**Université de Montréal**

# Bayesian codon models for detecting convergent molecular adaptation

par

**Sahar Parto**

Département de Biochimie et Médecine Moléculaire

Faculté de Médecine

Thèse présentée en vue de l'obtention du grade de doctorat en Bio-informatique

November 2017

# RÉSUMÉ

Modéliser le jeu combiné de la mutation et de la sélection au niveau moléculaire représente un des objectifs majeurs des sciences de l'évolution. L'acquisition massive de séquences génétiques au cours des dernières années a fourni un matériel abondant pour de telles analyses empiriques. Les modèles à codons sont de plus en plus utilisés en vue de fournir une description réaliste des processus de substitution des séquences codant pour les protéines. Parmi eux, les modèles mécanistes paramétrisent de façon séparée les effets mutationnels et sélectifs qui se combinent au sein du processus substitutionnel. Ces approches mécanistes caractérisent les effets sélectifs en s'appuyant sur un modèle explicite du paysage de fitness auquel la séquence protéique est soumise. Toutefois, jusqu'à présent, le paysage de fitness a toujours été considéré comme constant, alors qu'il existe des situations empiriques pour lesquelles le paysage de fitness subit en réalité des fluctuations écologiques au cours du temps. Lorsqu'une information empirique est par ailleurs disponible, concernant des différences systématiques de pression de sélection en fonction des fluctuations environnementales, il est alors possible de modéliser explicitement ces modulations du paysage de fitness.

Nous avons développé un modèle à codons mécaniste, dont le but est de détecter ces effets sélectifs différentiels dépendant des conditions environnementales. Ce modèle a été implémenté dans un cadre d'inférence bayésienne, et a tout d'abord été appliqué au cas de l'évolution du VIH. Le VIH évolue sous la pression du système immunitaire de son hôte humain. Notre modèle de sélection différentielle (DS) décrit les mécanismes détaillés de l'évolution du VIH sous les contraintes induites par le fond génétique de l'hôte (par exemple, le HLA). De ce fait,

il permet de trouver des associations entre adaptations du virus et profil HLA des hôtes. À long terme, notre approche permettra une meilleure compréhension du phénomène d'échappement du virus à la surveillance immunitaire de l'hôte, ce qui fournira alors des informations utiles en vue de l'élaboration d'un vaccin efficace contre le SIDA. Nous avons également appliqué notre modèle au gène de la Rubisco, une enzyme responsable d'une étape majeure de la photosynthèse. L'évolution de la Rubisco semble montrer des différences systématiques entre plantes dites C3 et C4, différences liées à des changements environnementaux. En utilisant le modèle DS, nous avons mis en évidence des effets systématiques d'adaptation convergente au niveau moléculaire, chez les espèces C4, par rapport aux espèces C3. Finalement, nous avons contrasté les résultats obtenus avec le modèle DS sur cet exemple avec ceux fournis par les modèles à codons classiques, basés sur l'estimation du $d_N/d_S$. Cette analyse comparée nous permet d'illustrer une différence conceptuelle fondamentale entre ces deux types de modèles à codons, concernant le type de régime sélectif que chaque type de modèle cherche à caractériser: à savoir, sélection directionnelle, contre adaptation continuelle.

**Mots clés** : Évolution, Mutation, Modèle à codon, Pression de sélection, Inférence bayésienne, Sélection différentielle, VIH, Rubisco

# ABSTRACT

Modeling the interplay between mutation and selection at the molecular level is one of the primary goals in molecular evolution. Massive acquisition of genetic sequence data in recent years has provided a wealth of information for such empirically-driven studies. Codon-based models are increasingly used to give a realistic description of the substitution process in protein-coding genes. Among them, the mechanistic codon-based modeling approach distinctly parameterizes mutational and selective effects bearing on the overall substitution process. These mechanistic approaches characterize the selective pressure by relying on an explicit model of the amino acid fitness landscape over the sequence. Thus far, a constant fitness landscape has generally been assumed. Yet, there are some situations in which the fitness landscape experiences some environmental fluctuations through time. When the empirical knowledge about the systematic difference in selective pressures is available, regarding the fluctuating environment, it is possible to explicitly model condition-specific amino acid fitness modulations.

In this thesis, we developed a codon-based model to capture these differential condition-specific selective effects on coding sequences. This model was implemented in a Bayesian framework and was first applied to HIV, which evolves under the selection pressure of the host immune system. Our Differential Selection (DS) model describes the detailed mechanisms of evolution of HIV under the constraints defined by host genetic backgrounds (e.g., Human Leukocyte Antigen). Therefore, it is possible to find associations between specific viral adaptations and specific HLA alleles of the hosts. Ultimately, our approach will enable us to understand better how the virus escapes from the host immune response, which will, in turn,

provide a useful guideline for designing an efficient vaccine against AIDS. We also applied the DS model on Rubisco, an enzyme responsible for a major step in photosynthesis. The evolution of Rubisco has been shown to be different in C3 and C4 plants, as a consequence of differing environmental conditions. We used the DS model to reveal the consistent patterns of convergent adaptation in Rubisco in C4 plants, compared to C3 plants. Finally, we contrasted our results from DS model with those obtained under classical codon models based on the estimation of $d_N/d_S$. This comparative analysis allows us to illustrate a fundamental conceptual difference between these two types of codon models, which are meant to detect different selective regimes: directional selection versus ongoing adaptation.

**Key words:** Evolution, Mutation, Codon models, Selective pressure, Bayesian inference, Differential Selection, HIV, Rubisco

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| DS | Differential Selection |
| HLA | Human leukocyte antigen |
| CTL | Cytotoxic T lymphocyte |
| MCMC | Markov chain Monte Carlo |
| LANL | Los Alamos National Laboratory |
| EC | Elite Controllers |
| CP | Chronic Progressors |
| Rubisco | Ribulose-1,5-biphosphate carboxylase/oxygenase |
| pp | Posterior probability |
| rbcL | Rubisco large subunit |
| rbcS | Rubisco small subunit |
| PNUE | Photosynthetic nitrogen use efficiency |
| RuBP | Ribulose 1,5 biphosphate |
| CAN | Capsid N-terminal |

**To my angles, Ryan and Keyaan, and my parents**

# ACKNOWLEDGEMENT

At this last stage of a long journey of this Ph.D., my thought is with all the people who helped me along the way. I am grateful to many people, some of them inspired me, some made these years remembering, and some of them helped me through the difficult times. First, I want to thank my supervisor Dr Nicolas Lartillot for the many things I have learned from him; the thoroughness when analyzing data, the respect for other's ideas, and the critical spirit. Also, I gratefully appreciate for his patience and support while I was struggling to balance research with family life.

I am very much grateful to Gertrude Burger, Franz Lang and Adrian Serohijos who kindly provided me with their lab facilities, while my supervisor had moved to France. My sincere gratitude is to all the people in the Cedergren Bioinformatics Center, who made it such an enriching environment. To Raphael Poujol, who were a big help and a pleasure to work with during the beginning years of my Ph.D. To Sandrine Moreira, Matt Sarrasin, Simon Laurin-Lemay, Jean-Francois Theroux, Nicolas Schweiger and finally, Pouria Dasmeh for their friendship, support, and help, both in software and hardware.

I would like to express my appreciation to Elaine Meunier, Sylvie Beauchemin, Lorraine Bidegare Charette, and Linda D'Astous. Their efficiency and kindness were much helpful along the administrative pathway within the university.

During the development of this dissertation, I received financial support from the Natural Sciences and Engineering Research Council of Canada and the biT fellowships for excellence

(a Canadian Institutes of Health Research strategic training program grant in bioinformatics). I am thankful to them.

Finally, I am forever beholden to my family. My parents are, of course, ultimately responsible for everything I have achieved. Words of appreciation cannot express the gratefulness I feel towards them. To my brothers, who were and are always by my side, on my mind, and in my heart. No one, however, helped me more constantly and directly in pursuing this path than my husband, Farshid. During all these years, he has been constantly my support.

# AT A GLANCE

Evolutionary theory is the conceptual foundation of biology. In the words of Theodosius Dobzhansky: "Nothing in biology makes sense except in the light of evolution"(Dobzhansky 1973). Since Darwin's seminal work, studies of organisms have been vastly improved. Molecular evolution is one of the most instrumental parts of these noble endeavors. It explores the transformation in DNA or protein sequences across generations, emphasizing on the processes leading to current and potentially prediction of future DNA and protein sequences.

In the 1960s, the homologous biomolecules from different species were used to derive their evolutionary history (Zuckerkandl and Pauling 1962; Zuckerkandl and Pauling 1965). In the following decades, groundbreaking improvement in DNA cloning and genome sequencing resulted in rapid accumulation of genetic sequence data which revolutionized phylogenetic and molecular evolutionary studies and established its position in many biotechnological applications. The advancement in computer science and the development of sophisticated analytical methods, together with the above improvements, eventually lead to the modern field of computational evolutionary biology. For example, the statistical analysis of evolutionary relationships among lineages is particularly useful for understanding the viral quasispecies epidemiology, transmission routes, their origin and subsequent evolution (Pybus and Rambaut 2009).

An acute need for more powerful statistical methods and efficient computational algorithms to enable a practical analysis and interpretation of evolutionary processes is the inspiration for this work. More precisely, the main goal of my Ph.D. work has been to develop Bayesian models of

protein-coding sequence evolution that can capture the consistent patterns of molecular adaptation that may occur in correlation with recurrent environmental changes (i.e., patterns of convergent evolution, or *differential selection* as a function of the environment). These differential selection models are somewhat sophisticated in their statistical design and their MCMC implementation.

