.

Type B evolution, however, varies for each simulation condition. There are three basic cases for Type B evolution (Figure 3.1):

Case 1: Most sites are under strong purifying selection, similar to Type A evolution. This is modeled by an L-shaped beta function. This case represents an “easy null” for tests formulated to detect positive selection.

Case 2: A larger proportion of sites evolving close to neutrality. This is modeled with a U-shaped beta function. This is considered a “hard” null for tests formulated to detect positive selection because the increased proportion of sites near neutral (ω≈1) are more likely to give a false signal for positive selection than when the majority of sites are under strong purifying selection.



Type B Shape Parameters (*p*,*q*)

	Case 1	Cases 2&3
a (black)	(0.5,2)	(0.5,0.5)
b (red)	(0.1,2)	(0.3,0.3)
c (blue)	(0.9,2)	(0.1,0.1)

Figure 3.1. Beta functions used for ω distributions in simulation studies. For each case, the black curve represents sub-case "a," the red curve sub-case "b," and the blue curve sub-case "c." For methods designed to detect positive selection, Case 1 is an "easy null," where few sites are evolving close to neutrally. Case 2 is a "hard null" where a greater number of sites are evolving nearly neutrally, but no positive selection is present. Case 3 is the alternative model, where a proportion of sites in the foreground branches are simulated under positive selection.

Case 3 (Positive Selection): Same as Case 2, but an extra site category under positive selection (ω=2).

For each case above, I simulate 3 different ω distributions (designated sub-cases “a,” “b,” and “c”) with slightly different shape parameters (Figure 3.1). In Case 1a, Type B evolution is simulated under the same conditions as Type A, so selection pressure is homogeneous across the phylogeny. Each sub-case is simulated both with a shift in codon bias (η_2=0.1: GC3≈10%) and without a shift in codon bias (η_2=0.5) across the phylogeny. In total, this design yields 3 cases x 3 different beta distributions per case x 2 models of codon bias = 18 evolutionary schemes. For each scheme, 50 replicate datasets are simulated.

Analysis at the Codon Level

Using codeml in the PAML package (Yang 2007), I analyze each simulated replicate under sites models M1a and M3, as well as branch-sites models Model A, Model B, and the modified Model A with ω_2=1 (Zhang et al. 2005) (see Appendices A and B). For branch-sites models, branches under Type 2 evolution are specified as foreground branches (*i.e.*, a foreground clade). Results from codeml are used to perform three likelihood ratio tests, which can be interpreted as tests for a shift in the distribution of functional constraints:

Test 1: M1a vs. Model A (d.f.=2)

Test 2: modified Model A (ω_2=1) vs. Model A (d.f.=1)

Test 3: M3 vs. Model B (d.f.=2)

The test statistics from these likelihood ratio tests are compared to a χ^2 distribution to obtain a *p*-value. Tests 1 and 2 (based on Model A) are for sites with a shift to positive selection in foreground branches while Test 3 (based on Model B) is for any shift in selection pressure, since background branches are not constrained to ω<1 (see Appendix B) (Yang and Nielsen 2002). See section 1.2.1 for additional descriptions of the branch-site models and foreground branches as well as likelihood ratio tests.

Analysis at the Amino Acid Level

For analysis of functional divergence at the amino acid level, I apply two different methods: Bivar (Susko et al. 2002) and FunDi (Gaston, unpublished method). For this analysis each dataset simulated at the codon level is translated into amino acids. Both Bivar and FunDi are then applied to each replicate to measure the effects of non-stationary evolution on the output of these methods.

Bivar separately measures evolutionary rates for two subtrees, which are defined *a priori*. For a given site, *i*, the evolutionary rate (r_i) is measured using a conditional mode estimate; *i.e.*, the rate with the greatest conditional probability given the sequence data and phylogeny. For my simulations, the two subtrees are the two different types of evolution (Type A and Type B). I use the rates estimated from the simulated data to compare the rate differences between subtrees using three different measures:

$$arsum = \sum_i |r_{iA} - r_{iB}|$$

$$alrsum = \sum_i |\log(r_{iA} / r_{iB})|$$

$$abrsum = \sum_i |r_{iA} - r_{iB}| / (r_{iA} + r_{iB})$$

where *r* is the rate of evolution at site *i* in a given subtree (A or B). These measures provide a test statistic for measuring the difference in rate distribution between the two subtrees (Susko et al. 2002).

FunDi uses a likelihood-based mixture model to determine the likelihood of functional divergence at each site. The mixture model has two components: 1) a standard evolutionary model and 2) a model that treats the two subtrees separately. To determine if some fraction of sites in a gene has been subject to functional divergence, I develop a likelihood ratio test based on FunDi. A test statistic equal to two times the difference in log likelihoods (*2*ΔlnL*) between a standard model (based on 100% of sites belonging to component 1 of the mixture model) and the full mixture model is calculated. This test statistic is used to test the hypothesis that a gene is under functional divergence compared to the null that the gene has no sites subject to functional divergence.

For both methods, I use parametric bootstrapping to determine a *p*-value for each test statistic. Because this bootstrapping also requires simulation, for clarity purposes I use the term “original replicate” to refer to the original simulated data, which is treated like a real gene, and “bootstrap sample” to refer to those datasets simulated for the purpose of parametric bootstrapping. For each original replicate, the amino acid alignment is used to estimate a maximum likelihood phylogeny using RAxML (Stamatakis 2006). Amino acid frequencies and the proportion of invariable sites are then estimated from the alignment using the codeml program in the PAML package. In addition, branch lengths and an alpha shape parameter are estimated by maximum likelihood under a WAG model using codeml. These estimated parameters are then used to simulate 100 bootstrap samples with stationary evolution and no shifts in evolutionary rate. Both Bivar and FunDi are applied to each bootstrap sample in the same manner as for the original replicate. Distributions created from the bootstrap samples for each test statistic (*arsum*, *alrsum*, and *abrsum* for Bivar and $2*\Delta lnL$ for FunDi) can be used to calculate *p*-values for the original replicates. Datasets with a *p*-value less than 0.05 are considered to have significant evidence for functional divergence at the amino acid level.

3.2.2 *Prochlorococcus* Genomic Data

The cyanobacterium *Prochlorococcus marinus* provides a good example of non-stationary evolution. *Prochlorococcus* lineages have diverged into two phylogenetically distinct ecotypes, a high-light adapted ecotype (HL) and a low-light adapted ecotype (LL). A shift in codon bias is observed between the two ecotypes; the most recently diverged strain has a GC content of 30.8% while the most deeply branching strain has a GC content of 50.7% (Kettler et al. 2007). This GC bias is reflected in biased synonymous codon usage, with the most recently diverged lineage having an effective number of codons (ENC) value of 49.0 and the most deeply branching having ENC = 58.2. Hereafter, compositional biases are summarized using %GC. Because the more recently evolved strains occupy a different habitat, this organism is a prime candidate for functional divergence analysis, yet the non-stationary GC content is a substantial deviation from the underlying model of all methods. For this reason, I choose genes from the *Prochlorococcus* core genome as an example of real data analysis under non-stationary conditions.

Alignments of Genomic Data and Generation of Phylogenies

For the analysis of genomic data, I begin with amino acid alignments of 1812 genes previously clustered into homologous groups using complete genomes from both *Prochlorococcus* and *Synechococcus* (Zhaxybayeva et al. 2009). I align nucleotide sequences from genomic data downloaded from GenBank (Table 2.8), using the amino acid alignments of Zhaxybayeva and colleagues (2009) as templates. For each gene, a maximum likelihood phylogeny is generated using RAxML (Stamatakis 2006). Because RAxML is start-point dependent, 10 trees are separately inferred under the GTR model with a gamma distribution for among-site rate variation. The tree with the best likelihood is used for the gene phylogeny.

Determination of a Core Genome Subset and a "Genome Tree"

The core genome is defined here as those genes found to be present in all 12 *Prochlorococcus* genomes. I concatenate nucleotide sequences for all core genes and generate a "genome tree" using RAxML. The genome tree is the best of 10 maximum likelihood inferences (GTR with gamma distribution) on the concatenated nucleotide sequence. In addition, I use PAUP* (Swofford 2003) to generate a consensus tree from all phylogenies of *Prochlorococcus* core genes.

Gene trees with phylogenies that are incongruent with the organismal phylogeny may yield false conclusions in phylogeny-based inference, such as in the inference of the strength and direction of natural selection pressure (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006). Although individual genes may have incongruent phylogenies, studies suggest that concatenation of multiple genes will be more robust to gene-specific lateral gene transfer events and thus are more likely to represent the evolutionary history of the organism (*e.g.,* Brown JR et al. 2001; Snel et al. 2005). For this reason, the genome tree is used as a representation of the organismal phylogeny. I then select the subset of gene trees having the same topology as the genome tree for further analysis of functional divergence. However, because small differences in phylogeny, due to either poor resolution or very local recombination events, are not expected to impact the inference of functional divergence (Anisimova et al. 2003), I am able to identify additional genes suitable for investigating the divergence of HL and LL *Prochlorococcus*. These additional genes are determined by a three-part bipartition analysis. This analysis filters the core genome for those genes that meet the following three conditions: 1) *Prochlorococcus* is monophyletic, 2) the HL ecotype is monophyletic, and 3) there is a bipartition between the two HL clades. Some topologies for this set of genes differ from the genome tree, but only due to "local" differences; *i.e.*, the structure of the tree required for testing functional divergence between HL and LL lineages is preserved.

Analysis of Prochlorococcus Genomic Data

For all genes in the subset, analysis of functional divergence is conducted at both the codon and amino acid levels. At the codon level, I use the codeml program from the PAML package (Yang 2007) to fit each alignment to sites-model M1a, branch-sites Model A, and modified Model A (ω_2=1), under the gene tree topology. For branch-sites models, all HL branches are specified as foreground branches (*i.e.*, a foreground clade). I also conduct an analysis in which only the branch between ecotypes is specified as the foreground branch. This analysis searches for an episodic shift in selection pressures between the two ecotypes. Using the likelihood for the data under each model, I conduct two likelihood ratio tests for positive selection: M1a vs. Model A (Test 1) and modified Model A (ω_2=1) vs. Model A (Test 2). The test statistics from these likelihood ratio tests are compared with a χ^2 distribution to determine a *p*-value for each test.

At the amino acid level, both Bivar and FunDi are applied in the same manner as described for the simulation study. HL and LL strains are specified as the two subtrees to be analyzed separately. For each gene, 100 parametric bootstrap samples are generated using parameter values estimated from the gene in the same manner as in the simulation study. For Bivar, the three ratios, *arsum*, *alrsum*, and *abrsum* are used as test statistics and compared to the distribution from the bootstrap analysis to obtain a *p*-value. For FunDi, a likelihood ratio test is performed as described above. Again, the test statistic is compared with the distribution from the bootstrap samples to determine a *p*-value for functional divergence.

3.3 Results and Discussion

3.3.1 Non-Stationary Codon Bias Alone Does Not Affect Explicit Tests Of Positive Selection Using Codon Models

In some organisms, codon bias is not stationary over evolutionary time (*e.g.,* Urbach et al. 1998; Moran 2003). However none of the current models of codon substitution account for this type of evolution; in these models, codon frequencies are assumed to be at equilibrium. Furthermore, codon bias is measured empirically from the entire alignment. Because frequency parameters are essential for the calculation of substitution probabilities, application of the current models to non-stationary data could lead to serious errors in the inference of natural selection. As part of this study, I explore how shifts in %GC (codon bias) may impact the inference of selection pressure based on the ω parameter of branch-sites codon models.

Each branch-site LRT (Tests 1, 2, and 3 in the methods) is evaluated in each of 18 evolutionary scenarios. I begin by isolating the simplest scenario (Case 1a), as this is the only case where selection intensity is homogeneous across the tree, thereby focusing strictly on the effect of non-stationary codon bias (Table 3.1). Here, Tests 1 and 2 do not show excessive rates of false positives when codon bias is stationary. Somewhat unexpectedly, this is also the case when a shift in codon bias is present. Test 3 yields a slightly higher false positive rate (as compared with Tests 1 and 2) when codon bias is stationary (8%) and this rate increases substantially under non-stationary codon bias (18%). These results indicate that while branch-site models designed to detect positive selection (Tests 1 and 2) are not negatively impacted by a shift in codon bias alone, Test 3, which tests for any shift in selection pressure, may yield an increased number of false positives under non-stationary codon bias.

Table 3.1. Percent replicates (*n*=50) that yield false positives when no shift in functional constraint is simulated.

Codon Bias	Test 1	Test 2	Test 3
Stationary	4	0	8
Non-Stationary	4	0	18

The constraints placed on model parameters under Tests 1 and 2 appear to give those tests some robustness to the type of model misspecification covered in this simulation. Test 3, on the other hand, does not constrain any site categories of the null model or the alternative model. Because the ω parameters for all site categories in both the null (M3) and alternative (Model B) are unconstrained, those models are more “flexible” in terms of their ability to be fit to the data that is simulated. In this case, the gap between the simulating model and both of the analytical models (M3 and Model B) is not small, particularly with respect to a shift in codon bias. It seems that the shift in codon bias in these data is “absorbed” by the variability in ω among lineages under Model B, with this variability being mistaken as a signal for a shift in the distribution of selective constraints.