These models were then applied to two central examples; first, to HIV (Human Immunodeficiency Virus) sequence, which is one of the fastest evolving viruses. This virus evolves under the constant selection pressure induced by the host immune response. Second, to the Rubisco sequence, which is an enzyme encoded in plant chloroplast and responsible for a major step in photosynthesis. The fact that these two example applications unfold over vastly different scales in terms of time, space, and rate of evolution, clearly illustrates that the method is sufficiently general to be broadly applicable to many other cases where condition-dependent selective effects are suspected.

Chapter one of this manuscript presents an introduction to probabilistic models in evolutionary biology. First, the statistical methods used in molecular phylogenetics are introduced, and then a brief review of the existing codon models used in molecular evolutionary studies will be presented.

In the second chapter, an introduction to HIV, its biological aspects and its relationship to the host genetic background will be given. One of the essential selective pressures imposed by the human immune response on HIV is mediated by the HLA (Human Leukocyte Antigen) system, and this provides a particularly striking example on which to apply our Differential Selection model. Accordingly, in this chapter, the Differential Selection model is introduced and applied to HIV.

The second application of Differential Selection model was conducted on Rubisco sequences in the more general context of a comparative analysis of current classical codon models (based on the estimation of the ratio of non-synonymous to synonymous substitution rates, $d_N/d_S$) and our Differential Selection model. To put things in context, chapter three contains an introduction to the biology and the biochemistry of Rubisco and the diversity of photosynthetic regimes in plants (so-called C3 and C4). This introduction is followed by the proper comparative analysis of alternative modeling strategies. This comparison will highlight an important point: how different types of codon models (here, Differential Selection versus $d_N/d_S$ codon models) formalize different types of selective regimes (inherently, diversifying versus directional selection). This, in turn, illustrates how codon models, which often referred to the classical $d_N/d_S$ models, have in fact the potential for a much broader scope of model-based investigation of the selective regimes experienced by protein-coding genes. Finally, chapter four presents concluding remarks and future perspectives.

The work presented in this dissertation has been published in the following articles:

**Chapter 2**

Parto, S. and N. Lartillot, 2017, Detecting consistent patterns of directional adaptation using differential selection codon models. BMC Evolutionary Biology, 17: p. 147.

**Chapter 3**

Parto, S. and N. Lartillot, 2017, Molecular adaptation in Rubisco: discriminating between convergent evolution and positive selection using mechanistic and classical codon models. PLOS ONE, 13 (2), e0192697.

# CHAPTER 1: INTRODUCTION

## 1.1 Statistical models of molecular evolution

### 1.1.1 Likelihood-based methods

Probabilistic models have played a central role to model the evolution of biological sequences. These models provide the ground for likelihood-based methods, which are now routinely used in phylogenetic studies. In the 1970s, Sforza and Edwards (Cavalli-Sforza and Edwards 1967) introduced the first maximum-likelihood approach for reconstructing phylogenetic trees. Their model is based on a Brownian-motion for describing the evolution of gene frequencies along the lineages of the phylogenetic tree. About ten years later in 1981, Felsenstein (Felsenstein 1981) developed the first Markovian model of DNA evolution by point substitutions and introduced computational methods for calculating and maximizing the likelihood under this model, for arbitrary phylogenies. This article opened a new field of research, leading to the development of increasingly sophisticated models and techniques for the estimation of phylogenetic trees from DNA sequences.

In contrast to previous methods, such as maximum parsimony, likelihood-based methods (i.e., maximum likelihood and Bayesian inference) represent a more principled approach. More precisely, their main advantage is to correctly separate and correctly articulate two distinct conceptual levels; first, the assumptions that are made about the underlying evolutionary processes (represented by the probabilistic model), and second, the general statistical methodology for

testing hypotheses and estimating parameters, conditional on those assumptions (the statistical paradigm, being either maximum likelihood or Bayesian inference). Thanks to this correct formalization, likelihood-based methods make it possible to go back and forth between assumptions, hypotheses (to be empirically tested), models, and empirical data. Thus, progressively, models can be designed based on current hypotheses, tested, rejected, and improved, by relaxing specific hypotheses or making new hypotheses, and so on. As can be attested by the vast literature published in the field of molecular evolution over the last 20 years, this statistically-controlled research cycle opens the way to a model-based progress of our understanding of molecular evolutionary processes.

### 1.1.2 Maximum likelihood versus Bayesian inference

As mentioned above, two alternative statistical paradigms have been proposed in the context of likelihood-based phylogenetics: these are represented by maximum likelihood estimation and Bayesian inference. Since the work presented here is exclusively Bayesian, a more detailed introduction will be given for this statistical paradigm.

Bayesian inference is a method in which Bayes theorem is used to update the probability of a hypothesis based on new emerging information. According to Bayes theorem, the probability distribution of the parameter θ given data (D) is

$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{\int_{\Theta} p(D|\Theta)p(\Theta)d\Theta} \qquad (1\text{-}1)$$

In the Bayesian parlance, $p(\theta|D)$ is referred as posterior probability, and $p(\theta)$ and $p(D|\theta)$ are called prior and likelihood, respectively. The denominator of the equation is recognized as the marginal likelihood. The prior distribution of a parameter is supposed to represent our knowledge of the parameter before having observed the data. On the other hand, the posterior probability represents our knowledge after seeing the data. So, Bayesian inference is presented as a method that allows one to incorporate prior knowledge into the analysis (through the prior). However, beyond the advantage of incorporating prior knowledge, another role of the prior is to allow for averaging over uncertainty (Huelsenbeck, Rannala et al. 2000). In fact, in many applications of Bayesian inference in phylogenetics, the priors that have been used so far are either non-informative or weakly informative, thus suggesting that this second role of the prior is perhaps more fundamental than the question of incorporating subjective prior knowledge, at least in Bayesian phylogenetics.

Bayesian inference has been introduced in molecular phylogenetics by Rannala and Yang (Rannala and Yang 1996) as an alternative to maximum likelihood. However, their method was useful only for very small datasets (few species). It was later improved by the explicit use of a Markov chain Monte Carlo (MCMC) method (Yang and Rannala 1997; Larget and Simon 1999). A Bayesian approach using MCMC has a considerable computation advantage over methods using maximum likelihood, especially on larger datasets. In fact, the introduction of MCMC is undoubtedly among the most significant developments in phylogenetics over the past two decades.

Combined with MCMC methods (Huelsenbeck, Larget et al. 2002), Bayesian inference makes it relatively easy to design complex hierarchical models, allowing for extensive modulations of the substitution process across genes, across sites and lineages of the phylogeny (Lartillot 2015). In

the present Ph.D. work, we will make extensive use of this advantage of Bayesian inference, in particular, by modulating the substitution process simultaneously across sites and environmental conditions. Computationally, such sophisticated models, raise the problem of integrating over the high-dimensional spaces implied by their large number of nuisance parameters (or random effects). For sufficiently complex models, this can be done only by using MCMC methods. In this context, the problem of controlling the statistical complexity of the model (e.g., avoiding over-parameterization) is naturally implemented through the use of hierarchical priors. The models introduced in this dissertation represent a particularly clear example of this 'rich-model' philosophy.

## 1.2 Models of molecular evolution

Substitution models describe the process of replacement of one nucleotide by another nucleotide over evolutionary times. These models are obviously simplifications of the real evolution, but they are widely used as good approximations. In particular, it should be emphasized here that the substitution process takes place at the level of the species: technically, a substitution is the replacement, at the level of the entire population, of an allele by another allele, itself initially produced by a point mutation. Thus, the substitution process is distinct from the mutation process, since it includes the combined effects of natural selection and random drift. This point is explicitly elaborated in the context of the so-called mutation-selection models (see Section 1.5.3). However, most of the substitution models that are currently used in phylogenetics, and which are now briefly introduced, are phenomenological: they directly parameterize the net rate of substitution, without trying to tease out the various evolutionary forces involved in this complicated process.

Substitution models are classified into three main groups according to the molecular units of substitution: nucleotide, amino acid or codon-based models. Regardless of the type of the model, and mostly for computational reasons, they all have assumed that the substitution process was Markovian. Generally, in Markovian models, the behavior of the process at time $t$ only depends on the current state. Thus, if the current state is specified, no additional information of the past states is needed to predict the future states, i.e., it is a memoryless process. Also, with rare exceptions (Kleinman, Rodrigue et al. 2010), the evolution of each site along the sequence is assumed to be independent of the state at all other sites. With these assumptions, the substitution process at each site is entirely characterized by an instantaneous rate matrix $Q$**.** This matrix has off-diagonal entries $Q_{ij}$ equal to the instant rates of substitution from nucleotide $i$ to nucleotide $j$.

Since the introduction of the first probabilistic models by Felsenstein (Felsenstein 1981), a long series of models of sequence evolution have been proposed. These models have introduced at least two distinct levels of complexity: either by invoking more and more complex substitution rate matrices or by introducing modulations of the substitution process across sites and branches. First, I give a brief introduction of the most classical substitution rate matrices that have been proposed in the literature. Then, I discuss the problem of how heterogeneity in substitution process across sites or branches has been modeled.

### 1.2.1 Models of nucleotide substitutions

The nucleotide substitution models have generally been parameterized in terms of two sets of parameters; nucleotide equilibrium frequencies, on one hand, and rates of nucleotide

exchangeability, on the other hand. The nucleotide equilibrium frequencies account for compositional constraints of the sequence. Technically, they represent the asymptotic composition of a long sequence of DNA evolving under the process specified by the substitution model of interest. The nucleotide exchangeability parameters, on the other side, capture the transient features of the substitution process. In practice, they typically correlate with the biochemical similarity between nucleotides. For example, transitions (among purines or pyrimidines) are biochemically more conservative, and also tend to happen more frequently, than transversions (from a purine to a pyrimidine or vice-versa).

The first DNA substitution model was proposed by Jukes and Cantor (Jukes and Cantor 1969). In this so-called JC69 model, all nucleotide substitutions have the same rate: i.e., $Q_{ij}=\alpha$ for all nucleotides *i* and *j.* The nucleotide frequencies implied by this model are thus all equal to 0.25 for the four nucleotides. Kimura (Kimura 1980) proposed a two-parameter model (K80) which distinguishes between transitions and transversions, so the instantaneous rate matrix is either equal to $\alpha$ for transitions, or $\beta$ for transversions. This model also has uniform equilibrium frequencies over the nucleotides. In 1981, Felsenstein (Felsenstein 1981) extended JC69 to a new model (F81), in which the nucleotide frequencies are not considered equal to 0.25 ($\pi_A \neq \pi_T \neq \pi_C \neq \pi_G$). Instead, they can be arbitrary (and are typically co-estimated with the phylogeny, by maximum likelihood). Later on, Hasegawa *et al.* (Hasegawa, Kishino et al. 1985), proposed a model (HKY85) by combining F81 and K80. In other words, they add the transition/transversion bias to the frequency-based nucleotide model of F81. Tamura and Nei (Tamura and Nei 1993) developed another model (TN93) that is able to distinguish between the rate of A↔G transition and C↔T transition, whereas transversions occur at the same rate, but different from the two transition rates.

Finally, the GTR model (general time-reversible) is the most general, independent, finite-site and time-reversible model, which was proposed by Tavaré (Tavaré 1986). The nucleotide frequencies ($\pi$) and the nucleotide exchangeability parameters ($\rho$) for all pairs of nucleotides are arbitrary, under the constraint that $\rho_{ij}=\rho_{ji}$ (to ensure reversibility) and that the nucleotide frequencies sum to 1. Thus, the *Q* matrix for GTR model is as follows

$$Q_{ij} = \begin{bmatrix} \bullet & \rho_{AC}\pi_C & \rho_{AG}\pi_G & \rho_{AT}\pi_T \\ \rho_{CA}\pi_A & \bullet & \rho_{CG}\pi_G & \rho_{CT}\pi_T \\ \rho_{GA}\pi_A & \rho_{GC}\pi_C & \bullet & \rho_{GT}\pi_T \\ \rho_{TA}\pi_A & \rho_{TC}\pi_C & \rho_{TG}\pi_G & \bullet \end{bmatrix} \qquad (1\text{-}2)$$

where $\sum_{1}^{4} \pi_j = 1$. The diagonal elements of *Q* are defined such that the sum of the elements of each row equals to zero.

This model is the most independent substitution model and widely used in evolutionary studies. Using GTR model reduces the number of the parameters by assuming time reversibility. The GTR model is used for describing the mutational process in all codon models developed in this thesis.

### 1.2.2 Models of amino acid replacements

For many analyses, particularly over broad evolutionary distances, the evolution of protein-coding sequences is often directly modeled on the amino acid level. The main justification for this recoding of the data is that synonymous substitutions tend to be mutationally saturated. In this context, the state space of the Markov process of point substitution is now equal to 20, instead of 4. In contrast to nucleotide substitution models, whose parameters are generally co-estimated with the tree in a typical phylogenetic analysis, models of amino acid replacement were first designed using empirical approaches. Firstly, Dayhoff *et al.* (Dayhoff 1978) proposed a model of amino acid replacement, in which substitution rates are derived from protein alignment with more than 85% similarity. The Dayhoff matrices, known as PAM, are mainly used by database search methods.

A few years later, Jones *et al.* (Jones, Taylor et al. 1994) used a similar methodology to derive a substitution matrix, particularly adapted to transmembrane proteins. This matrix has entirely different values from those of Dayhoff, suggesting a significant difference in amino acid substitutions patterns for membrane proteins, compared to soluble proteins. Later on, Adachi and Hasegawa (Adachi and Hasegawa 1996) used mitochondrial proteins of 20 vertebrates to obtain a substitution matrix that can be efficiently used for mitochondrial protein sequence alignments. On the other hand, Henikoff and Henikoff (Henikoff and Henikoff 1992) applied a different approach to derive the BLOSUM matrices, using local, ungapped alignment of distant protein sequences.

Recently, structural constraints of proteins were also used to develop new evolutionary models. For example, Lio *et al.* (Lio, Goldman et al. 1998) and Thorne *et al.* (Thorne, Goldman et

al. 1996) implemented the protein secondary structure and Pollock *et al.* (Pollock, Taylor et al. 1999) 3D structure information in their amino acid substitution models.

### 1.2.3 Codon substitution models

Thus far, many of the models above have been mainly used for reconstructing phylogenies or estimating divergence time. Another field of evolutionary studies focuses on characterizing the forces acting on genetic sequences. In particular, much effort has been devoted to the characterization of the selective regimes acting on protein-coding sequences (Jukes and King 1979). This objective has been achieved by the development of codon-based models. I will go to the details of these models in section 1.5 after other important general concepts behind current phylogenetic models are introduced.

## 1.3 Modeling variation across sites and across lineages

For many reasons, the substitution process is likely to vary substantially among genes, or sites within proteins, or even closely related species. The possible biological causes of this variation are abundant: different proteins have different functions. Different sites are either buried or exposed in the 3D structure, and as a result, are likely to accept different amino acids. Finally, the mutation rate, the GC bias, and the force of selection are all known to vary among species, depending on variation in life-history traits, such as generation-time, or population size. As a result, it has quickly been realized in the phylogenetic community that accounting for the resulting variation among sites, or between lineages, is essential. Doing so, however, substantially increases the complexity

of the models and requires taking care of the problem of controlling its statistical regularity (i.e., avoiding over-parameterization). Here, we briefly introduce the statistical and computational strategies that have been used for accommodating variation across sites and branches. Then, we more specifically point out the solutions that have been used in the models developed here.

### 1.3.1 Variation across sites

Heterogeneity in substitution process across sites was first proposed by Yang (Yang 1993). In this model, which accounts for variation in the overall rate of substitution across sites (thus, there are slow-evolving and fast-evolving sites), heterogeneity is formalized as random effects across sites; each site has its rate. Site-specific rates are assumed to be *iid* (*Independent and Identically Distributed*) from a gamma distribution, whose shape parameter is estimated by maximum likelihood. The distribution of these rates, as random effects, across sites is estimated, and then the rate of each site is inferred given that distribution.

Since then, other aspects of substitution process have been modeled as site-specific effects; equilibrium frequencies (Lartillot and Philippe 2004), relative exchange rate (Pagel and Meade 2004) or amino acid fitness propensities in the context of mechanistic codon models (Rodrigue, Philippe et al. 2010). In most of these models, a finite or infinite mixture model has been used. However, parametric models, in which each site has a potentially distinct substitution process, have also been proposed. For example in (Lartillot and Philippe 2004), a distinct equilibrium frequency profile (over amino acids) is defined for each site, and all these frequency profiles are assumed to be *iid* from a Dirichlet distribution. The parameters of the Dirichlet distribution are

themselves endowed with a prior (and thus are estimated from the dataset). Apart from the fact that the random effects are multi-dimensional and are not explicitly integrated by discretization, but instead, are implicitly integrated by MCMC sampling, this model is formally very close to the gamma-distributed rate model of Yang (Yang 1993). This model-design strategy (based on a parametric model of site-specific random effects) will be used in the models introduced in this Ph.D. work as it allows joint inference of all the parameters to be performed simultaneously, which lets us implement modulations of amino acid fitness across coding sites.

### 1.3.2 Variation across branches

Regarding variation among branches (lineages), branch-specific substitution matrices (Yang and Roberts 1995) (or at least branch-specific equilibrium frequencies (Galtier and Gouy 1998)) can be defined. For large trees and small sequences, however, this approach can quickly result in a large number of parameters to be estimated with a relatively small amount of information available for each of them. Several approaches have been proposed to deal with this problem, which is fundamentally a problem of statistical regularization of the model. First, using a principal-component analysis, models have been developed in which the number of branch-specific parameters has been reduced (Groussin, Boussau et al. 2013). Alternatively, an approach based on partitioning branches into a small number of categories has been proposed (Foster 2004), where each category defines a substitution process. In these models, MCMC methods are used to average over possible partitions of branches by sampling from the posterior probability over all possible partition schemes. Finally, more complex models have been proposed invoking random point processes (Blanquart and Lartillot 2006; Blanquart and Lartillot 2008) or Brownian processes

(Lartillot and Poujol 2011) along the phylogeny. In contrast to the dimension-reduction or the partition approaches, these process-based approaches control the statistical regularity of the model, not by explicitly limiting the number of parameters, but instead, by inducing correlations among parameter values across neighboring branches, through a random process explicitly defined as a function of evolutionary time.

Here, in this research, our Differential Selection model (presented in chapter 2) assumes variation among branches. On the other hand, the branches are a priori assigned to a small number of categories, based on external information. Thus, the parameterization of the model is naturally controlled by the inherently small number of conditions.

## 1.4 Markov chain Monte Carlo

As mentioned before, Markov chain Monte Carlo (MCMC) sampling methods have been vastly used as Bayesian inference becomes extensively popular in complex phylogenetic studies. MCMC is a general computing method that generates the inference of Bayesian posterior probabilities, in a high dimensional space.