3.3.2 Combined Effects of Non-Stationary Codon Bias and Selection Pressure Can Negatively Impact Inferences of Selection Pressure Using Codon Models

In real gene sequences, a shift in codon bias is often accompanied by a shift in the rate of evolution (*e.g.,* Dufresne et al. 2005; Moran 2003). Branch-site codon models accommodate non-stationary selective constraint by allowing some sites to shift among categories of selection pressure in foreground branches. However, a shift in selective constraint may not involve positive selection, as a change in functional constraint alone is sufficient to yield a change in evolutionary rate. An unresolved question is whether such changes, when accompanied by non-stationarity in other aspects of substitution, might result in false positives for likelihood ratio tests for positive selection. Here, I examine the combined effect of non-stationary functional constraint and codon frequencies on the inference of selection pressure.

Table 3.2. Percent replicates with significant likelihood ratio tests ($p \leq 0.05$) under non-stationary evolution.

	Test 1		Test 2		Test 3	
	S	N	S	N	S	N
Null (Case 1a)	4	4	0	0	8	18
Case 1 (b, c)	24	52	1	0	73	87
Case 2 (a,b,c)	94	95	0	9	85	98

Note: The complete results can be found in Appendix F. For this table, values have been averaged over sub-cases. Case 1a is presented separately because it is the null, while "Case 1" is the average of Cases 1b and 1c. For each test, conditions with stationary (S) and non-stationary (N) codon bias are shown.

Five simulation scenarios are characterized by either a small shift in the distribution of selection pressures (Cases 1b and 1c) or a large shift in selection pressures (Cases 2a, 2b, and 2c) in the foreground clade. In no case does the shift involve positive selection, hence they represent null scenarios with respect to explicit tests for positive selection (Tests 1 and 2). These scenarios are relevant to the question of codon bias because they are generated with either the presence or absence of a shift in codon frequencies. Note that individual results for sub-cases a, b, and c are shown in Appendix F, whereas results provided here (Table 3.2) are averaged over sub-cases. Results for Case 1 represent an average over scenarios b and c while results for Case 2 represent an average over scenarios a, b, and c.

For Test 1, even a small shift in the distribution of selection pressure (Case 1b and 1c) causes a large increase in the false positive rate for positive selection (Table 3.2). It is interesting that even when codon frequencies do not shift and there is only a slight shift in selection pressure (Case 1b and 1c), the false positive rate is high (24%). The addition of non-stationary codon bias to this condition causes false positive rates to nearly double (52%). There is also an increase in the false positive rate in Case 2, where a large shift in selective pressures occurs (*i.e.,* when there is a larger proportion of sites evolving near neutrality, with $\omega \approx 1$). However, here the overall effect of non-stationary codon bias is small for Test 1 because the false positive rate is unacceptably high (94%) even when codon frequencies are stationary. Although here I focus on false positives based on LRT results and not parameter estimates, it is interesting to note that in every case, average estimates for ω_2 are greater when codon bias is non-stationary (Table 3.3), further suggesting that the shift in codon bias is being absorbed through variation in ω. My findings are consistent with a previous study, simulated under very different conditions, which found high false positive rates for Test 1 (40-70%) depending on topology, branch lengths, and selection pattern (Zhang 2004). A subsequent study found that this LRT is better suited as a test for relaxed constraint than positive selection (Zhang et al. 2005). Indeed, my study confirms this is the case as even when there is a shift in codon bias, Test 1 has a consistently low false positive rate under the null (Case 1a in Table 3.2).

Table 3.3. Mean and standard error values for ω_2 in foreground branches estimated under Model A for stationary and non-stationary codon bias and different shifts in selection pressure.

	Stationary		Non-Stationary	
	mean	SE	mean	SE
Null (Case 1a)	1.12	0.07	1.29	0.13
Case 1 (b, c)	1.07	0.05	1.10	0.06
Case 2 (a, b, c)	1.00	0.00	1.19	0.03
Case 3 (a, b, c)	1.08	0.01	1.44	0.05

Note: The true value of ω_2 is not greater than 1 in any of these cases. The values of ω_2 in the table are the average of separate maximum likelihood estimates taken over all the replicates of a given simulation case.

Test 3 behaves similarly to Test 1, with non-stationary frequencies causing an increase in rejection rate of the null hypothesis (Table 3.2). The effect of non-stationary codon bias is smaller in Test 3, at least in part because the false positive rate is higher than for Test 1 when there is no shift in codon frequencies. Model B is the alternative model for Test 3 and, although it allows for positive selection (as does the null model M3), site categories are not constrained to ω>1. Thus, Test 3 is not an explicit test for positive selection in foreground branches (Yang and Nielsen 2002). Based on the formulation of the null and alternative models, Test 3 should be viewed as a test for a shift in selection pressure, such as a relaxation of selective constraints. Hence, rejection of the null under Test 3 in Case 1 and Case 2 should not be interpreted as evidence of positive selection and consequently they do not represent false positives for positive selection (unlike Tests 1 and 2).

As tests for relaxed constraint, both Test 1 and Test 3 perform reasonably well. When codon bias is non-stationary, Test 1 may be preferred because it yields fewer false positives under the null case (Case 1a) and does not show an increase in false positive rate under non-stationary codon bias alone. However, Test 3 has higher power when the shift in selection pressure is small (Case 1b and 1c). Therefore, under stationary codon bias, Test 3 may provide a better understanding of the shifts in selection pressure experienced by a set of genes. Both tests appear to have increased power under non-stationary codon bias. Because this increase results from a violation of model assumptions, it can be considered a systematic error in the direction of the alternative model. While increased power is normally desirable, one can imagine a case where a shift in selection pressure is marginal at best. If this hypothetical dataset also experiences a shift in codon bias, the statistical significance of the shift in selection pressure will be falsely inflated. Furthermore, given that the relationship between the assumptions of the model and the true generating process will be unknown for real data, it seems imprudent to rely on systematic errors in the direction of the alternative hypothesis as means of achieving good power.

Test 2 is known to be a more conservative test for positive selection (Zhang et al. 2005). When a small shift in selection pressure is present, the false positive rate is very low regardless of whether codon bias is stationary or non-stationary (Table 3.2). For a large shift in selection pressure combined with a shift in codon frequencies, the false positive rate increases to 9%, which is only marginally larger than the level of the test. However, this increase is entirely due to one sub-case (see Case 2c in Appendix F), which is simulated under the most extreme shift in selection pressure. Taking Case 2c on its own, the false positive rate is estimated to be 24% when codon bias is non-stationary (see Appendix F). These results indicate that even though Test 2 is a conservative test for positive selection, it may yield false positives due to model misspecification under strong non-stationary evolution.

Case 3 (a, b, and c) is relevant to assessing the power of each test to detect positive selection when it is truly present in the data (Table 3.4). Tests 1 and 3 have high power in Case 3. This further validates the potential of these LRTs as tests for a shift in selective pressure among lineages. Unfortunately Test 2, the only defensible test for a shift involving positive selection in the foreground clade, has low power. The best performance of Test 2 is under non-stationary conditions, but simulations under Cases 1 and 2 suggest this reflects systematic error in the direction of the alternative hypothesis. Thus it seems that the conservative quality of Test 2 in the face of a shift in codon bias is due to generally low power of the test. This finding of low power for Test 2 is consistent with previous analysis of this LRT under different conditions (Zhang et al. 2005).

Table 3.4. Percent replicates with significant likelihood ratio tests under positive selection and with stationary and non-stationary codon bias.

	Test 1		Test 2		Test 3	
	S	N	S	N	S	N
Null (Case 1a)	4	4	0	0	8	18
Case 3 (a,b,c)	93	94	2	23	96	100

Note: The complete results can be found in Appendix F. For this table, values have been averaged for Case 3. "S" indicates that codon bias is stationary across the phylogeny while "N" indicates non-stationary codon bias.

3.3.3 Testing for Functional Divergence Under Non-Stationary Evolution at the Amino Acid Level

Methods designed to detect functional divergence on the amino acid level do not take into account synonymous substitution. For this reason, they may be less sensitive than codon models to a shift in GC3 content. Inagaki and Roger (2006) show heterogeneity in codon usage may lead to phylogenetic artifacts under codon models, and that amino acid models are more reliable under such conditions. However, because amino acid models do not utilize all of the sequence information, they may have lower power to detect some substitution processes. Here, I evaluate the ability of amino acid models to detect a shift in functional constraint as well as test their sensitivity to shifts in codon bias.

Both programs for analyzing functional divergence at the amino acid level, Bivar and FunDi, are applied to each of 18 evolutionary scenarios described in the methods. Bivar employs three summary statistics to quantify the difference in evolutionary rates among sites between two subtrees. Two of the three measures (*alrsum* and *abrsum*) yield reliable results in my simulations (see Appendix G for the full set of results). The first ratio, *arsum*, yields excessively high levels of false positives (36%) even when evolution is stationary (Case 1a). Since results from the other two ratios, *alrsum* and *abrsum*, are similar (Appendix G), I will review the *abrsum* values for discussion.

As methods based on amino acid models are designed to detect any shift in evolutionary rate, the only null scenario is Case 1a, where evolutionary rate is homogeneous throughout the phylogeny. With respect to data simulated under the null hypothesis, Bivar yields substantially fewer false positives (2%) as compared to FunDi (14%). All other cases (1b, 1c, 2a-c, 3a-c) are used to investigate their power to detect functional divergence. In Table 3.5, I show only the results from conditions with stationary codon bias in order to investigate power for detecting functional divergence without any possible error caused by a shift in composition. Because there is no apparent pattern within each case (see Appendix G), I have calculated averages over sub-cases. Both methods have moderate power to detect a shift in evolutionary rate (Bivar: 20-54%; FunDi: 30-41%). Bivar has somewhat higher power than FunDi in Cases 1 and 2, but not in Case 3. Recall that Case 3 differs from Cases 1 and 2 by the addition of a fraction of sites subject to positive selection. These results therefore suggest that FunDi may be better suited to detecting functional divergence in real datasets subject to positive selection. Because amino acid models do not take into account synonymous substitutions, they may be less sensitive to some substitutions that result in a shift in functional constraint. Indeed, these amino acid level methods seem to have lower power than codon level models designed to test for a shift in constraint. For instance, codon-based Test 3 detects a shift in constraint in a minimum of 73% of replicates, which is substantially higher than either amino acid method in any scenario.

Table 3.5. Percent of replicates with signals for functional divergence ($p \leq 0.05$) for amino acid methods under stationary codon frequencies.

	Bivar (*abrsum*)	FunDi
Null (Case 1a)	2	14
Case 1 (b,c)	58	29
Case 2 (a,b,c)	55	43
Case 3 (a,b,c)	31	35

Note: Complete results can be found in Appendix G. For this table, values are averaged over sub-cases. Case 1a is presented separately because it is the null, while "Case 1" is the average of Cases 1b and 1c.

To determine whether non-stationary codon bias impacts detection of functional divergence at the amino acid level, I apply both amino acid level methods (Bivar and FunDi) to simulations with non-stationary codon bias. In these simulations, false positive rates increase substantially when a shift in codon bias is present. The false positive rate for Bivar (measured under Case 1a) increases from 2% to 16 % when a shift in codon bias is introduced. The false positive rate for FunDi increases from 14% to 36% when there is a shift in codon bias. Outside of the null, functional divergence is detected in more replicates when a shift in codon bias is present (Table 3.6). These results suggest that tests for functional divergence that are formulated at the amino acid level also are impacted by non-stationary codon bias usage (in this setting the shift in codon bias is manifested as non-stationary amino acid frequencies).

Table 3.6 Percent replicates in which functional divergence is detected by using amino acid methods under stationary and non-stationary codon bias.

	Bivar (*abrsum*)	FunDi
Stationary	47	36
Non-Stationary	58	46

Note: Complete results can be found in Appendix G. Here, values are averaged over all conditions with a shift in evolutionary rate.

There are other aspects of the substitution process, in addition to codon bias, that are not included in the analytical models and thus may also have impacted the involved statistical tests. Simulation at the codon level in this study is based on generating sequences with a specified distribution of nonsynonymous and synonymous substitutions, without regard for the specific amino acid encoded. Tests for functional divergence at the amino acid level use models that permit one-step changes between amino acid states requiring more than one synonymous change. Furthermore, both Bivar and FunDi take into account differential amino acid exchangeabilities. It is therefore possible that the observed impact of non-stationary codon bias on amino acid methods for functional divergence detection is related to this additional gap between the generating and analytical models. To investigate this, I conduct an additional simulation study, generated at the amino acid level, in order to determine whether a shift in composition alone leads to false positives for amino acid models (See Appendix H for detailed methods and results). For this study, I simulate a shift in amino acid frequencies equivalent to the shift in codon frequencies used for the main simulation. Under these conditions, Bivar yields only 2% false positives and FunDi yields 6%. These results indicate that a shift in amino acid frequencies alone does not impact amino acid level methods for functional divergence analysis. Taken together with the other simulations, these results highlight the importance of the gap between the generating and analytical models; the larger the gap, the more serious the effect on any involved statistical tests. The problem remains that for any real dataset, we will not know the true generating model. We do know, however, that it is most certainly more complex than the analytical models in hand.