The Metropolis-Hastings algorithm (Metropolis, Rosenbluth et al. 1953; Hastings 1970) is the most general and the most employed sampling method in phylogenetic MCMC. However, for complex models, simple MCMC using Metropolis-Hastings (and each time, computing the full likelihood), is expensive. To achieve more efficient MCMC mixing, over the years, data-augmentation and parameter-expansion methods have been proposed (Mateiu and Rannala 2006; Lartillot and Poujol 2011). During data augmentation, a complete substitution history (or mapping)

is sampled for all sites and along the whole tree, conditional on the current value of the parameters. The parameters are then resampled by Metropolis-Hastings updates, given the current mapping. Finally, the mapping itself is resampled regularly given the current value of the model parameters. This MCMC strategy immensely simplifies the computation burden, and it is used in the implementation of the models presented in this Ph.D. project. More specifically, the much higher efficiency of the augmentation strategy is because the likelihood under the augmented state (i.e., the probability of the substitution mapping given the parameters of the model) can be expressed in terms of a relatively compact sufficient statistic.

In practice, the implementation of the models introduced here relied on a package developed by Lartillot and Poujol (Lartillot and Poujol 2011). In addition to relying on data augmentation and parameter expansion, this package uses a generic paradigm for representing models and their factorization, the so-called graphical model paradigm (Höhna, Landis et al. 2016). This paradigm allows us to handcraft hierarchical models by combining building blocks: each block (each node of the graph) represents a variable of the model, whose distribution is parameterized by the variables that are represented by its parent nodes in the graph. Generic routines are then proposed by the package to implement a combination of MCMC updates, given the structure of the graph.

## 1.5 Codon models

### 1.5.1 Motivation

Modeling the interplay between mutation and selection at the molecular level is one of the major goals in molecular evolutionary studies. Estimation of evolutionary patterns from homologous sequences is crucial for understanding the evolutionary processes like mutation rates, selective effects, or random drift.

In the case of protein-coding sequences, mutation processes can result in both synonymous (silent) substitutions, which leave the amino acid sequence unchanged, and non-synonymous substitutions, which result in an amino acid replacement. Interestingly, mutation processes are blind to the coding structure of the sequence, and therefore affect both synonymous and non-synonymous substitutions in the same manner. In contrast, selection differs markedly between the two types of substitutions. At least in some cases, it is reasonable to assume that there is virtually no selection on synonymous changes. If this hypothesis holds true, then one can use synonymous substitutions to measure the rate of evolution before the effect of natural selection: this rate is merely equal to the mutation rate. Non-synonymous substitutions, on the other hand, reflect the rate of evolution after selection has acted on the protein. Comparing the two types of substitutions, synonymous and non-synonymous, then gives an estimate of the relative strength of selection, independently of the mutation rate. Technically, this idea has been implemented using codon models.

## 1.5.2 Classical codon models

Goldman and Yang (Goldman and Yang 1994) and Muse and Gaut (Muse and Gaut 1994) independently introduced similar models of codon evolution for measuring the intensity of natural selection pressure acting on protein-coding sequences. These models typically estimate the overall rate of non-synonymous ($d_N$) and synonymous ($d_S$) substitutions (or their ratio, ω) across phylogeny, using an alignment of protein-coding sequences.

In the case of the Muse and Gaut (and Goldman and Yang) formalism, if we assume the point mutation from codon *a* to *b* as $\mu_{ab}$, the substitution rate matrix *Q* is as follows

$$\boldsymbol{Q_{ab}} = \begin{cases} \mu_{\boldsymbol{ab}} & \textbf{Synonymous} \\ \mu_{\boldsymbol{ab}} \times \boldsymbol{\omega} & \textbf{Non} - \textbf{synonymous} \\ \boldsymbol{0} & \boldsymbol{a} \textbf{ and } \boldsymbol{b} \textbf{ differ at more than one site} \end{cases} \quad (1\text{-}3)$$

The global parameter ω captures the average $d_N/d_S$ across the protein-coding sequence and along the whole phylogenetic tree and is employed as a measure of the strength and direction of selection pressure. When ω =1, natural selection is not affecting the substitution rate, and thus evolution is consistent with neutrality. When ω>1, non-synonymous mutations are more likely to reach fixation than synonymous mutations, suggesting that the non-synonymous substitutions have been beneficial to the organism and that selection acting on new non-synonymous variants is positive (on average). Conversely, ω<1 is an indication of negative, or purifying, selection.

In the first versions of these models, ω is a global parameter, capturing the average $d_N/d_S$ ratio over all sites and all branches. However, this model is not very powerful to detect positive selection, because not all sites in the sequence are under equal selective pressure. Therefore, some modifications of the original models allow ω to change among sites (Nielsen and Yang 1998; Anisimova, Bielawski et al. 2001), in the hope to detect positive selection concentrated over few sites of the protein. Statistically, these so-called *site models* are a particular case of the random effects models discussed above (see section 1.3.1). Positively selected sites are typically identified based on the (empirical Bayes) posterior probability of having $d_N/d_S$ greater than 1 (Yang, Wong et al. 2005). An alternative fixed effects approach has also been proposed, which was shown to give similar results, compared to the empirical Bayes approach (Kosakovsky Pond and Frost 2005).

Codon models also have been modified to allow for variation in selective pressure among lineages. In the first versions, these *branch-specific* models average the value of ω over all sites within a gene, thus allowing variation in $d_N/d_S$ only along branches (Yang and Nielsen 1998). More sophisticated versions were then developed, allowing for a combination of branch- and site-variation (*branch-site* models) (Yang and Nielsen 2002; Zhang, Nielsen et al. 2005). In practice, *branch-site* models aim is finding an episode of positive selection along pre-specified branches of the tree (foreground branches), affecting only a fraction of the sites. To do so, they assume that the remaining branches of the tree (background branches) are under a non-adaptive regime. They can then use these background branches to estimate the default strength of selection in the absence of positive selection.

An alternative approach to detect episodes of diversifying selection does not specify the foreground and background branches a priori.It instead, relies on a random effects likelihood (REL) strategy, which integrates over all possible selective patterns across branches for each site (Kosakovsky Pond, Murrell et al. 2011). This method was subsequently combined with a fixed effects strategy to account for variation across sites to detect episodic positive selection thus leading to a mixed effects model of evolution (MEME (Murrell, Wertheim et al. 2012)). The MEME model uses a mixture model with two classes of ω ratio associated to each branch in the phylogenetic tree. Although this approach has a good power to detect positive selection at individual sites, it is difficult to specify on which branch precisely locates the episodic selection.

*Branch-site* models have been used extensively in the past decade. They have been able to find some cases of putative episodes of adaptive evolution on some branches (Sawyer, Emerman et al. 2004; Kosiol, Vinař et al. 2008). For example, Kosiol *et al.* (Kosiol, Vinař et al. 2008) used a likelihood ratio test based on codon substitution models that allow for selective pressure variation across sites and along the branches of the tree, to identify molecular adaptation in mammalian genomes. First, using site models, they found a certain fraction (2.4%) of genes under positive selection globally over the tree for all mammals. Then, using *branch-site* models, they further identified some genes under episodic diversifying selection over specific branches. Less than 1% of genes showed positive selection on specific branches, showing that it is difficult to identify specific branches for specific sites. *Branch-site* models also potentially suffer from a high rate of false positives, caused by multinucleotide mutations (Venkat, Hahn et al. 2017).

### 1.5.3 Mechanistic models

All of the codon models discussed above try to detect an *acceleration* of non-synonymous rate, compared to the synonymous rate of substitution (taken as the neutral rate). Regarding the selective regime, such an acceleration of non-synonymous substitution rate is typically the result of ongoing adaptation (Mustonen and Lassig 2009), i.e., adaptation to a continually changing ecological environment or ongoing evolutionary Red-Queens. This means that the protein-coding sequences are evolving and adapting to a continually fluctuating fitness landscape. On the other hand, these codon models do not try to explicitly model the fitness landscape at the level of the protein sequence. Also, they do not explicitly formalize the detailed population-genetic mechanisms that are responsible for the intricate patterns of synonymous and non-synonymous substitutions observed in empirical sequences. These codon models also ignore the differences between different pairs of non-synonymous amino acid replacements resulting from point mutations.

Therefore, a viable alternative to these classical codon models is to derive a general mechanistic form of the codon substitution process based on first principles of population genetics, referred as mutation-selection codon models. The selection coefficient is defined as fitness landscape in these mutation-selection models. Many of the codon substitution models developed to date have the same fitness landscape for all positions. However, there are both theoretical and empirical motives to believe that selection acting on the level of protein-coding sequences is strongly site-specific, thus resulting in a marked differentiation of the stationary distribution across positions. The original motivations of the model presented by Halpern and Bruno (Halpern and

Bruno 1998), in 1998, were, in fact, to account for site specificities of amino acid fitness landscape within the mutation-selection modeling framework.

In the model proposed by Halpern and Bruno (Halpern and Bruno 1998), the substitution rate between codons is defined as the product of mutation rate and fixation probability. The fixation probability, itself, is dependent on scaled selection coefficient, $S^i_{a_1 a_2}$. This coefficient is associated with a mutant protein with the amino acid $a_2$ encoded by codon $c_2$, in a wild-type population where the amino acid $a_1$ encoded by codon $c_1$ is fixed at that position. Therefore, the instantaneous substitution rate matrix is as follows

$$\mathbf{Q}^{i}_{c_1 c_2} = \begin{cases} \mu_{c_1 c_2} & \textbf{Synonymous} \\ \mu_{c_1 c_2} \times \dfrac{S^i_{a_1 a_2}}{1 - e^{-S^i_{a_1 a_2}}} & \textbf{Non} - \textbf{Synonymous} \\ \mathbf{0} & \boldsymbol{c_1}\ \textbf{and}\ \boldsymbol{c_2}\ \textbf{differ at more than one site} \end{cases} \quad (1\text{-}4)$$

The mutation-selection model of Halpern and Bruno depends on the definition of the complete array of site-specific fitness vectors: for each site, a vector of 20 fitness parameters, for the 20 amino acids. This raises the question of how to empirically estimate the value of these fitness vectors. There are some strategies to be used in this direction:

- Re-parameterizing the process in terms of equilibrium frequencies. Such that, equilibrium frequencies at each site are identified with the observed frequencies at each column of the alignment (Halpern and Bruno 1998).

- Estimating the site-specific fitness profiles directly (along with the mutation rate matrix and the other parameters of the model) by maximum likelihood (Tamuri, Goldman et al. 2014).

However, both these models can cause a potential problem of extensive and over-parameterization (the so-called infinitely many parameters trap), in which any extra observed information added to the model would change the overall form of the model (Rodrigue 2013).

- A statistically more acceptable approach is to consider the site-specific fitness profiles as random effects across sites: thus, we need to define a law for those random effects and integrate them over this statistical law, using MCMC, like the approach of Rodrigue *et al.* (Rodrigue, Philippe et al. 2010) using a non-parametric mixture.

In the work presented here, which will be described in more details in chapter 2, we developed a model, similar to those of (Rodrigue, Philippe et al. 2010) and (Halpern and Bruno 1998), with the additional feature that site-specific fitness profiles are modulated across the tree. Like in (Rodrigue, Philippe et al. 2010), in these models, site-specific fitness profiles are modeled as random effects and are integrated by MCMC. On the other hand, instead of using a non-parametric mixture, we considered a parametric model in which, site-specific fitness profiles are assumed to be *iid* from a Dirichlet distribution. Although this parametric approach is less flexible and less general than the non-parametric mixture model used in (Rodrigue, Philippe et al. 2010), it is computationally faster, given the fact that the model itself is very complicated.

# CHAPTER 2: DETECTING CONSISTENT PATTERNS OF DIRECTIONAL ADAPTATION USING DIFFERENTIAL SELECTION CODON MODELS

The mechanistic codon models mentioned in the previous chapter (Halpern and Bruno 1998; Rodrigue, Philippe et al. 2010) represent an essential step in the modeling of the interplay between mutation, selection, and random drift. However, one fundamental assumption of these models is that the fitness landscape is constant through time. However, most biological systems do not evolve under constant selection pressure. Fitness effects change according to time and space, which in turn lead to phenotypic adaptations. A particularly situation where this occurs is when there are recurrent environmental changes, which may lead to repeated (or convergent) patterns of molecular adaptation. Capturing fluctuations in selection and, more specifically, consistent patterns of adaptation, is a challenge for models of sequence evolution.

In this chapter, we introduce a Differential Selection (DS) model which can capture differential fitness effects acting on coding sequences (section 2.5). This Differential Selection model is able to tease out the amino acids under differential selection pressure in different environments (here in HIV in different host-dependent HLA (Human Leukocyte Antigen) types).

The model has been developed in C++ framework, based on existing building blocks. The scripts for analyzing the data and results were programmed in Perl, Python and R environment.

In the following, we first give an introduction to HIV, its evolution and how it can be related to the host HLA type. Then, we present the DS model and its application to HIV, which was published in BMC Evolutionary Biology (Parto and Lartillot 2017).

## 2.1 HIV (Human Immunodeficiency Virus)

The Human Immunodeficiency Virus (HIV), is not only the causative agent of the deadly disease, AIDS (Acquired Immuno Deficiency Syndrome) but also one of the fastest evolving organisms (Sharp and Hahn 2010). Therefore, HIV has been of great interest for researchers in both medicine and evolutionary studies. In this chapter, the Differential Selection codon model developed during this Ph.D. is applied to HIV to find consistent patterns of its convergent evolution as a response to host immunity.

HIV is an enveloped, single-stranded RNA lentivirus, with 9.75 kb genome and 100-120 nm in diameter. Its genome consists of 9 genes *(gag, pol, env, tat, rev, nef, vif, vpr, and vpu),* shown in figure 2-1*,* which encode 19 proteins. Three genes, *gag*, *pol,* and *env*, contain the necessary information for the production of new viral particles. The remaining genes code for regulatory proteins that control the ability of the virus to infect cells (Kuiken C. 2008). The functions of the HIV-1 proteins are listed in table 2-I.

HIV has an extraordinary ability to adapt. This adaptation is due to its genetic diversity which is first the result of its fast replication cycle and large population size, and second its high mutation rate of $3\times10^{-5}$ per nucleotide base per cycle of replication, which is much larger than in eukaryotes (Jenkins, Rambaut et al. 2002). Recurrent recombination (Robertson, Hahn et al. 1995), genetic drift (Voronin, Holte et al. 2009) and natural selection driven by the immune system of the host even intensify this diversity. They produce additional mechanisms for viruses to share beneficial mutations between individuals in a population and even within one individual in a relatively short time (Shankarappa, Margolick et al. 1999; Rambaut, Posada et al. 2004). Large population permits

a higher rate of mutation, and its optimum is mostly dependent on the effects of beneficial mutations rather than deleterious ones (Jiang, Mu et al. 2010). These beneficial mutations have more chance to be fixed, while deleterious mutations are more commonly removed from the population (Elena, Wilke et al. 2007).

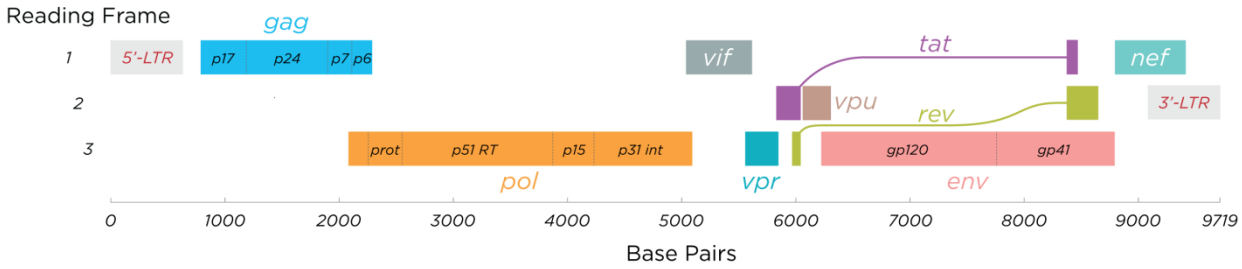



Figure 2-1. HIV-1 gene map.

There are two variants of HIV, which are slightly different in their genome structure; HIV-1 and HIV-2. HIV-1 diversity is further divided into three phylogenetic groups M, N and O. The M (major) group, which has a global (worldwide) distribution, has diverged in the 1930's (Korber, Muldoon et al. 2000) and from then has split into 9 subtypes A to K (also called clades). Estimating the age of the last common ancestor of the HIV-1 M group was done using data collected over two decades. Studies on the relation between HIV subtype and disease progression has shown that subtype D results in a faster progression of the disease (Baeten, Chohan et al. 2007; Ssemwanga,

Nsubuga et al. 2013), but host genetic background and immune response should also be considered as vital factors modulating disease progression.

Table 2-I. HIV-1 proteins and their functions (Rambaut, Posada et al. 2004).

| Proteins | Designation and size | Function |
|---|---|---|
| **Gag** | P24 | Capsid (CA), structural protein |
| | P17 | Matrix (MA) protein |
| | P7 | Nucleocapsid (NC), helps in reverse transcriptase |
| | P6 | Role in budding |
| **Pol** | P66, P51 | Reverse transcriptase (RT) |
| | P10 | Protease (PR), posttranslational processing of viral protein |
| | P31 | Integrase (IN), viral cDNA integration |
| **Env** | gp120 | Envelope surface protein |
| | gp41 | Envelope transmembrane protein |
| **Tat** | p14 | Transactivation |
| **Rev** | p19 | Regulation of viral mRNA expression |
| **Nef** | p27 | Pleiotropic, can increase or decrease virus replication |
| **Vif** | p23 | Increase virus infectivity and cell-to-cell transmission |
| **Vpr** | p15 | Helps in virus replication |
| **Vpu** | p16 | Helps in virus release |

## 2.2 HIV phylogeny and evolution

Phylogenetic studies have been critical for understanding the biology and evolution of HIV (Hillis 1999). Its vast diversity and high mutation rate have made HIV a good candidate for

phylogenetic studies, such as determining HIV origin (Korber, Muldoon et al. 2000), tracking routes of transmission (McCloskey, Liang et al. 2014; Wei, Xing et al. 2015), identifying drug resistance (Rios, Delgado et al. 2007; Brooks, Niznick et al. 2013) and developing effective vaccines (Novitsky, Smith et al. 2002). Moreover, phylogenetic approaches have been used to assess within- and among-host HIV diversity (Castro-Nallar, Pérez-Losada et al. 2012).

The fixation of mutations in the viral population is affected by natural selection, which is either positive (diversifying) or negative (purifying). In the absence of natural selection, the evolutionary history of the viral genome is only affected by mutation and random drift (Yang and Bielawski 2000). It is generally accepted that non-synonymous substitutions are more likely to make a biological change than synonymous ones and are subject to selection along with the corresponding viral trait. Although, selection (mostly purifying) on synonymous substitutions has been detected in HIV-1 genome (Mayrose, Stern et al. 2013; Zanini and Neher 2013). Under this paradigm and using phylogenies, several studies have tried to identify molecular determinants of natural selection in HIV at the molecular level (Templeton, Reichert et al. 2004; Pond, Frost et al. 2006).

Assuming homogenous selection pressure over all lineages of the phylogenetic tree is biologically unrealistic, since some selection forces may have occurred episodically during the evolution of the virus. For that reason, some branches may show stronger positive selection than other branches (i.e., selection is lineage-specific). This might happen when a virus cross-infects a new host population (from different species), or even switches to a new individual within one species. Positive selection is also likely to be mainly concentrated in small regions (gene-specific,

such as *env* genes with known deleterious mutations) or even in specific sites on the genome (site-specific).