3.3.4 *Prochlorococcus* Genes Appear Subject to Functional Divergence Between HL and LL Ecotypes

Here I analyze a subset of the core genome of the cyanobacterium *Prochlorococcus*, which is well known to contain genes subject to functional divergence (among HL and LL ecotypes) as well as strong non-stationary evolution with respect to codon bias. *Prochlorococcus* genes from previously clustered genomic data are processed in order to obtain a subset of the core genome with the same basic evolutionary history as the organism. Out of a total of 1812 genes, 1005 genes are present in all *Prochlorococcus* strains and thus make up the core genome. This estimate is within the range of sizes previously reported for this genus. It is smaller than the Kettler and colleagues (2007) estimate of 1273 core genes, likely due to the use of *Synechococcus* in the initial clustering of my data and the difference in clustering methods (Zhaxbayeva et al. 2009). Early studies estimated the core genome to be even larger (*e.g.,* Dufresne et al. 2003), but do not use information from all 12 genomes. On the other hand, my estimate is much larger than the core proteome identified by Paul and colleagues (2010). However, the E-value cutoff used in their study was extremely conservative (1 x 10^{-20}).

Each gene in the core genome is compared with the genome tree in order to identify genes with the same basic evolutionary history as the organism. Figure 3.3

shows the genome tree generated from the concatenation of all core genes as well as the consensus tree from all core gene phylogenies. While both trees clearly separate the strains in the HL and LL ecotypes, other parts of the tree are poorly resolved. For example, the consensus tree contains a polytomy for MIT9211 and CCMP1375 within the LL grade. In addition, in the consensus tree, some branches in the HL clade have low support, which may be caused by small branch lengths in this part of the phylogeny. Because parts of the consensus tree have poor resolution, most genes do not have topologies that are congruent with the genome tree. In fact, only 97 out of 1005 core genes have the exact same topology as the genome tree. This may be due, in part, to short branch lengths. In addition, recombination events may yield gene phylogenies that differ from organismal phylogenies. While small phylogenetic differences, whether due to poor resolution or local recombination, are not a problem for phylogenetic-based inference, a large difference may negatively impact codon models (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006) and the effects on amino acid models are not known.

In order to obtain more than 97 genes for functional divergence analysis while not allowing genes trees to be drastically different than the genome tree, I conduct a three-part bipartition analysis. This analysis filters for genes with phylogenies that have three bipartitions that match those in the genome tree: 1) between *Prochlorococcus* and *Synechococcus*, 2) between HL and LL ecotypes, and 3) between the two HL clades. Based on this analysis, I find 585 genes with the same fundamental evolutionary pattern as the genome tree. This subset of genes from the *Prochlorococcus* core genome is used in the analysis of functional divergence.

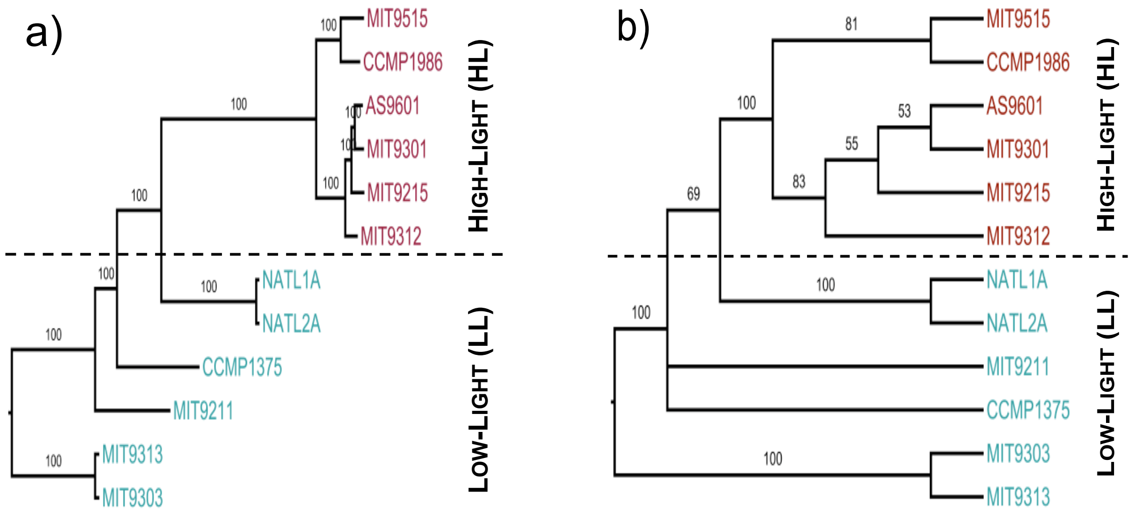


Figure 3.3. a) Genome tree generated from concatenation of all 1005 *Prochlorococcus* core genes. b) Consensus tree of topologies inferred from separate analysis of 1005 core genes.

The subset of 585 core genes are analyzed using the two likelihood ratio tests designed to detect positive selection in part of the tree: Test 1 and Test 2. Using Test 1, a signal for positive selection is observed in 76% of genes when the entire HL clade is specified as foreground and 90% of genes when just the branch between ecotypes is specified (Table 3.7). However, results from previous studies (Zhang et al. 2005) and simulations presented in this chapter suggest that Test 1 is more appropriate as a test for altered constraint than positive selection. In addition, results from my simulations show that non-stationary evolution, such as that present in *Prochlorococcus*, is associated with an increase in false positives for this LRT. For these reasons, this is likely an overestimate of the number of genes under positive selection. According to the more conservative Test 2, no genes in the subset contain a signal for positive selection when the whole HL clade is specified as foreground and only 24% when the branch between ecotypes is specified. The increased number of genes under positive selection when the foreground branch is the branch between ecotypes may indicate that an episodic shift in selection pressure is more likely to have occurred in *Prochlorococcus* than an actual shift in the entire HL clade. However, because even this conservative test yields increased false positive rates under non-stationary evolution, it is impossible to infer from these results a reliable estimate of the proportion of *Prochlorococcus* core genes that have experienced positive selection. Clearly there is signal in these data that is worth further investigation. New models that explicitly model phylogenetic shifts in codon bias will be required for further analysis under the codon model framework.

Table 3.7. Percent of genes (n=585) with significant results ($p \leq 0.05$) for codon level analysis of shifts in selection pressure in a subset of the *Prochlorococcus* genome.

Foreground Branches	Test 1	Test 2
All HL	76	0
Between Ecotypes	90	24

Analysis at the amino acid level also suggests functional divergence in a substantial proportion of genes in the *Prochlorococcus* core genome. Under Bivar, the *abrsum* measure detects functional divergence in 268 (46%) genes. The likelihood ratio test used for FunDi yields positive results in 108 (18%) genes of the subset of the core genome. While results from Bivar and FunDi differ, the results of my simulations suggest a preference for Bivar. Bivar has the desirable qualities of moderate power and only small elevation of false positive rates under non-stationary codon bias. To offset potentially elevated false positive rates, I recomputed the results for Bivar under a more stringent level of the test by setting $\alpha = 0.01$. Under this condition, 172 (29%) genes yield a signal for functional divergence. This value is similar to the codon-level Test 2 results with the branch between ecotypes specified as foreground. However, upon closer analysis, I find that only 50 genes test positive for a shift in functional constraint for both Bivar and codon-level Test 2. While my findings are preliminary and should be followed up with tests that employ models that can explicitly accommodate non-stationary codon bias, they suggest core genes might play an important role in ecotype divergence. Although many previous studies focus on the role of the flexible genome in the divergence of *Prochlorococcus* ecotypes (*e.g.,* Kettler et al. 2007; Rocap et al. 2003), more recent studies suggest that the core genome may be important to niche differentiation in other bacteria (Sarkar and Guttman 2004; Dunn et al. 2009). It seems this could be the case for HL and LL divergence in *Prochlorococcus* as well.

3.4 CONCLUSIONS

Models of sequence evolution can be useful tools for the analysis of genetic data. However, one must carefully consider the assumptions made by these models and determine whether they are acceptable for a given dataset. My results show that non-stationary evolution may lead to false conclusions in the inference of selection pressure at the codon level. Even the most conservative of codon models yields significant levels of false positives where sequence evolution is strongly non-stationary. It is therefore recommended that current models of codon evolution should not be used when there is a large shift in codon bias within a dataset.

One possible solution to this problem is to implement models of codon evolution that take into account non-stationary composition. Currently, non-stationary models exist at the nucleotide level (*e.g.,* Yang and Roberts 1995; Galtier and Gouy 1998), but have yet to be developed at the codon level. A major concern in implementing such models is that the increase in parameters to account for multiple sets of codon frequencies is too large. However, this can be avoided by using an approximation based on estimates of nucleotide frequencies at each of the three codon positions (F3x4) (Yang 2007). This reduces the burden of the estimation from 61 to just 9 frequency parameters per set. One can imagine a model where these frequencies could then be calculated empirically for two parts of the phylogeny, specified *a priori* (*i.e.,* for two sub-trees each containing some contemporary sequences) or estimated using a maximum likelihood framework. Maximum likelihood estimation would permit greater flexibility in modeling shifts in codon bias; for some examples of analogous models at the nucleotide level see models N1 and N2 of Yang and Roberts (1995). However, maximum likelihood estimation will incur substantial increases in computational burden. Implementation of such a model would allow for detection of functional divergence that takes advantage of codon-level information, but datasets would no longer be subject to the negative effects of model misspecification due non-stationary codon bias.

While amino acid models may be a viable alternative when a shift in composition is present, synonymous information is not utilized, which may decrease the power of those models in some settings. In addition, my simulation studies reveal negative effects on the involved statistical tests when sequences were simulated at the codon level and analyzed with amino acid models. My results therefore highlight the need for the integration of information from both codon and amino acid levels. Ideally both types of information could be taken into account. Some work has been done in this area, using empirical codon models to simultaneously account for synonymous substitution and varying amino acid exchangeabilities (*e.g.,* Schneider et al. 2005; Doron-Faigenboim and Pupko 2007; Kosiol et al. 2007). Such models could be extended to permit shifts in the distribution of selection pressures among branches, thereby permitting the formulation of explicit LRTs. One difficulty, however, is that the interpretation of the ω parameter is no longer straightforward when empirical exchangeabilities are incorporated into codon models (Doron-Faigenboim and Pupko 2006; Kosiol et al. 2007).

While models of functional divergence can provide valuable information concerning adaptively significant genetic variation, underlying model assumptions cannot be ignored. Based on my simulations, I provide the following suggestions for using these models:

- Test 1 should not be used as a test for positive selection, as it is sensitive to both shifts in codon bias and shifts in the distribution of selection pressures that do not involve positive selection.
- Test 2 is fairly robust, but has low power.
- Test 3 is a powerful test for functional divergence, but should only be used when codon frequencies are stationary. Furthermore, at α=0.05 the false positive rate will be slightly above the level of the test. I therefore recommend using α=0.01 to control for this.
- Pay particular attention to evolutionary processes that may violate model assumptions (*e.g.,* non-stationary evolution, recombination, etc.). All the statistical tests, whether based on codon or amino acid models, were negatively affected by errors arising from model misspecification.

- Simulate under conditions relevant to the data at hand in order to thoroughly understand the limits of candidate methods.
- Tests based on amino acid and codon models should be complimentary, but the specific tests should be chosen judiciously.

CHAPTER 4: THE *PROCHLOROCOCCUS cpeB* GENE AS AN EXAMPLE OF ANALYSIS UNDER NON-STATIONARY EVOLUTION

4.1 INTRODUCTION

Niche differentiation in bacteria is often accompanied by divergence at the gene level. Strains of the cyanobacteria species *Prochlorococcus marinus* have diverged to form two distinct ecotypes, a high-light adapted ecotype (HL) and a low-light adapted ecotype (LL), which have different chlorophyll b/a ratios (Moore et al. 1995; Partensky et al. 1997; West and Scanlan 1999; Rocap et al. 2002). Physiological differences such as optimal light intensity, optimal temperature, metal tolerance, and nutrient utilization allow members of the two ecotypes to thrive at different depths of the water column (Mann et al. 2002; Tolonen et al. 2006; West et al. 2001; Moore et al. 2002; Ahlgren et al. 2006).