A well-known example of positive selection on HIV genome is the selection pressure induced by host immune system, via the Human Leukocyte Antigens (HLA). The fact that the HLA is, in turn, evolving in response to infections results in an immunological arms race that is typical of host-pathogen interactions (Lam, Hon et al. 2010).

## 2.3 HLA and escape mutations

HIV mostly infects cells which bear the CD4 co-receptor ($CD4^+$). Early infection leads to acute high-level viremia. After 2-4 weeks, the immune response results in reduced intracellular and circulating HIV (viral set points) which is due to specific lysis of HIV-infected cells by $CD8^+$ cytotoxic-T lymphocytes (CTL). The elimination of infected cells is mediated by the presentation of virus epitopes by the HLA (Human Leukocyte Antigen) molecules on the surface of the cells and recognition of the epitopes by CTLs (McMichael and Rowland-Jones 2001). HLA genes are one of the most polymorphic regions in the human genome, which makes the HLA molecules able to present various epitopes, depending on the genetic background of individuals.

Importantly HIV is capable of evading the CTL response because of its rapid rate of mutation in or around these epitopes, some of which are selected to decrease antigen presentation and CTL recognition. This mechanism of HIV evolution is called *escape mutation* (Goulder and Watkins 2004). So both the process of antigen presentation to CTL as well as CTL escape are HLA-restricted (Carlson, Brumme et al. 2008). CTL escape mutations occur at sites within HLA-

restricted epitopes where an amino acid substitution may abolish epitope-HLA binding, reduce CTL recognition or generate antagonistic CTL response (McMichael and Rowland-Jones 2001).

The progression of the disease is not equal in all infected individuals. It has been observed that some patients are classified as fast progressors (Peretz, Alter et al. 2005), whereas it takes a long time to progress the disease in some HIV hosts (called long-term non-progressors), while some individuals do not show any infection at all (Levy 1993). This difference in the response to HIV infection appears to be due in part to the HLA genetic background of the host. For example, it has been observed that individuals carrying the HLA-B57, HLA-B27 and HLA-B5801 alleles are among the most resistant individuals against AIDS. On the other hand, HLA-B35 and HLA-B5802 background lead to much weaker resistance (Goulder, Bunce et al. 1996; Kaslow, Carrington et al. 1996; Carrington, Nelson et al. 1999; Matthews, Prendergast et al. 2008; Pereyra, Addo et al. 2008).

The molecular mechanism behind this heterogeneity among hosts, depending on their HLA genetic background, is directly related to the process of accumulation of escape mutations by the virus. More precisely, it is the result of a tradeoff between replication and immune escape, which can be more or less advantageous to the virus, depending on the specific HLA background (Boutwell, Rowley et al. 2009). More specifically, the spectrum of viral peptides presented by some HLA types is such that the only escape mutations available for the virus tend to result in a much-reduced rate of replication. In this context, some escape mutations are still selectively favored and reach fixation. However, the viral set point of the mutant virus is lower, leading to a slower progression of the disease (Chopera, Woodman et al. 2008). Furthermore, transmission of

less fit viruses into new hosts may lead to lower viral loads and a better clinical outcome in the new hosts (Chopera, Woodman et al. 2008; Goepfert, Lumm et al. 2008).

## 2.4 HLA-associated HIV evolution

For all reasons mentioned above, identifying correlations between HLA alleles and HIV escape mutations at the population level would be very useful, as it could help us understand the patterns of HIV adaptation to the host background. This, in turn, could have critical applications, in particular, the design of more efficient vaccines against AIDS.

Moore *et al.* (Moore, John et al. 2002) first demonstrated some evidence of association of HIV polymorphism with particular HLA alleles. They used a multivariate analysis with logistic regression for each residue to assess the association between the presence of polymorphism and the HLA alleles at the population level. However, they did not take HIV phylogeny into account. Their findings were confirmed in other studies using statistical methods that account for the phylogeny (Brumme, Brumme et al. 2007; Matthews, Prendergast et al. 2008). Mathews *et al.* (Matthews, Prendergast et al. 2008) identified the association between *gag*-specific polymorphism and HLA-B allele and between pol-specific polymorphism and HLA-C allele which leads to reduced viral load.

In 2009, Miura *et al.* (Miura, Brumme et al. 2009) showed that the replication rate of HIV in EC (Elite Controllers) is significantly reduced, associated with the distinct HLA alleles which provoke escape mutations in EC. They found that the proportion of HLA-associated escape mutations is high in EC but it is significantly lower than CP (Chronic Progressors).

Carlson *et al.* (Carlson, Brumme et al. 2008) developed a statistical model of HIV evolution, called a phylogenetic dependency network. They simultaneously accounted for epistatic effects of HLA alleles, codon covariation, and phylogeny to identify sites of HLA-mediates selection pressure. They provided a dependency network for the *gag* gene, which predicts that 41% of P17 and 20% of P24 codons are under selective pressure from at least one HLA allele.

All these developments, however, even those accounting for the phylogeny of the virus, have been conducted primarily from a statistical or machine-learning perspective. In part for that reason, none of them relies on an explicit model of the underlying *molecular evolutionary process*. The Differential Selection model introduced in the next chapter represents an attempt in this direction. It takes the form of a fully specified mutation-selection codon model for the virus, in which selection at the amino acid level at each coding site is modulated as a function of the HLA background. Applying this model to HIV sequences from hosts with known HLA background allows us to detect some positions that are under HLA-dependent selection.

## 2.5 Detecting consistent patterns of directional adaptation using differential selection codon models

### 2.5.1 Abstract

**Background:** Phylogenetic codon models are often used to characterize the selective regimes acting on protein-coding sequences. Recent methodological developments have led to models explicitly accounting for the interplay between mutation and selection, by modeling the

amino acid fitness landscape along the sequence. However, thus far, most of these models have assumed that the fitness landscape is constant over time. Fluctuations of the fitness landscape may often be random or depend on complex and unknown factors. However, some organisms may be subject to systematic changes in selective pressure, resulting in reproducible molecular adaptations across independent lineages subject to similar conditions.

**Results:** Here, we introduce a codon-based differential selection model, which aims to detect and quantify the fine-grained consistent patterns of adaptation at the protein-coding level, as a function of external conditions experienced by the organism under investigation. The model parameterizes the global mutational pressure, as well as the site- and condition-specific amino acid selective preferences. This phylogenetic model is implemented in a Bayesian MCMC framework. After validation with simulations, we applied our method to a dataset of HIV sequences from patients with known HLA genetic background. Our differential selection model detects and characterizes differentially selected coding positions specifically associated with two different HLA alleles.

**Conclusion:** Our differential selection model is able to identify consistent molecular adaptations as a function of repeated changes in the environment of the organism. This model can be applied to many other problems, ranging from viral adaptation to the evolution of life-history strategies in plants or animals.

***Keywords:*** *HIV, evolution, selection, HLA, virus adaptation, Bayesian, MCMC*

## 2.5.2 Background

Statistical models of molecular evolutionary processes are now widely used to analyze the interplay between mutation and selection. Often, these models are formulated at the codon level, thus relying on the contrast between synonymous and non-synonymous substitutions to leverage out an estimation of the strength of selection acting at various levels (nucleotide, amino acids, codon usage) of protein-coding sequences.

The first codon models, proposed independently by Goldman and Yang (Goldman and Yang 1994) and Muse and Gaut (Muse and Gaut 1994), relied on a simple aggregate parameter, $\omega = d_N/d_S$, to capture the overall strength of selection, globally over the protein-coding sequence and over the phylogenetic tree. Subsequent elaborations on these original models allowed for variation in $d_N/d_S$ among sites (Nielsen and Yang 1998; Anisimova, Bielawski et al. 2001) or among lineages (Yang 1998), or both (Yang, Wong et al. 2005; Zhang, Nielsen et al. 2005). This increases the sensitivity and the resolution of the detection of selective regimes. However, all of these models still do not discriminate between alternative amino acids. Instead, they essentially put all non-synonymous substitutions on the same level (Rodrigue, Philippe et al. 2010).

In this direction, Halpern and Bruno (Halpern and Bruno 1998) and also Thorne *et al.* (Thorne, Choi et al. 2007) have proposed an alternative codon modeling strategy, allowing for site- and amino acid-specific selective effects. The model of Halpern and Bruno also has a clear mechanistic interpretation, being derived from first principles of population genetics. Specifically, the rate of substitution between codons is seen as the product of the mutation rate and the fixation probability. In turn, the fixation probability is made explicitly dependent on the selection coefficient of the mutation under consideration. Selection coefficients are obtained from an explicit fitness

landscape, in which the fitness of each amino acid is allowed to be different at each coding site. Technically, the model, therefore, invokes, at each coding site, a normalized vector of 20 amino acid fitness coefficients, collectively referred to as the site-specific fitness profile. In the original version of Halpern and Bruno (Halpern and Bruno 1998), site-specific amino acid fitness profiles were empirically estimated based on observed amino acid frequencies. Since then, a statistically more sophisticated version of this model was developed in a Bayesian framework by Rodrigue *et al.* (Rodrigue, Philippe et al. 2010). They used a non-parametric approach to integrate over the uncertainty about site-specific selective features (now seen as random effects across sites), and to capture the unknown law of amino acid fitness profiles across sites. The importance of accounting for modulation of selection across sites by introducing site-specific amino acid fitness profiles was demonstrated by Bayes factor computation and posterior-predictive tests (Rodrigue, Philippe et al. 2010). Of note, more phenomenological variants of this modeling approach, also with site-specific amino acid fitness contributions but without the population-genetic justification of Halpern and Bruno's paradigm, have been explored (Robinson, Jones et al. 2003; Rodrigue, Philippe et al. 2010).