Along with differences in physiological characteristics, *Prochlorococcus* ecotypes are separated by several genomic differences, including genome size and base composition. Lineages in the HL clade have much smaller genomes than members of the LL ecotype (Hess et al. 2001; Dufresne et al. 2005). The phenomenon of genome reduction is well documented in endosymbionts, but this is the first case observed in a free-living organism. In endosymbionts, genome reduction is often accompanied by a shift in base composition. *Prochlorococcus* also possesses this characteristic; genomic GC contents vary from 30.8% in a HL strain to 50.7% in a LL strain (Kettler et al. 2007; Hess et al. 2001). For more information on the ecology and evolution of *Prochlorococcus*, see Section 1.1.

While the divergence of *Prochlorococcus* into two separate ecotypes makes it a prime candidate for studies that examine shifts in selection pressure, its genomic characteristics violate the assumptions of current methods available for such analysis. Models of evolution at the codon level do not account for codon bias that is non-stationary among lineages. In Chapter 3, I show that branch-sites codon models can yield false positives for positive selection under non-stationary evolution, leading to false biological conclusions. Despite the fact that the non-stationary nature of the *Prochlorococcus* genome violates model assumptions, a previous study used codon models to analyze the *cpeB* gene, concluding that there is evidence for positive selection (Zhao and Qin 2007).

The *cpeB* gene encodes the *β* subunit of phycoerythrin (PE), which is a phycobiliprotein associated with the light-harvesting structures, phycobilisomes. However, unlike its closest relative, *Synechococcus*, strains of *Prochlorococcus* do not use phycobilisomes for light harvesting, instead they harvest light via a chlorophyll antenna. Therefore, the function of *cpeB* in *Prochlorococcus* is unknown. Although the function of the *cpeB* gene is not known, multiple studies have analyzed patterns of selective pressure in this gene, examining both divergence from *Synechococcus* (Ting et al. 2001) and divergence of HL and LL ecotypes (Zhao and Qin 2007). The interest in *cpeB* stems from the notion that it has likely been co-opted to serve a novel function within *Prochlorococcus*. In this Chapter, I reanalyze the *cpeB* gene, using both information acquired in Chapter 3 and additional simulations specific to this gene, in order to draw conclusions about patterns of functional constraint.

4.2 METHODS

4.2.1 Alignment and Analysis of *cpeB* Gene Sequences

Although Zhao and Qin (2007) use environmental sequences in their analysis, I choose to use sequences from genomic data, based on the assumption that these sequences are more reliable. Nucleotide sequences for *cpeB* genes are extracted from all 12 *Prochlorococcus* genomes (See Table 2.8) and downloaded from GenBank. Additional *cpeB* sequences from strains PAC1, PAC2, and TAK9803 (Accession Numbers: AJ272069.1, AJ237612.1, and AJ304838.1) are also downloaded from GenBank. All nucleotide sequences are translated into amino acid sequences, which are aligned using T-Coffee (Notredame et al. 2000). This amino acid alignment is manually edited in Jal-View (Waterhouse et al. 2009) and is then used as a template to align nucleotide sequences. A maximum likelihood phylogeny is generated in PAUP* (Swofford 2003) using an HKY85 model with a gamma distribution for among-site rate variation. In addition, a neighbor-joining phylogeny is also generated using paralinear (LogDet) distances, which are thought to be less sensitive to compositional shifts (Lake 1994; Lockhart et al. 1994).

Branch length estimates are measured using a non-stationary nucleotide model (Yang and Roberts 1995) implemented in baseml of the PAML package (Yang 2007). This model allows frequency parameters to be estimated separately for different branches of the tree. The user may specify the number of frequency parameter sets as well as the nodes to be used in frequency estimation for each set. In this case, I specify two sets of frequency parameters, one for the HL clade, and one for the LL grade and the root node. Using the maximum likelihood topology for the *cpeB* gene, new branch lengths are estimated using this non-stationary model combined with a stationary correction for transition/transversion bias.

Recombination is known to lead to false positives in phylogeny-based inference (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006) and testing for recombination is therefore an important step in any analysis of functional constraint where recombination is plausible. If recombination is detected, each fragment in the gene should be analyzed separately (Scheffler et al. 2006). Because the phylogeny of *Prochlorococcus* is largely asymmetric, I use the two methods not adversely affected by asymmetric topology or non-stationary evolution (based on simulations in Chapter 2): GARD-MBP, and GENECONV.

Using the codeml program from the PAML package, several codon models are fit to the *cpeB* data. These models include sites models M1a and M3 as well as branch-sites Model A and Model B. In addition, HL and LL sequences are separately fit to model M3 in order to estimate parameters for use in a simulation study. For the whole dataset, three likelihood ratio tests are performed: 1) M1a vs. Model A, 2) modified Model A (ω_2=1) vs. Model A, and 3) M3 vs. Model B. For branch-sites models, the HL clade is specified as foreground. Based on previous studies (Yang and Nielsen 2002; Zhang et al. 2005) and the results of simulations in Chapter 3, Tests 1 and 3 are best interpreted as tests for a shift in functional constraint. Test 2 is designed as a test for positive selection in foreground branches (Zhang et al. 2005). For more information on codon models and likelihood ratio tests, see Section 1.2.1 and Chapter 3.

Amino acid models are less sensitive to shifts in codon bias than codon models (see Chapter 3). For this reason, I conduct an analysis of functional divergence at the amino acid level using the program Bivar (Susko et al. 2002). Of the two methods I evaluate in Chapter 3, I find Bivar is more powerful and has lower type-I error rates in a setting comparable to this one. For this analysis, I use the HL strains and LL strains to designate the required subtrees. Based on my simulations in Chapter 3, I choose to employ only the *abrsum* test statistic. This statistic is used as a measure of the difference in rate distributions between the two subtrees:

$$abrsum = \sum_i |r_{i1} - r_{i2}| / (r_{i1} + r_{i2})$$

where *r* is the conditional mode rate estimate at site *i* for a given subtree (1 or 2). After the test statistic is calculated, parametric bootstrapping, with 100 bootstrap samples, is used to determine a *p*-value for functional divergence. For more information on Bivar, as well as the parametric bootstrapping methodology, see Section 3.2.1.

4.2.2 Simulation 1: Estimation of Selection Pressure

In order to determine whether the amount of heterogeneity present in the *cpeB* gene can cause false positives in the inference of selection pressure, a set of simulations is performed based on parameter values estimated from the *cpeB* gene sequences. Using the INDELible program for non-stationary simulation, I simulate alignments of 170 codons in length (the length of the *cpeB* gene), using the maximum likelihood phylogeny estimated for the *cpeB* gene with branch lengths estimated using a non-stationary nucleotide model and rescaled to represent the mean number of substitutions per codon site. Distributions for ω are estimated separately for HL and LL ecotypes under model M3 and codon frequencies are measured empirically for each ecotype. Using these values, I specify separate models for the simulated HL and LL strains. Two classes of parameters, ω distribution parameters and codon bias frequency parameters, are varied in this study (Figure 4.1). For each of these classes, simulations are conducted in which there is either no change (the entire tree is simulated under the HL parameter values) or changes occur according to the estimated difference between HL and LL ecotypes (Figure 4.1). As I want to investigate the null scenario, there is an exception to the above design: any category of sites estimated to have an ω value greater than 1 is simulated under ω=0.9 to avoid a true signal for positive selection. For each combination of parameter values, 100 replicates are simulated.

		ω values	
		Homogeneous	**Separate HL/LL**
Codon bias	**Homogeneous**	Same GC Same ω	Same GC Δ in ω
	Separate HL/LL	Δ in GC Same ω	Δ in GC Δ in ω

Figure 4.1 Design of simulation studies. The diagram indicates the study is comprised of four cases. In cases where a single parameter value is homogeneous over the entire tree, the empirical estimate of that parameter value from the HL clade is used.

All simulated sequences are analyzed using codeml under sites models M1a as well as branch-sites Model A and the modified Model A where ω_2 is fixed at 1. For branch-sites models, all branches in the simulated HL clade are specified as foreground.

Two likelihood ratio tests are then performed: M1 vs. Model A (Test 1) and Model A (ω_2=1) vs. Model A (Test 2).

4.2.3 Simulation 2: Estimation of Branch Length

Non-stationary codon bias may cause errors not only for likelihood ratio test results, but also for estimation of model parameters, such as branch lengths. The purpose of Simulation 2 is to investigate the effect of non-stationary evolution on the estimation of branch length under the selected codon models. Here I employ the same four basic conditions as in Simulation 1 (Figure 4.1). However, for each condition, the branch between ecotypes is adjusted to 2, 3, 5, or 10 substitutions per codon site. For each of the resulting 16 simulation conditions, 100 replicate datasets are simulated. For each dataset, branch lengths are estimated under a sites model (M1a) as well as a branch-sites model (Model A).

4.3 RESULTS AND DISCUSSION

4.3.1 Analysis of Real *cpeB* Gene Sequences

The presence of recombination events within a gene is known to be associated with false positives in the inference of positive selection (Anisimova et al. 2003; Shriner et al. 2003; Scheffler et al. 2006). This occurs because the differences in substitution parameters in a segment of recombinant DNA may mimic genetic variability generated by a difference in selection pressure among sites. Tests for recombination using GENECONV and GARD-MBP do not yield evidence for recombination within the *cpeB* gene. Lacking evidence of recombination, I proceed with a joint analysis of all alignment positions using the standard tests formulated at the codon and amino acid levels.

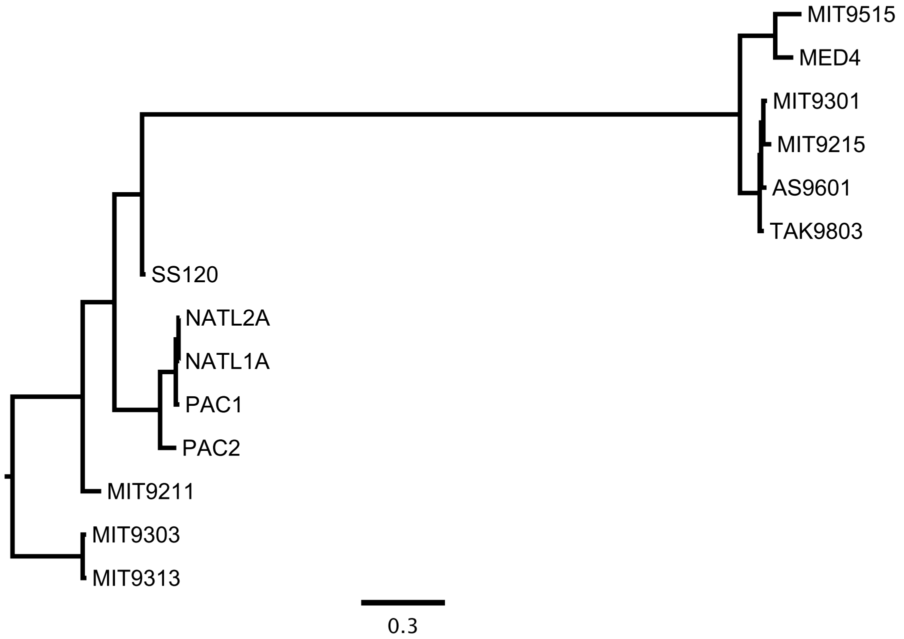


Figure 4.2. Maximum likelihood phylogeny of *Prochlorococcus cpeB* gene. Branch lengths are measured under a non-stationary nucleotide model and multiplied by three to approximate substitutions per codon site. Note the very long branch length between HL and LL lineages for this gene.

Figure 4.2 shows the maximum likelihood phylogeny for the *cpeB* gene with branch lengths estimated under a non-stationary nucleotide model. Notice that the branch between ecotypes is substantially longer than any other branch in the phylogeny, even when estimated under a model that accounts for non-stationary composition. Table 4.1 shows estimates for this branch length under a variety of nucleotide and codon models. When the phylogeny is estimated using LogDet distances, which are less sensitive to compositional shifts, this branch length is found to be much shorter than under any other model (1.6 subst./codon site). On the other hand, under codon models, which do not take into account non-stationary codon bias, the estimation for this branch length is extremely high (11.7-50 subst./codon site). These results indicate that models that do not account for compositional heterogeneity tend to yield larger estimates of this branch length when there is a shift in the evolutionary process. Moreover, it is very unlikely that the larger estimates in Table 4.1 are reliable.

Table 4.1. Estimates for the internal branch between ecotypes for the *cpeB* gene under a variety of models.

Estimation method	Branch Length
LogDet	1.6
baseml (Non-Stationary)	5.5
HKY85	6.42
M1	11.7
Model A	50

Note: Branch length is in substitutions per codon site. Estimates from nucleotide models (LogDet, baseml, and HKY85) are multiplied by three to approximate substitutions/codon site.

The *cpeB* dataset is fitted to sites models M1a and M3 as well as branch-sites Model A and Model B. In addition, the HL and LL strains are separately fit to model M3 in order to obtain parameters for use in the simulation studies. Based on parameter estimates from the entire tree under M3, Model A, and Model B, there is no evidence for a fraction of sites under positive selection (Table 4.2). A signal for positive selection does occur under M3 when HL strains are analyzed separately. The separate analysis of HL strains yield strong signal with ω_2=4.95. However, the fraction of sites having evolved under ω_2 is not large (2%). Based on parameter estimates alone, the signal for positive selection in the *cpeB* gene is only marginal.

Table 4.2 Parameter estimates for *cpeB* gene under codon models.

Model	Parameter Estimates	Log Likelihood
M1a	ω_0=0.03635, p_0=0.96753 ω_1=1.00000, p_1=0.03247	-3159.803856
M3 (k=2)	ω_0=0.00780, p_0=0.61206 ω_1=0.09493, p_1=0.38794	-3119.948037
Model A	ω_0=0.00927, p_0=0.74503 ω_1=1.00000, p_1=0.00510 ω_{2FG}=1.00000, p_2=0.24987	-3092.832398
Model B	ω_0=0.00067, p_0=0.27947 ω_1=0.02593, p_1=0.11729 ω_{2FG}=0.19787, p_2=0.60324	-3071.174536
HL: M3 (k=3)	ω_0=0.01770, p_0=0.73560 ω_1=0.39141, p_1=0.24266 ω_2=4.94987, p_2=0.02174	-1340.177146
LL: M3 (k=3)	ω_0=0.00067, p_0=0.71127 ω_1=0.02048, p_1=0.07322 ω_2=0.02066, p_2=0.21551	-1915.765556

Note: Likelihood scores for separate HL and LL analyses are not comparable with the other likelihood scores in the table because the sample of sequences is not the same.

Evidence for historical changes in selection pressures can be formally assessed using LRTs. I have applied three such tests at the codon level. In addition, I have applied an amino acid level test for a shift in functional constraint. At the codon level, Tests 1 and 3 are significant whereas Test 2, the only defensible test for positive selection in the foreground clade, is not significant (Table 4.3). I argued in Chapter 3 that Tests 1 and 3 are better interpreted as tests for functional divergence. Hence, these results suggest a signal for functional divergence among HL and LL *cpeB* gene sequences. However, even that interpretation is not immune to errors arising from the presence of non-stationary codon usage (Chapter 3). Furthermore, the single test applied at the amino acid level (Bivar) is not significant (p=0.15). Given that Bivar is also a formal test for functional divergence, and that it appears to be less sensitive to non-stationary codon usage (see Chapter 3), the significant results for Tests 1 and 3 must be examined more closely.

Further evaluation of the empirical results for *cpeB* is carried out using simulation, presented in the next section, based on parameter values derived from the *cpeB* gene. The benefits of this approach are twofold. First, by using simulation studies based on the values derived from this specific gene I can investigate the possibility that the results for Tests 1 and 3 may have been negatively impacted by non-stationary codon frequencies in this particular case. Second, Zhao and Qin (2007) also found a significant result for Test 1 for *cpeB* gene sequences, but they interpreted the results as significant evidence for positive selection. A directed simulation study can address the question of whether such an interpretation could be valid in this particular case. While simulations in Chapter 3 suggest otherwise, those simulations are more general, with no direct connection to the parameter values characteristic of these *cpeB* gene sequences.

Table 4.3. Results of likelihood ratio tests for *cpeB* gene.

Null	Alternate	2*ΔlnL	d.f.	*p*-value
M1a	Model A	133.94	2	0
Model A (ω_2=1)	Model A	0.00	1	1
M3	Model B	97.5	2	0

4.3.2 The Impact of Non-Stationary Codon Usage on Inferences About Selection Pressure

The goal of this simulation study is to determine whether the degree of non-stationarity present in the *cpeB* gene could have negatively impacted the inference of selection pressure using codon models. Parameter values used for simulations are estimated independently for HL and LL ecotypes (Table 4.4). Because the simulation software requires site categories with equal proportions across the phylogeny, the proportions estimated from the HL strains are used in this simulation study for all parts of the tree. In addition, because ω_2 for the HL data is estimated to be greater than one, this value is scaled back to ω_2=0.9 for simulations. This ensures that no sites are simulated under positive selection, so I can measure false positive rates for the likelihood ratio tests.

Table 4.4. Parameter values used in simulations based on the *cpeB* gene.

		HL	LL
Omega distribution	$p_0 = 0.76356$	$\omega_0 = 0.0177$	$\omega_0 = 0.00067$
	$p_1 = 0.24266$	$\omega_1 = 0.39141$	$\omega_1 = 0.02066$
	$p_2 = 0.02174$	$\omega_2 = 0.9$	$\omega_2 = 0.02048$
GC content		28%	44%

Note: HL and LL empirical codon frequencies are used in the simulations; %GC is presented here as a summary statistic.

Test 1 was originally proposed as an explicit test for positive selection and has been interpreted as such by Zhao and Qin (2007) in their analysis of the *cpeB* gene. However, general simulations conducted in Chapter 3 suggest that this LRT is better interpreted as a test for relaxed constraint. The results from this simulation show that the latter interpretation is true in the case of the *cpeB* gene as well. The majority of replicates yield significant results when there is a shift in the ω distribution, but no positive selection (Table 4.5). This is consistent with the findings from Chapter 3, as well as with previous studies (Zhang et al. 2005).

If instead we view Test 1 as a test for altered selective pressures (regardless of the presence of positive selection) then Test 1 has high power, with 94% of cases detected (Table 4.5). This value increases to 100% when non-stationary selection pressure is combined with a shift in codon bias. However, as seen in Chapter 3, this indicates a systematic error in the direction of the alternative, which should not be relied upon to provide power. It is noteworthy that under homogeneous selection pressure the shift in codon bias present in the *cpeB* gene causes a 32% false positive rate. This is in contrast to my results in Chapter 3 (which found no effect from non-stationary codon bias alone), and is likely because in this case more sites are evolving near neutrality, and thus are more likely to yield a false signal for the alternative model than sites under strong purifying selection. These results therefore indicate that a false positive for a shift in selection pressure for the *cpeB* gene cannot be ruled out.

Table 4.5. Percent replicates (n=100) for Simulation 1 with significant results ($p \leq 0.05$) for likelihood ratio test M1a vs. Model A.

	Same ω values	Δ ω values
Same GC	1	94
Δ GC	32	100

In Chapter 3, the LRT for Test 2 (Model A (ω_2=1) vs. Model A) is found to be the only defensible test for positive selection in part of the phylogeny. Note that while Zhao and Qin (2007) did not apply this test in their study, my application of Test 2 to *cpeB* finds no evidence of positive selection based on this LRT. Here, none of the simulated conditions yield a significant number of false positives for Test 2, indicating that the degree of heterogeneity present in *cpeB* is not enough to cause false positives for this more conservative test. However, it is important to remember that the simulation studies in Chapter 3 revealed that Test 2 has extremely low power.

4.3.3 The Impact of Non-Stationary Codon Usage on the Estimation of Branch Lengths

Non-stationary codon bias may cause errors not only for likelihood ratio test results, but also for estimation of branch lengths. The *cpeB* gene, like many *Prochlorococcus* genes, has an especially long branch between ecotypes (see Figure 4.2). My analysis of those gene sequences reveals that estimations of that branch length are very sensitive to model formulation (Table 4.1). It is possible that under non-stationary evolution, variation associated with a shift in codon bias may be accounted for through an inflated branch length estimate. Therefore, I conduct a simulation to explore the effects of a shift in the evolutionary process on branch length estimation using codon models.

Here, models M1a and Model A are used to estimate the branch length between simulated ecotypes for each of the 16 conditions presented in the methods (Table 4.6). When selective pressures are stationary across the phylogeny, the mean estimated branch lengths for both models are close to the true simulated branch lengths, and errors (2 standard deviations) are consistently less than half of the mean. Surprisingly, the presence of a shift in codon usage alone does not seem to have much effect on the estimation of branch lengths. However, when a shift in ω distribution is present, branch length estimates become inflated as well as increasingly variable. It is interesting to note that under model M1a, estimates are more accurate and less variable when obtained under the combined effects of non-stationary codon bias and selection pressure than under non-stationary selection pressure alone. Estimates of branch lengths under Model A are more inflated than under M1a, with the worst estimate (both with respect to mean and standard deviation) obtained under the combined effects of non-stationary codon bias and selection pressure (*e.g.,* mean=37.8 subst./codon site when simulated length is 10).

Table 4.6. Average branch lengths (in substitutions per codon site) for Simulation 2 estimated under models M1a and Model A.

		Same ω values				Δ ω values			
		M1a		Model A		M1a		Model A	
	BL	mean	2stdv	mean	2stdv	mean	2stdv	mean	2stdv
Same GC	2	2.3	1.0	2.3	1.0	3.4	2.0	4.0	4.3
	3	2.2	1.3	2.2	1.3	5.7	3.3	6.8	4.2
	5	5.5	2.4	5.5	2.4	9.1	5.0	11.5	7.0
	10	10.7	4.4	10.8	4.4	18.2	8.6	23.0	11.7
Δ GC	2	2.1	0.9	2.1	0.9	3.3	1.5	3.3	5.3
	3	3.3	1.6	3.2	1.5	4.5	6.1	7.3	8.3
	5	5.7	2.6	5.7	2.7	6.7	2.8	16.5	14.5
	10	9.8	4.8	10.0	4.9	10.0	4.0	37.8	20.2

Note: For each set of parameters, 100 replicates are analyzed.

Branch length estimation becomes increasingly inaccurate as the true length of the branch increases (Table 4.6). Here I use estimates under Model A in the condition simulated with both non-stationary codon bias and selection pressure as an example. As a percentage of the actual branch length, longer branches are, on average, more inflated under Model A than shorter branches (65% larger when branch length is 2 subst./codon site compared with 278% larger when 10 subst./codon site). The uncertainty of the estimate is also larger for longer branches. A positive relationship between branch length and estimation error is expected because information relevant to the true length of the branch is lost as sequences become saturated with multiple substitutions at a site.

Analyses of the real *Prochlorococcus cpeB* gene suggest both a shift in selection pressures and a relatively long branch. Results from this simulation help explain the extremely large branch lengths estimated under codon models (Table 4.1). In addition, inaccurate branch length estimation is expected to be associated with problems estimating other parameter values from the data. Indeed, parameter estimates from this simulation indicate that, under non-stationary evolution, estimates of the proportion of sites having a shift in selection pressure becomes heavily inflated (Model A estimates 25% of sites with an ω_2=1 in foreground when simulated condition is 2% of sites having ω_2=0.9) (see Appendix I). These results support those found in Simulation 1 for Test 1 in suggesting that model misspecification due to non-stationary evolution leads to a false signal for the alternative model due to problems with the estimates of model parameter values. It seems likely that Test 2 is impacted by this as well, but because the test has such low power under ideal conditions, the impact due to the model misspecification investigated here is small.

4.4 CONCLUSIONS

Previous studies have examined functional constraint in the *cpeB* gene. Ting and colleagues (2001) used pairwise comparisons of d_N/d_S ratios to investigate selection pressure in the *cpeB* gene, finding that these ratios are consistently greater than one. In contrast, the gene encoding the other phycoerythrin subunit, *cpeA*, has d_N/d_S ratios consistently lower than one. This finding suggests that the *cpeB* gene has evolved under positive selective pressures. However, pairwise comparisons oversimplify the substitution process and have no way of accounting for a shift in codon bias, which can negatively impact estimates of d_N and d_S. Another study, conducted by Zhao and Qin (2007), used maximum likelihood branch-site models to carry out Test 1, concluding that the HL *cpeB* branch has experienced positive selection. However, their application of codon models violates the assumption that codon frequencies are homogeneous in all lineages, making this conclusion unfounded. Although Zhao and Qin used a different dataset (environmental rather than genomic sequences), the same patterns of non-stationary evolution exist. In addition, my genomic sequences are more complete, resulting in a more reliable alignment. I found no statistical support for a history of positive selection, and although my results indicate the there is a signal for a simple shift in selection pressure, simulations suggest a false positive cannot be ruled out for this hypothesis. Furthermore, results from amino acid level analysis provide no additional evidence for a shift in functional constraint in the *cpeB* gene of *Prochlorococcus*. Taking together both simulation and real data analysis, I conclude that previous accounts of positive selection in *cpeB* are not as strongly supported by the data as have been suggested.

Chapter 3 highlights the potential negative effects of model misspecification. The simulations carried out in this chapter differ in that they focus on parameters relevant to the *cpeB* gene. These gene-specific simulations provide an understanding of the limits of model-based tests that the more general simulations in Chapters 2 and 3 cannot. While the true gap between the generating model for *cpeB* and the analytical models must be larger than investigated here, this approach reveals how un-modeled variation may be “absorbed” into an inflated branch length estimate. Because codon frequencies are averaged over the entire alignment, when there is a large shift in both codon usage and selection pressures in one part of the phylogeny, there may be systematic error in the estimation of model parameters such as the proportion of sites for each class in the mixture models and the values of ω. These findings highlight the need for users of codon models to carefully examine all parameter estimates for signs of model misspecification.