This modeling approach, although reasonably complex, still leaves an important aspect of protein evolution aside, by assuming that the fitness landscape is constant through time. Yet, many ecological situations suggest that fitness landscapes undergo substantial fluctuations through time (Mustonen and Lässig 2008). Two alternative approaches are possible, to relax this specific assumption. First, fluctuations of the fitness landscape could be modeled as a purely latent effect (e.g., Markov-modulated models) (Gascuel and Guindon 2007), thus without relying on any extra information about the environmental or ecological drivers of the fluctuations. Secondly, in some situations, empirical knowledge is available, regarding varying conditions across sampled genetic

sequences. In this context, it is, in principle, possible to explicitly model condition-specific amino acid fitness modulations. The present work is an attempt at modeling such effects.

A clear-cut example where robust empirical knowledge about varying selective environments is available is the evolution of viral sequences as a function of the genetic background represented by the hosts. For example, the analysis of patterns of selection, using $d_N/d_S$ codon models in a phylogenetic maximum likelihood framework, has shown the substantial role of fluctuating selection in the emergence of new mutations and the ability of HIV-1 to escape from immune system (Nielsen and Yang 1998; Edwards, Holmes et al. 2006; Salemi, Burkhardt et al. 2007). HIV-1 is capable of evading the CTL (Cytotoxic T-Lymphocyte) response because of its rapid rate of mutation in HLA-restricted epitopes, called escape mutation. Escape mutation gives the virus the ability to adapt to different selective forces in different individuals and in response to drugs (Schweighardt, Wrin et al. 2010), which makes the design of a vaccine very difficult.

Therefore, understanding the evolution of HIV-1 within the human body, which is both rapid and under strong selection, helps designing more effective vaccines against HIV-1 and control its evolution. On the other hand, the high rate of mutation of HIV-1 enables the virus to produce a genetically diverse population in each host, called quasi-species (Carlson, Brumme et al. 2008), which makes it possible for the virus to adapt to its host even within a single round of infection. In this direction, the correlation between HLA alleles and HIV polymorphisms has been paid a lot of attention in recent years, from population-based studies (Moore, John et al. 2002; Altfeld and Allen 2006; Carlson and Brumme 2008) to studies taking phylogeny into account (Brumme, Tao et al. 2008; Rousseau, Daniels et al. 2008). A method, called the Phylogeny Dependency Network, was introduced to analyze HLA-mediated escape in HIV-1 (Carlson, Brumme et al. 2008). This method accounts for the phylogeny, the correlation between coding sites and linkage

disequilibrium between HLA alleles. On the other hand, it only takes the information of the tips of the phylogenetic tree into account. More fundamentally, it does not rely on an explicit model of the underlying molecular evolutionary processes. Another phylogenetic model has been used by (Tamuri, dos Reis et al. 2009) *et al.* to identify host dependent selective constraints for viruses. These authors specified different host-dependent substitution rates along the phylogenetic tree and used a maximum likelihood approach, combined with a likelihood-ratio test, to identify positions under differential selection between hosts. This method, first formulated directly at the amino acid level, was then generalized to account for the coding structure (Tamuri, Goldman et al. 2014).

Here, we introduce a codon model able to capture site- and condition-specific amino acid fitness effects. In this differential selection (DS) model, which is implemented in a Bayesian inference framework, a site and branch heterogeneous selection factor is invoked to estimate the substitution rate at the codon level of aligned HIV-1 sequence. As the population-genetics of viral populations is complex and challenging to model quantitatively, we explored two alternative strategies for deriving the codon substitution process, either using a phenomenological approach or using a mechanistic derivation as in Halpern and Bruno (Halpern and Bruno 1998). Our DS model was then used to investigate how the fluctuating environment provided by the diversity of human HLA background affects HIV-1 sequence evolution. We illustrate how our approach finds consistent patterns of viral adaptation, in terms of how selection acts at specific positions, modulating amino acid preference as a function of the HLA background.

## 2.5.3 Materials and Methods

### HIV-1 data

A dataset of 333 *Gag* sequences (443 codons) of HIV-1 subtype B from 41 HIV-infected individuals with known HLA types were obtained from the Los Alamos National Laboratory (LANL) HIV database [111]. Each patient is represented by eight sequences on average. We also downloaded the information about the HLA types of the patients. About 35% of the sequences are from HLA B57+ patients. Recombinant sequences were excluded from the study by choosing an internal option in the LANL HIV databases to remove all known CRFs (Circulating Recombinant Forms). The amino acid alignment of the sequences provided by the source was downloaded, manually corrected (misplaced amino acids were relocated, and misaligned regions were deleted) and used for back aligning the DNA sequences at the codon level. The dataset is provided in Appendix A.

## Phylogenetic tree estimation

Primarily for computational reasons, the method introduced here assumes a fixed tree topology. However, owing to the relatively short length of the coding sequences, the tree topology may not be known with high confidence. In addition, there is the question of whether the sequences corresponding to a given patient should form a monophyletic group. This may not always be the case, in particular, because of tree reconstruction errors, a problem which can be alleviated simply by constraining the monophyly of each patient during the tree reconstruction. However, non-monophyly could also be real, being caused by complicated multiple infection patterns between individuals. In this case, constraining the monophyly might result in misspecification of the reconstructed tree topology.

To check the robustness of our method to these potential sources of error, we tested alternative methods for reconstructing the phylogenetic tree and conducted independent analyses under these alternative tree topologies. Specifically, a first tree topology (T1) was obtained directly from the LANL website. This tree was estimated using the neighbor-joining algorithm (Saitou and Nei 1987). A second tree (T2) was reconstructed using MrBayes (version 3.2.6) (Huelsenbeck and Ronquist 2001; Ronquist, Teslenko et al. 2012), under the GTR+Γ substitution model and constraining the monophyly of the groups corresponding to sequences belonging to a given patient. A third tree (T3) was estimated, still using MrBayes, under the same substitution model, but without imposing any constraint on the tree topology. In MrBayes, we ran MCMC chains for 1500000 cycles. The average standard deviation of split frequencies reaches the value less than 0.05, and the Potential Scale Reduction Factor (PSRF) (Gelman and Rubin 1992), which should approach 1.0 as the two runs converge, was equal to 1.001 and 1.000 for the two chains.

In the case of tree T1 and T3, we observed 20 and 23 cases of non-monophyletic patients, respectively. In both cases, we applied a greedy algorithm for excluding the smallest possible set of sequences such that each patient is then represented by a monophyletic group of sequences. This was done using the following recursive procedure: first, the number of sequences from each host pending from each node was determined recursively at each node, from the tips toward the root. During this recursive scan, wherever a group pending from a given node was not monophyletic, the sequences belonging to the host with the smallest number of sequences pending from that node were flagged. Finally, in a backward recursive scan of the tree, from root to tips, the flagged sequences were removed from the dataset. Application of this method leads to the elimination of 20 and 23 out of 333 sequences in the cases of tree T1 and T3. Altogether, T1, T2, and T3 have respectively 313, 333 and 310 tips (sequences). The RF (Robinson-Foulds) distance

(Robinson and Foulds 1981) of these tree topologies is shown in table 2-II. The Newick format of all phylogenetic trees, which were used in downstream analyses, is given in Appendix B.

Finally, for the three topologies, the branches of the phylogenetic tree were divided into four conditions according to the host HLA types.

Table 2-II. RF (Robinson-Foulds) distances between tree T1, T2, and T3. RF is calculated using (Boc, Diallo et al. 2012).

| | **T1** | **T2** | **T3** |
|---|---|---|---|
| **T1** | 0 | 233 | 220 |
| **T2** | 233 | 0 | 7 |
| **T3** | 220 | 7 | 0 |

## Model

**Notations** – We consider a coding sequence of length $N$ ($N$ being the number of coding positions, or equivalently $3N$ is the number of nucleotide sites). The number of conditions (e.g., HLA types) is defined by $K$. All the indices used in this paper conform to the following conventions:

- Codon positions (sites) $i \in [1, N]$
- Conditions $k \in [1, K]$
- Codon states $c \in [1, 61]$
- Nucleotide states $n \in [1,4]$
- Amino acid states $a \in [1, 20]$

## Model of codon substitution

The rate of evolution by point substitution is the result of a complex interplay between mutation, selection, and random drift. Drawing inspiration from previous developments in statistical molecular evolution (Goldman and Yang 1994; Muse and Gaut 1994; Halpern and Bruno 1998; Robinson, Jones et al. 2003; Rodrigue, Philippe et al. 2010), we modeled this process at the codon level, as a multiplicative combination of mutation rates and selective effects (the latter implicitly including the contribution from random drift).

The mutation process is assumed to be homogenous over time and along the sequence. It is modeled as a Markovian general time-reversible process, parameterized regarding the relative exchange rates ($\rho$) between nucleotides and the stationary probability (equilibrium frequency) of the target nucleotide ($\pi$). Thus, the rate of substitution from nucleotide $n_1$ to nucleotide $n_2$ is equal to:

$$\boldsymbol{Q_{n_1 n_2}} = \frac{\mathbf{1}}{\boldsymbol{Z}} \boldsymbol{\rho_{n_1 n_2}} \, \boldsymbol{\pi_{n_2}} \qquad (2\text{-}1)$$

Where Z is the normalization factor:

$$\boldsymbol{Z = \sum_{n_1}^{n_2} \rho_{n_1 n_2} \pi_{n_2}} \qquad (2\text{-}2)$$

The set of relative exchangeabilities between nucleotides is constrained to be symmetric:

$$\boldsymbol{\rho_{n_1 n_2} = \rho_{n_2 n_1}} \qquad \boldsymbol{for\ all\ n_1, n_2 = 1..4} \qquad (2\text{-}3)$$

In addition, it is normalized:

$$\boldsymbol{\sum_{n_1}^{n_2} \rho_{n_1 n_2} = 1} \qquad (2\text{-}4)$$

The vector $\pi$ of equilibrium frequencies is also with the constraint

$$\boldsymbol{\sum_{n} \pi_{n} = 1} \qquad (2\text{-}5)$$

The selective forces, on the other hand, are both condition- and position-specific. The modulations across conditions and positions are mediated exclusively by the encoded amino acid sequence. Accordingly, for each position *i* and each condition *k*, we introduce an array of 20 non-negative fitness factors, $F^{ik} = (F_a^{ik})_{a\in[1,20]}$, one for each amino acid. In the following, these 20-dimensional vectors will be referred to as amino acid *fitness profiles*. Thus, we have distinct fitness profiles across positions, and for a given position, the fitness profile over the 20 amino acids is further modulated across conditions. How these fitness profiles are defined in practice is explained in more detail below (section; Definition of the amino acid selective effects).