CHAPTER 5: CONCLUSION

Prochlorococcus provides an interesting example of ecological divergence and is therefore an attractive organism for studying genomic divergence. However, characteristics of the *Prochlorococcus* genome, including non-stationary evolution and recombination, may violate underlying assumptions of models commonly used in evolutionary analysis. In this thesis I separately explored the two main evolutionary forces acting upon *Prochlorococcus* genomes: functional divergence (via substitution) and recombination. This thesis provides preliminary information about the role of these forces in the divergence of HL and LL *Prochlorococcus* ecotypes as well as the performance of commonly used statistical methods on *Prochlorococcus* gene sequences. The specific conclusions for each study are discussed in the last section of each analysis chapter. In fulfillment of the Dalhousie University Faculty of Graduate Studies requirement for a conclusion chapter, I provide a short summary of the main findings from Chapters 2, 3, and 4 in this section of the thesis. The reader is referred to each chapter for in-depth discussion of the involved issues.

In Chapter 2, I evaluated methods for recombination detection under a variety of scenarios that are relevant to functional divergence. I found that tree shape (more specifically asymmetric topology) has the largest impact on false positive rates. However, the extent of this effect differs among the different methods. In addition, non-stationary evolution was found to have a minor effect on some methods. Because the methods measure parameter values from entire alignments, some types of evolutionary heterogeneity among sites may result in false conclusions. Based on the results of my simulations I was able to construct a set of general guidelines for those wishing to use statistical methods for detecting within-gene recombination. These guidelines are presented at the end of Chapter 2.

Because *Prochlorococcus* has diverged into two separate ecotypes, and there are several complete genomes from members of each ecotype, it is an attractive candidate to study shifts in functional constraint at the molecular level. However, current methods designed to detect these shifts do not account for the non-stationary evolution that is characteristic of *Prochlorococcus*. In Chapter 3, I conducted a series of simulations to explore the effects of non-stationary evolution on the inference of selection pressure. I found that under extreme shifts in selection pressure and non-stationary codon bias, even a very conservative codon model might yield false results, leading to false biological conclusions. I used my extensive simulation study to make general recommendations for the use of model-based tests for divergence in the distribution of selection pressures. These recommendations are provided at the end of Chapter 3.

Because current codon models are not robust to non-stationary evolution, they are not appropriate for use on *Prochlorococcus* genes. Despite this fact, studies have appeared in which these models were used to make claims for positive selection. In Chapter 4, I revisit one such case (the *cpeB* gene) and employ simulations based on parameter values estimated from that particular gene. I find that previous claims for positive selection cannot be verified because the amount of heterogeneity present in the *cpeB* gene may cause a false positive for the test used. This analysis provides a specific example of the issues associated with the analysis of real data that is subject to non-stationary evolution, such as *Prochlorococcus* data.

REFERENCES

Ahlgren NA, Rocap G, Chisholm SW. 2006. Measurement of *Prochlorococcus* ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. Environ. Microbiol. 8:441-454.

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18:1585-1592.

Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229-1236.

Aris-Brosou S, Bielawski JP. 2006. Large-scale analyses of synonymous substitution rates can be sensitive to assumptions about the process of mutation. Gene 378:58-64.

Bao L, Gu H, Dunn KA, Bielawski JP. 2008. Likelihood-based clustering (LiBaC) for codon models, a method for grouping sites according to similarities in the underlying process of evolution. Mol. Biol. Evol. 25:1995-2007.

Barker GLA, Handley BA, Vacharapiyasophon P, Stevens JR, Hayes PK. 2000. Allele-specific PCR shows that genetic exchange occurs among genetically diverse *Nodularia* (cyanobacteria) filaments in the Baltic Sea. Microbiology 146:2865-2875.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. U.S.A. 102:14332-14337.

Bielawski JP, Yang Z. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J. Mol. Evol. 59:121-132.

Bielawski JP, Yang Z. Maximum likelihood methods for detecting adaptive protein evolution. In: Statistical methods in molecular evolution—Nielsen R, ed. (2005) New York: Springer-Verlag. 103–124.

Bielawski JP, Dunn KA, Yang Z. 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution: Approximate and maximum-likelihood methods lead to different conclusions. Genetics 156:1299-1308.

Blouin C, Boucher Y, Roger AJ. 2003. Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. Nucleic Acids Res. 31:790-797.

Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL, Case RJ, Doolittle WF. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu. Rev. Genet. 37:283-328.

Brown CJ, Garner EC, Keith Dunker A, Joyce P. 2001. The power to detect recombination using the coalescent. Mol. Biol. Evol. 18:1421-1424.

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. 2001. Universal trees based on large combined protein sequence data sets. Nat. Genet. 28:281-285.

Campbell L, Nolla HA, Vaulot D. 1994. The importance of *Prochlorococcus* to community structure in the Central North Pacific Ocean. Limnol. Oceanogr. 39:954-961.

Chan CX, Beiko RG, Ragan MA. 2006. Detecting recombination in evolving nucleotide sequences. BMC Bioinformatics 7:412.

Chen SL, Hung CS, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer RR, Ozersky P. 2006. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach. Proc. Nat. Acad. Sci. U.S.A. 103:5977-5982.

Chisholm SW, Olson RJ, Zettler ER, Goericke R, Waterbury JB, Welschmeyer NA. 1988. A novel free-living prochlorophyte abundant in the oceanic euphotic zone. Nature 334:340-343.

Chisholm SW, Frankel SL, Goericke R, Olson RJ, Palenik B, Waterbury JB, West-Johnsrud L, Zettler ER. 1992. *Prochlorococcus marinus* nov. gen. nov. sp.: An oxyphototrophic marine prokaryote containing divinyl chlorophyll a and b. Arch. Microbiol. 157:297-300.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF, Chisholm SW. 2006. Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311:1768-1770.

Dammeyer T, Bagby SC, Sullivan MB, Chisholm SW, Frankenberg-Dinkel N. 2008. Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. Curr. Biol. 18:442-448.

Didelot X, Maiden MCJ. 2010. Impact of recombination on bacterial evolution. Trends Microbiol. 18:315-322.

Dorman K. 2007. Identifying dramatic selection shifts in phylogenetic trees. BMC Evol. Biol. 7(Suppl 1):S10.

Doron-Faigenboim A, Pupko T. 2007. A combined empirical and mechanistic codon model. Mol. Biol. Evol. 24:388-297.

Dufresne A, Garczarek L, Partensky F. 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol. 6:R14.

Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, Barbe V, Duprat S, Galperin MY, Koonin EV, Le Gall F, et al. (21 co-authors). 2003. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc. Natl. Acad. Sci. U.S.A. 100:10020-10025.

Dunn KA, Bielawski JP, Yang Z. 2001. Substitution rates in Drosophila nuclear genes: Implications for translational selection. Genetics 157:295-305.

Dunn KA, Bielawski JP, Ward TJ, Urquhart C, Gu H. 2009. Reconciling ecological and genomic divergence among lineages of *Listeria* under an "extended mosaic genome concept". Mol. Biol. Evol. 26:2605-2615.

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

Fitch WM, 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. Biochem. Genet. 4:579-593.

Fletcher W, Yang Z. 2009. INDELible: A flexible simulator of biological sequence evolution. Mol. Biol. Evol. 26:1879-1888.

Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. Science 315:476-480.

Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18:866-873.

Galtier N, Gouy M. 1998. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15:871-879.

Gaucher EA, Miyamoto MM, Benner SA. 2001. Function–structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. Proc. Natl. Acad. Sci. U.S.A. 98:548-552.

Gaucher EA, Gu X, Miyamoto MM, Benner SA. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem. Sci. 27:315-321.

Goericke R, Welschmeyer NA. 1993. The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. Deep Sea Res. I 40:2283-2294.

Goldman N, Yang ZH. 1994. Codon-based model of nucleotide substitution for protein-coding DNA-sequences. Mol. Biol. Evol. 11:725-736.

Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. 16:1664-1674.

Gu X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. Mol. Biol. Evol. 18:453-464.

Gu X. 2006. A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. Mol. Biol. Evol. 23:1937-1945.

Hanage WP, Fraser C, Spratt BG. 2005. Fuzzy species among recombinogenic bacteria. BMC Biol. 3:6.

Hein J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. 98:185-200.

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

Hess WR, Rocap G, Ting CS, Larimer F, Stilwagen S, Lamerdin J, Chisholm SW. 2001. The photosynthetic apparatus of *Prochlorococcus*: Insights through comparative genomics. Photosynthesis Res. 70:53-71.

Husmeier D, McGuire G. 2003. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. Mol. Biol. Evol. 20:315-337.

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

Jakobsen IB, Easteal S. 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput. Appl. Biosci. 12:291-295.

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. 2006. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science 311:1737.

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J et al. (14 co-authors). 2007. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genetics 3:2515-2528.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170-179

Kneip C, Voss C, Lockhart PJ, Maier UG. 2008. The cyanobacterial endosymbiont of the unicellular algae *Rhopalodia gibba* shows reductive genome evolution. BMC Evol. Biol. 8:30.

Knudsen B, Miyamoto MM. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc. Natl. Acad. Sci. U.S.A. 98:14512-14517.

Konstantinidis KT, DeLong EF. 2008. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. The ISME Journal 2:1052-1065.

Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes. Annu. Rev. Microbiol. 55:709-742.

Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. Mol. Biol. Evol. 24:1464-1479.

Kosakovsky Pond SL, Muse SV. 2005a. Site-to-site variation of synonymous substitution rates. Mol. Biol. Evol. 22:2375-2385.

Kosakovsky Pond SL, Muse SV. 2005b. HyPhy: Hypothesis testing using phylogenies. Statistical Methods in Molecular Evolution :125-181.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol. Biol. Evol. 23:1891-1901.

Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. 2005. The net of life: Reconstructing the microbial phylogenetic network. Genome Res. 15:954-959.

Kuo CH, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res. 19:1450-1454.

Lake JA. 1994. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proc. Natl. Acad. Sci. U.S.A. 91:1455-1459.

Lane CE. 2007. Bacterial endosymbionts: Genome reduction in a hot spot. Curr. Biol.17:R508-R510.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. 2:150-174.

Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc. Natl. Acad. Sci. U.S.A. 101:11013-11018.

Liu H, Nolla HA, Campbell L. 1997. *Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean. Aquat. Microb. Ecol. 12:39-47.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605-612.

Lockhart PJ, Steel MA, Barbrook AC, Huson DH, Charleston MA, Howe CJ. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. 15:1183-1188.

Lodders N, Stackebrandt E, Nübel U. 2005. Frequent genetic recombination in natural populations of the marine cyanobacterium *Microcoleus chthonoplastes*. Environ. Microbiol. 7:434-442.

Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC. 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. 73:152-160.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19:1-7.

Mann EL, Ahlgren N, Moffett JW, Chisholm SW. 2002. Copper toxicity and cyanobacteria ecology in the Sargasso Sea. Limnol. Oceanogr. 47:976-988.

Mann NH, Cook A, Millard A, Bailey S, Clokie M. 2003. Marine ecosystems: Bacterial photosynthesis genes in a virus. Nature 424:741-741.

Mann NH, Clokie MRJ, Millard A, Cook A, Wilson WH, Wheatley PJ, Letarov A, Krisch HM. 2005. The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus* strains. J. Bacteriol. 187:3188-3200.

Martin DP. 2009. Recombination detection and analysis using RDP3. Methods Mol. Biol. 537:185-205.

Martin D, Rybicki E. 2000. RDP: Detection of recombination amongst aligned sequences. Bioinformatics 16:562-563.

Marttinen P, Baldwin A, Hanage WP, Dowson C, Mahenthiralingam E, Corander J. 2008. Bayesian modeling of recombination events in bacterial populations. BMC Bioinformatics 9:421.

Maynard Smith J. 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. 34:126-129.

Maynard Smith J, Smith NH. 1998. Detecting recombination from gene trees. Mol. Biol. Evol. 15:590-599.

McCutcheon JP, McDonald BR, Moran NA. 2009. Origin of an alternative genetic code in the extremely small and GC–Rich genome of a bacterial symbiont. PLoS Genetics 5:e1000565.

McManus GB, Dawson R. 1994. Phytoplankton pigments in the deep chlorophyll maximum of the Caribbean Sea and the Western Tropical Atlantic Ocean. Mar. Ecol. Prog. Ser. 113:199-206.

Minin VN, Dorman KS, Fang F, Suchard MA. 2005. Dual multiple change-point model leads to more accurate recombination detection. Bioinformatics 21:3034-3042.

Miyata T, Yasunaga T. 1980. 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. 16:23-36.

Mooers AØ, Holmes EC. 2000. The evolution of base composition and phylogenetic inference. Trends in Ecol. Evol. 15:365-369.

Moore LR, Chisholm SW. 1999. Photophysiology of the marine cyanobacterium *Prochlorococcus*: Ecotypic differences among cultured isolates. Limnol. Oceanogr. 44:628-638.

Moore LR, Goericke R, Chisholm SW. 1995. Comparative physiology of *Synechococcus* and *Prochlorococcus*: Influence of light and temperature on growth, pigments, fluorescence and absorptive properties. Mar. Ecol. Prog. Ser. 116:259-275.

Moore LR, Rocap G, Chisholm SW. 1998. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature 393:464-467.

Moore LR, Post AF, Rocap G, Chisholm SW. 2002. Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. Limnol. Oceanogr. 47:989-996.

Moran NA. 2003. Tracing the evolution of gene loss in obligate bacterial symbionts. Curr. Opin. Microbiol. 6:512-518.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11:715-724.

Narra HP, Ochman H. 2006. Of what use is sex to bacteria? Curr. Biol. 16:R705-R710.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418-426.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929.

Notredame C, Higgins DG, Heringa J. 2000. T-coffee: A novel method for fast and accurate multiple sequence alignment1. J. Mol. Biol. 302:205-217.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299-304.

Pamilo P, Bianchi NO. 1993. Evolution of the zfx and zfy genes: Rates and interdependence between the genes. Mol. Biol. Evol. 10:271-281.

Partensky F, La Roche J, Wyman K, Falkowski PG. 1997. The divinyl-chlorophyll a/b-protein complexes of two strains of the oxyphototrophic marine prokaryote *Prochlorococcus*–characterization and response to changes in growth irradiance. Photosynthesis Res. 51:209-222.

Partensky F, Hoepffner N, Li WKW, Ulloa O, Vaulot D. 1993. Photoacclimation of *Prochlorococcus* sp.(prochlorophyta) strains isolated from the North Atlantic and the Mediterranean Sea. Plant Physiol. 101:285-296.

Paul S, Dutta A, Bag SK, Das S, Dutta C. 2010. Distinct, ecotype-specific genome and proteome signatures in the marine cyanobacteria *Prochlorococcus*. BMC Genomics 11:103.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc. Natl. Acad. Sci. U.S.A. 98:13757-13762.

Pupko T, Galtier N. 2002. A covarion-based method for detecting molecular adaptation: Application to the evolution of primate mitochondrial genomes. Proc. R. Soc. London ser. B. 269:1313-1316.

Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. 2002. Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18:S71.

Ragan MA. 2001. Detection of lateral gene transfer among microbial genomes. Curr. Opin. Genet. Dev. 11:620-626.

Rispe C, Delmotte F, van Ham RCHJ, Moya A. 2004. Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. Genome Res. 14:44-53.

Rocap G, Distel DL, Waterbury JB, Chisholm SW. 2002. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl. Environ. Microbiol. 68:1180-1191.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424:1042-1047.

Sarkar SF, Guttman DS. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. Appl. Environ. Microbiol. 70:1999-2012.

Sawyer S. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. 6:526-538.

Scanlan DJ, Hess WR, Partensky F, Newman J, Vaulot D. 1996. High degree of genetic variation in *Prochlorococcus* (prochlorophyta) revealed by RFLP analysis. Eur. J. Phycol. 31:1-9.

Scheffler K, Martin DP, Seoighe C. 2006. Robust inference of positive selection from recombining coding sequences. Bioinformatics 22:2493-2499.

Schneider A, Cannarozzi GM, Gonnet GH. 2005. Empirical codon substitution matrix. BMC Bioinformatics 6:134.

Shalapyonok A, Olson RJ, Shalapyonok LS. 1998. Ultradian growth in *Prochlorococcus* spp. Appl. Environ. Microbiol. 64:1066-1069.

Shriner D, Nickle DC, Jensen MA, Mullins JI. 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet. Res. 81:115-121.

Snel B, Huynen MA, Dutilh BE. 2005. Genome trees and the nature of genome evolution. Annu. Rev. Microbiol. 59:191-209.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

Strehl B, Holtzendorff J, Partensky F, Hess WR. 1999. A small and compact genome in the marine cyanobacterium *Prochlorococcus marinus* CCMP 1375: Lack of an intron in the gene for tRNA (leu) UAA and a single copy of the rRNA operon. FEMS Microbiol. Lett. 181:261-266.

Suchard MA, Weiss RE, Dorman KS, Sinsheimer JS. 2002. Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. Syst. Biol. 51:715-728.

Suchard MA, Weiss RE, Dorman KS, Sinsheimer JS. 2003. Inferring spatial phylogenetic variation along nucleotide sequences. J. Am. Stat. Assoc. 98:427-437.

Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. 2005. Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. PLoS Biol. 3:e144.

Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biology 4:1344-1357.

Susko E, Inagaki Y, Field C, Holder ME, Roger AJ. 2002. Testing for differences in rates-across-sites distributions in phylogenetic subtrees. Mol. Biol. Evol. 19:1514-1523.

Suzuki K, Handa N, Kiyosawa H, Ishizaka J. 1995. Distribution of the prochlorophyte *Prochlorococcus* in the Central Pacific Ocean as measured by HPLC. Limnol. Oceanogr. 40:983-989.

Swofford DL. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). v. 4.0. Sinaur Associates, Sunderland, MA, USA .

Ting CS, Rocap G, King J, Chisholm SW. 2001. Phycobiliprotein genes of the marine photosynthetic prokaryote *Prochlorococcus*: Evidence for rapid evolution of genetic heterogeneity. Microbiology 147:3171-3182.

Tolonen AC, Aach J, Lindell D, Johnson ZI, Rector T, Steen R, Church GM, Chisholm SW. 2006. Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. Mol. Syst. Biol. 2:53

Urbach E, Scanlan DJ, Distel DL, Waterbury JB, Chisholm SW. 1998. Rapid diversification of marine picophytoplankton with dissimilar light-harvesting structures inferred from sequences of *Prochlorococcus* and *Synechococcus* (cyanobacteria). J. Mol. Evol. 46:188-201.

Vaulot D, Marie D, Olson RJ, Chisholm SW. 1995. Growth of *Prochlorococcus*, a photosynthetic prokaryote, in the Equatorial Pacific Ocean. Science 268:1480-1482.

Wang HC, Spencer M, Susko E, Roger AJ. 2007. Testing for covarion-like evolution in protein sequences. Mol. Biol. Evol. 24:294-305.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189-1191.

Webb A, Hancock JM, Holmes CC. 2009. Phylogenetic inference under recombination using Bayesian stochastic topology selection. Bioinformatics 25:197-203.

West NJ, Scanlan DJ. 1999. Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the Eastern North Atlantic Ocean. Appl. Environ. Microbiol. 65:2585-2591.

West NJ, Schonhuber WA, Fuller NJ, Amann RI, Rippka R, Post AF, Scanlan DJ. 2001. Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by in situ hybridization using 16S rRNA-targeted oligonucleotides. Microbiology 147:1731-1744.

Wiuf C, Christensen T, Hein J. 2001. A simulation study of the reliability of recombination detection methods. Mol. Biol. Evol. 18:1929-1939.

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168:1041-1051.

Worobey M. 2001. A novel approach to detecting and measuring recombination: New insights into evolution in viruses, bacteria, and mitochondria. Mol. Biol. Evol. 18:1425-1434.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568-573.

Yang Z. 2002. Inference of selection from multiple species alignments. Curr. Opin. Genet. Dev. 12:688-694.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586-1591.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46:409-418.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908-917.

Yang Z, Roberts D. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12:451-458.

Yang Z, Nielsen R, Goldman N, Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Zeidner G, Bielawski JP, Shmoish M, Scanlan DJ, Sabehi G, Béjà O. 2005. Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. Environ. Microbiol. 7:1505-1513.

Zhang J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. Mol. Biol. Evol. 21:1332-1339.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472-2479.

Zhao F, Qin S. 2007. Comparative molecular population genetics of phycoerythrin locus in *Prochlorococcus*. Genetica 129:291-299.

Zhaxybayeva O, Doolittle WF, Papke RT, Gogarten JP. 2009. Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus*. Genome Biology and Evolution 2009:325-339.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. 2006. Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. Genome Res. 16:1099-1108.

APPENDIX A: Sites Models

Table A1. Codon models for among site variation in selection pressures (sites models) as implemented in the codeml package of PAML.

Model	Parameters for ω distribution
M0	ω
M1a	p_0 ($p_1 = 1 - p_0$) ω_0<1, ω_1=1
M2a	$p_{0,}$ p_1 ($p_2 = 1 - p_0 - p_1$) ω_0<1, ω_1=1, ω_2>1
M3	$p_{0,}$ p_1 ($p_2 = 1 - p_0 - p_1$) ω_0, ω_1, ω_2
M4	p_0, p_1, p_2, p_3, p_4 ω_0=0, ω_1=⅓, ω_2=⅔, ω_3=1, ω_4=3
M5	α, β
M6	p_0 ($p_1 = 1 - p_0$) α_0, β_0, α_1
M7	*p, q*
M8	p_0 ($p_1 = 1 - p_0$) *p, q,* ω_S>1
M9	p_0 ($p_1 = 1 - p_0$) *p*, *q*, α, β
M10	p_0 ($p_1 = 1 - p_0$) *p*, *q*, α, β
M11	p_0 ($p_1 = 1 - p_0$) *p*, *q*, μ, σ
M12	$p_{0,}$ p_1 ($p_2 = 1 - p_0 - p_1$) ω_0=0, μ_2, σ_1, σ_2
M13	$p_{0,}$ p_1 ($p_2 = 1 - p_0 - p_1$) σ_0, σ_1, σ_2

Notes: Parameters in parentheses are not free parameters. The three distributions used are the beta distribution (*p*, *q*), the gamma distribution (α, β), and the normal distribution (μ, σ).

APPENDIX A: continued

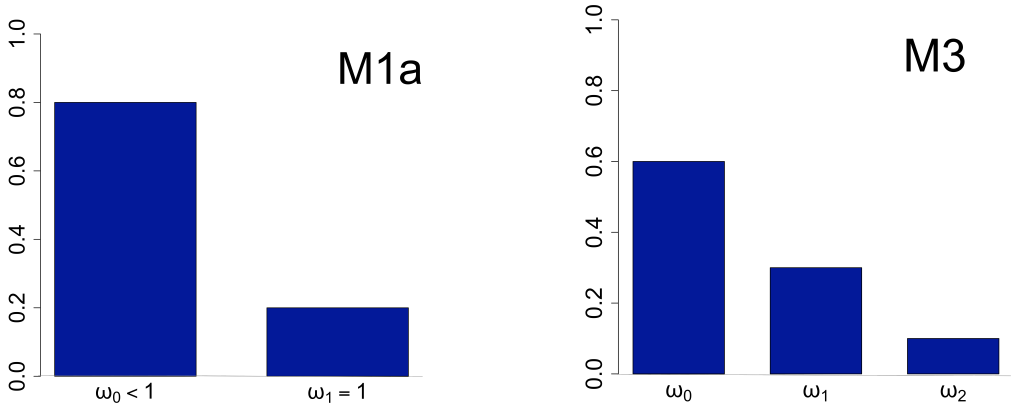


Figure A1. Graphical representation of sites models M1a and M3. Each bar represents a category of sites in the model under different selective pressures, while the height of the bar represents a hypothetical proportion of sites in that category. These proportions (p_i in Table A1) are free parameters in the model, whose values are estimated by maximizing the likelihood of the data. The selection pressure for each category (ω_i) may be an unconstrained parameter, a constrained parameter, or it may be fixed to a pre-specified value, depending on the model. Note that under M1a, site categories are constrained to ω_0<1 and ω_1=1 while under model M3, all site category ω values are unconstrained.

APPENDIX B: Branch-Sites Models

Table B1. Overview of branch-sites Model A.

		Model A		Modified Model A	
Site Class	Proportion	BG ω	FG ω	BG ω	FG ω
0	p_0	$\omega_0 < 1$	$\omega_0 < 1$	$\omega_0 < 1$	$\omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p_1)\, p_0/(p_0 + p_1)$	$\omega_0 < 1$	$\omega_2 \geq 1$	$\omega_0 < 1$	$\omega_2 = 1$
2b	$(1 - p_0 - p_1)\, p_0/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 \geq 1$	$\omega_1 = 1$	$\omega_2 = 1$

Notes: Model A is designed to test for a shift to positive selection from background (BG) to foreground (FG) branches. Modified Model A is used in "Test 2" as a null. For a full description of Model A, see Yang and Nielsen (2002) and Zhang et al. (2005)

Table B2. Overview of branch-sites Model B.

		Model B	
Site Class	Proportion	BG ω	FG ω
0	p_0	ω_0	ω_0
1	p_1	ω_1	ω_1
2a	$(1 - p_0 - p_1)\, p_0/(p_0 + p_1)$	ω_0	ω_2
2b	$(1 - p_0 - p_1)\, p_0/(p_0 + p_1)$	ω_1	ω_2

Notes: Model B is designed to test for any shift in selective pressures between BG and FG branches. For a full description of Model B, see Yang and Nielsen (2002)

Table B3. Overview of branch-sites Models C and D.

		Model C		Model D	
Site Class	Proportion	Clade 1	Clade 2	Clade 1	Clade 2
0	p_0	$\omega_0 < 1$	$\omega_0 < 1$	ω_0	ω_0
1	p_0	$\omega_1 = 1$	$\omega_1 = 1$	ω_1	ω_1
2	$p_2 = 1\text{-}p_0\text{-}p_1$	ω_{2A}	ω_{2B}	ω_{2A}	ω_{2B}

Notes: Models C and D are clade models designed to detect shifts in selective pressures between two clades. For a full description of Models C and D, see Bielawski and Yang (2004).

APPENDIX B: continued

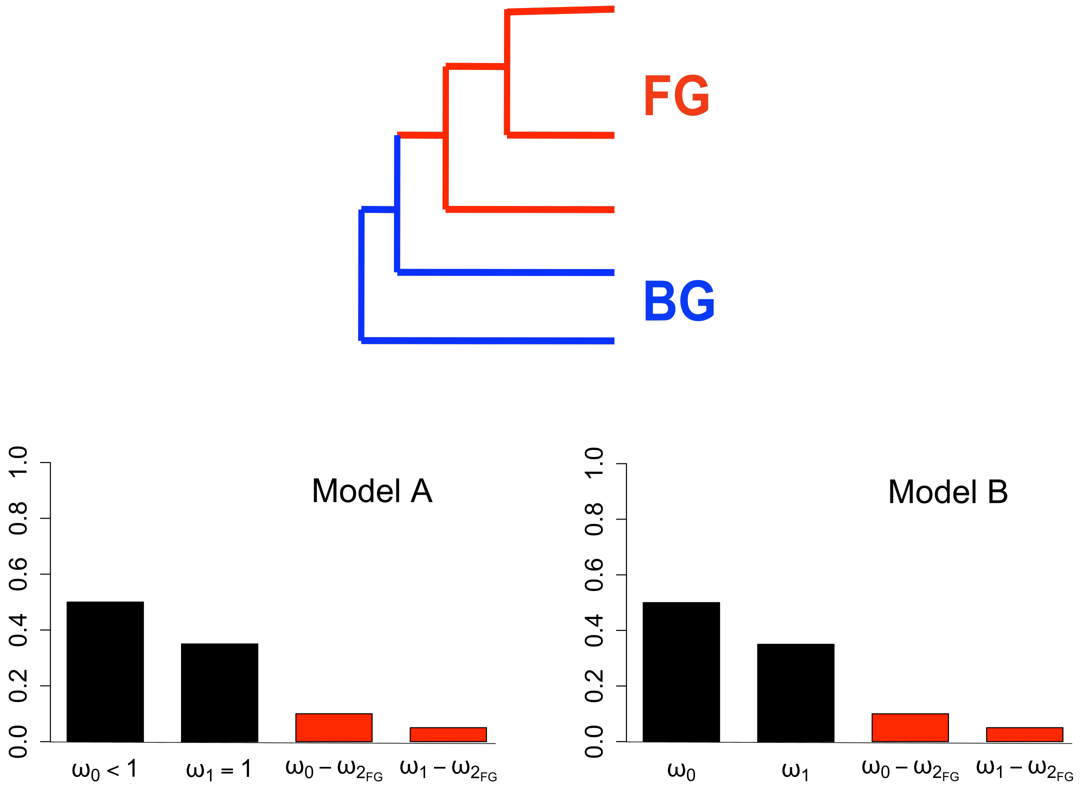


Figure B1. Graphical representation of branch-sites Model A and Model B. Each bar represents a model category under different selective constraints. Bar height represents a hypothetical proportion of sites in that category. These proportions (p_i in Tables B1 and B2) are free parameters in the model, whose values are estimated by maximizing the likelihood of the data. The selection pressure for each category (ω_i) may be an unconstrained parameter, a constrained parameter, or it may be fixed to a pre-specified value depending on the model. Black bars represent site categories in BG and FG branches (*i.e.,* ω_i homogeneous over all branches of the phylogeny). Red bars represent categories with unique levels of selection pressure (independent ω_i parameters) in FG branches (see phylogeny). Note that Model A has the following constraints: ω_0<1, ω_1=1, and ω_2>1. However, for Model B, ω values for site categories are not constrained.

APPENDIX C: Amino Acid Models

Table C1. Summary of main methods for detecting functional divergence on the amino acid level.

Reference	Name	Types	Sites	Gene
Gu 1999	DIVERGE	I	Empirical Bayes	LRT
Gaucher et al. 2001	--	I	Substitution Distribution	10% of sites
Knudsen & Miyamoto 2001	--	I	LRT	NA
Lopez et al. 2002	--	I	Substitution Distribution	NA
Susko et al. 2002	Bivar	I	Substitution Distribution	Parametric Bootstrapping
Pupko & Galtier 2002	--	I	Substitution Distribution	n% of sites (from binomial)
Gu 2006	DIVERGE	II	Empirical Bayes	LRT
Gaston unpublished	FunDi	I & II	Empirical Bayes	**LRT

Notes: "Name" column is the name given to either the program or the software (in the case of DIVERGE), if applicable. "Types" indicates which type of functional divergence the method is designed to detect. "Sites" and "Gene" columns indicate the methods used for inferences at the site and gene levels.

**Not developed by the authors, but tested in this thesis

APPENDIX D: Positive Selection Simulation

It is known that recombination events can cause false signals for positive selection. However, it is not known if the reverse is true. Although the results from Chapter 2 indicate that recombination detection methods do not yield excess false positives when positive selection is present in a part of the tree, positive selection present in the entire alignment may more closely resemble a signal for recombination more closely. Here, I test whether recombination detection methods are negatively impacted when positive selection is present in the whole phylogeny.

Methods

Sequence alignments are simulated in which a fraction of sites are under positive selection throughout the entire phylogeny. I simulate datasets under both asymmetric and symmetric phylogenies (See Figure 2.1). Here, the entire phylogeny is simulated under the same evolutionary model. A U-shaped beta function (p=0.5, q=0.5) is used to model those sites under purifying selection. An additional discrete site category (10%) under positive selection (ω=5) is added. For both asymmetric and symmetric trees, 50 replicate datasets are simulated. These are then compared to the null case (Case 1a), in which there is no positive selection.

Results

None of the methods tested yield higher rates of false positives when positive selection is present throughout the phylogeny (Table D1). These results indicate that, methods for recombination detection are generally robust to this form of positive selection.

APPENDIX D: continued

Table D1. Percent replicates (*n*=50) with signals for recombination for simulation with sites under positive selection in whole phylogeny.

		GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	Chimaera
Symmetric	Null	0	2	2	2	14	16
	Positive Selection	0	2	2	0	10	4
Asymmetric	Null	0	58	10	0	32	36
	Positive Selection	0	46	4	2	20	18

Note: For both symmetric and asymmetric trees, the internal branch length is equal to 0.3 subst./codon site. This results in root-to-tip lengths of 1.2 subst./codon (symmetric) and 2.7 subst./codon (asymmetric).

APPENDIX E: Recombination Results from Non-Stationary Simulations

Table E1. Percent false positives (n=50) for Simulation 3 under a symmetric tree with a) homogeneous codon bias or b) a shift in codon bias ($\eta_2 = 0.1$) using different methods of the detection of recombination events.

a)

Case	GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	CHIMAERA
1a	0	4	10	0	12	14
1b	0	0	2	0	12	20
1c	0	2	2	0	14	16
2a	0	2	2	2	10	16
2b	0	0	0	2	8	8
2c	0	2	4	2	8	12

b)

Case	GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	CHIMAERA
1a	0	6	6	0	10	14
1b	0	6	6	2	16	20
1c	0	2	2	0	2	10
2a	0	10	10	0	20	28
2b	0	2	2	0	10	14
2c	0	16	16	2	8	14

APPENDIX E: continued

Table E2. Percent false positives (*n*=50) for Simulation 3 under an asymmetric tree with a) homogeneous codon bias or b) a shift in codon bias ($\eta_2 = 0.1$) using different methods of the detection of recombination events.

a)

Case	GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	CHIMAERA
1a	0	58	10	0	32	36
1b	0	92	8	4	20	32
1c	2	80	10	0	28	40
2a	0	80	20	4	44	48
2b	0	78	4	0	32	42
2c	0	88	10	2	34	36

b)

Case	GARD-MBP	GARD-SBP	RDP	GENECONV	MaxChi	CHIMAERA
1a	2	58	18	2	38	44
1b	0	90	8	0	34	42
1c	0	72	6	2	14	38
2a	0	78	18	6	46	48
2b	0	66	12	0	40	44
2c	0	64	18	2	40	46

Appendix F: Codon Level Analysis of Non-Stationary Simulations

Table F1: Percent replicates (*n*=50) with significant likelihood ratio tests (p<0.05) based on branch-site codon models when a) no shift in codon bias is simulated and b) when a large shift in codon bias is simulated (η_1=0.5, η_2=0.1).

a)

	M1 vs. ModelA	ModelA(w=1) vs. ModelA	M3 vs. ModelB
Case1a	4	0	8
Case1b	6	2	78
Case1c	52	0	68
Case2a	98	0	92
Case2b	88	0	78
Case2c	96	0	86
Case3a	96	0	96
Case3b	90	0	94
Case3c	94	6	98

b)

	M1 vs. ModelA	ModelA(w=1) vs. ModelA	M3 vs. ModelB
Case1a	4	0	18
Case1b	24	0	76
Case1c	82	0	98
Case2a	98	0	100
Case2b	94	4	98
Case2c	92	24	96
Case3a	98	10	100
Case3b	98	8	100
Case3c	86	52	100

Appendix G: Amino Acid Level Analysis of Non-Stationary Simulations

Table G1. Percent replicates (*n*=50) with signals for functional divergence ($p \leq 0.05$) under Bivar and FunDi with different shifts in selection pressure and with a) homogenous codon bias and b) a shift in codon bias.

a)

	Bivar			FunDi
	arsum	*alrsum*	*abrsum*	*2ΔlnL*
Case 1a	36	4	2	14
Case 1b	38	56	54	14
Case 1c	68	62	62	44
Case 2a	80	64	68	48
Case 2b	76	30	30	34
Case 2c	80	64	66	46
Case 3a	74	50	50	38
Case 3b	60	10	12	18
Case 3c	72	24	30	48

b)

	Bivar			FunDi
	arsum	*alrsum*	*abrsum*	*2ΔlnL*
Case 1a	64	18	16	36
Case 1b	80	74	78	10
Case 1c	92	86	84	36
Case 2a	96	94	94	70
Case 2b	96	34	32	50
Case 2c	94	62	66	66
Case 3a	78	52	52	40
Case 3b	88	36	34	52
Case 3c	96	22	26	48

APPENDIX H: Amino Acid Level Simulation

The sensitivity of amino acid models to non-stationary codon bias may be due to the discrepancy between the simulating model and the analyzing model. Because sequences are simulated on the codon level, evolutionary rate is simulated in terms of the ratio of nonsynonymous to synonymous substitutions. However, at the amino acid level, the sequences are analyzed based on a substitution matrix, which takes into account differences in replacement frequencies between amino acids. To see if this difference in models accounts for the impact of non-stationary codon on amino acid models observed in Chapter 3, I conduct a small simulation at the amino acid level.

Methods

Fifty replicate datasets are simulated with non-stationary amino acid frequencies, but stationary evolutionary rate. The same asymmetric tree used for the codon level simulations, with two types of evolution, is used here as well. Amino acid frequencies are measured from the codon level simulations so that the shift in frequencies is comparable to that in the non-stationary codon simulations. Sequences are simulated under a WAG model with the proportion of invariable sites set to 0.24 and the alpha parameter for among-site rate distribution set to 0.7. These are the average values estimated from the simulated replicates for Case 1a. Each replicate data set is analyzed with both FunDi and Bivar as described in Chapter 3. Parametric bootstrapping is used to determine a *p*-value for each dataset.

Results

My results indicate that neither method is significantly impacted by a shift in amino acid composition. FunDi yielded 6% false positives while Bivar yielded only 2%. This indicates that the increase in false positives associated with a shift in codon bias observed in Chapter 3 is likely due to a discrepancy between generating and analytical models rather than sensitivity of the models to non-stationary composition.

APPENDIX I: Parameter Estimates for *cpeB* Simulations

Table I1. Parameter estimates under Model A for simulations based on *cpeB* gene.

		Same ω		$\Delta\omega$	
	BL	p_{FG}	ω_2	p_{FG}	ω_2
Same GC	2	0.001	1.15	0.02	1.15
	3	0.001	1.14	0.03	1.2
	5	0.001	1.08	0.03	1.16
	10	0.001	1.01	0.03	1.1
Δ GC	2	0.03	1.13	0.24	1
	3	0.03	1.02	0.24	1
	5	0.03	1.15	0.24	1
	10	0.03	1.11	0.25	1

Notes: Values are averaged over 100 replicates. The parameter p_{FG} indicates the sum of site categories with a shift to ω_2 in foreground branches.