Given a mutation matrix and a set of amino acid fitness profiles, we considered two alternative approaches for expressing substitution rates between codons as a function of the fitness of the

amino acids. The first is a phenomenological approach, while the second is more mechanistic in its inspiration.

## Phenomenological model (M1)

The phenomenological model is similar, in its general form, to the models explored by Rodrigue *et al.* (Rodrigue, Philippe et al. 2010), or, in a slightly different parameterization, to the models considered in Robinson *et al.* (Robinson, Jones et al. 2003). Specifically, consider a given position *i* along the sequence, and a given condition *k* along the tree. Consider also two codons, $c_1$ and $c_2$, differing only at one position and with nucleotides $n_1$ and $n_2$ at that position. These two codons encode for amino acids $a_1$ to $a_2$, respectively. Then, the rate of substitution between these two codons is given by:

$$\boldsymbol{R}_{\boldsymbol{c_1 c_2}}^{\boldsymbol{ik}} = \boldsymbol{Q}_{\boldsymbol{n_1 n_2}} \times \sqrt{\frac{\boldsymbol{F}_{\boldsymbol{a_2}}^{\boldsymbol{ik}}}{\boldsymbol{F}_{\boldsymbol{a_1}}^{\boldsymbol{ik}}}} \qquad (2\text{-}6)$$

Thus, according to this model, the rate of substitution is proportional to the mutation rate, while being influenced by the selection operating at the amino acid level, through the fitness factors $F_a^{ik}$: the substitution rate is higher (resp. lower) than the neutral substitution rate if the fitness of the final amino acid is greater (resp. smaller) than the fitness of the initial amino acid. Note that, if the two codons are synonymous, i.e., if $a_1=a_2$, then the substitution rate is merely equal to the mutation rate defined by the nucleotide transition matrix *Q*. Finally, the model considers only point substitutions, and therefore, the substitution rate is assumed to be equal to zero between codons differing at more than one nucleotide position. Thus, all together:

$$R^{ik}_{c_1c_2} = \begin{cases} Q_{n_1n_2} & \textbf{Synonymous} \\ Q_{n_1n_2} \times \sqrt{\frac{F^{ik}_{a_2}}{F^{ik}_{a_1}}} & \textbf{Non} - \textbf{synonymous} \\ 0 & \boldsymbol{c_1} \textbf{ and } \boldsymbol{c_2} \textbf{ differ at more than one site} \end{cases} \quad (2\text{-}7)$$

This formulation ensures that the average number of synonymous substitutions per unit length is equal to 1. Here, the selection factor modulates the rate of non-synonymous substitution.

## Mechanistic model (M2)

The second approach is inspired by a mechanistic argument based on first principles of population genetics, as initially suggested by Halpern and Bruno (Halpern and Bruno 1998). Consider again the substitution rate between codon $c_1$ to $c_2$ at site $i$ and condition $k$. First, we define a scaled selection coefficient (scaled by effective population size $N_e$), associated with codon $c_2$, seen as a mutant in the context of a population in which the wild-type allele is $c_1$. This scaled selection coefficient is given by:

$$S^{ik}_{a_1a_2} = ln\left(\frac{F^{ik}_{a_2}}{F^{ik}_{a_2}}\right) \quad (2\text{-}8)$$

Then, the rate of substitution between codon $c_1$ and $c_2$ is given by the product of the mutation rate and the relative fixation probability $P$ (i.e., relative to neutral). This fixation probability is itself dependent on the scaled selection coefficient. Using the classical diffusion approximation, this relative fixation probability can be expressed as:

$$\boldsymbol{P_{fix} = \frac{S^{ik}_{a_1 a_2}}{1 - e^{-S^{ik}_{a_1 a_2}}}} \qquad (2\text{-}9)$$

So that the rate of substitution between codons is given by

$$\boldsymbol{R^{ik}_{c_1 c_2} = \begin{cases} Q_{n_1 n_2} & \textbf{Synonymous} \\ Q_{n_1 n_2} \times \frac{S^{ik}_{a_1 a_2}}{1 - e^{-S^{ik}_{a_1 a_2}}} & \textbf{Non - synonymous} \\ 0 & \boldsymbol{c_1} \textbf{ and } \boldsymbol{c_2} \textbf{ differ at more than one site} \end{cases}} \qquad (2\text{-}10)$$

Again, we see that the rate of substitution is higher (resp. lower) than the neutral substitution rate if the non-synonymous mutation leads to an increase (resp. a decrease) in the fitness of the sequence.

## Definition of the amino acid selective effects

In principle, the amino acid fitness profiles associated with each site and each condition $F_a^{ik}$, could be considered as independent arrays, both across sites and across conditions. However, most of the amino acid conservation (due to purifying selection) observed along the sequence is in fact condition-independent. Against this globally invariable fitness background, the modulations of the fitness landscape induced by condition-dependent effects (such as the HLA type of the host) are likely to be comparatively small. In this context, considering amino acid selective effects as entirely independent random effects across conditions would imply that the invariable background

would be re-estimated independently for each condition, potentially resulting in a loss of statistical power. Therefore, as a more powerful alternative, we explicitly defined an amino acid selection in terms of a log-additive superposition of a global background and condition-dependent differential selective effects, as follows. First, a baseline or global fitness profile is defined for each position. That is, for position *i*, we define a 20-dimensional vector ($G_a^i$), for *a=1...20*. This vector is drawn from a uniform Dirichlet distribution independently at each site. This baseline defines the fitness landscape under condition 0, which is therefore taken as our reference condition (black branches in figure 2-2).

Next, the selection is modulated across conditions using condition-specific differential selection profiles. Thus, for position *i* in condition *k*, we define a 20-dimensional vector ($D_a^{ik}$), for *a=1...20*. Unlike the baseline profiles, which are positive (and sum to 1), those differential selection effects can be positive or negative. A positive (resp. negative) coefficient means that the fitness of the corresponding amino acid is increased (resp. decreased) in the target condition, compared to the reference condition. The differential selection profiles are drawn *iid* from a Normal distribution of mean 0 and condition-specific variance $\sigma_k^2$.

Altogether, the condition-specific fitness profiles are constructed as follows:

$$\begin{aligned} \boldsymbol{F_a^{i0} = G_a^i} \\ \boldsymbol{F_a^{i1} = G_a^i e^{D_a^{i1}}} \\ \boldsymbol{F_a^{ik} = G_a^i e^{(D_a^{i1} + D_a^{ik})}} \\ \boldsymbol{k = 2 \ldots k} \end{aligned} \qquad (2\text{-}11)$$

Note that we have used a two-level system for introducing the differential effects (i.e., a different equation for *k=1* and *k>1*). This is motivated by the fact that we need to discriminate

both among branches that are between hosts and within the same host, and among hosts with differing HLA backgrounds. Thus, it reflects the differential between within-host ($D_a^{i1}$) and between-host ($G_a^i$) selection regions, while representing specific selective features more associated explicitly with differing HLA backgrounds $(D_a^{ik})_{k=2\ldots K}$. In the case of HIV-1, we consider two focal HLA backgrounds (B57+ and B35+), against a default B57-/B35- background. Thus, we define a total of four different conditions (*K=4*), and the branches of the tree are partitioned according to four different selection regimes (figure 2-2): first, we distinguished between the branches connecting the host-specific groups of sequences (between-patient condition) and the branches within each host-specific group of sequences (within-patient condition). Among the latter set of branches, we further distinguished among patients according to their HLA-type: either between HLA-B57+ and HLA-B57- patients, or between HLA-B35+ and HLA-B35- patients. The HLA-B57 type is known to be associated with the control of viremia (Migueles, Sabbaghian et al. 2000; Altfeld, Addo et al. 2003) whereas HLA-B35 is known as the HLA related to the fast progression of the disease (Itescu, Mathur-Wagh et al. 1992; Flores-Villanueva, Hendel et al. 2003).

An important point should be emphasized concerning the statistical formalization of the fitness landscape and its modulations across sites and conditions. Conceptually, the arrays of global and condition-specific fitness effects should be considered, not as parameters, but as random effects across sites, which are integrated over a distribution (respectively, a Dirichlet and a Normal distribution for the global and differential effects). This integration is done implicitly, through the MCMC sampling (see below). As a result, the aim of the model introduced here is not to achieve accurate and asymptotically consistent point estimation of site- and condition-specific fitness effects: in most cases, the information for inferring such fitness effects will be limited. Instead, it

is to draw an inference based on the complete posterior distribution. A more specific objective is to single out those relatively few cases for which there is sufficient information to infer, with high posterior probability, the presence of a differential selective effect between two conditions. One crucial property of this type of inference is to allow for a reasonable control of the fraction of false discoveries among those cases that are selected based on a high posterior probability of a differential effect. This is something which is investigated through posterior predictive simulations (see below).

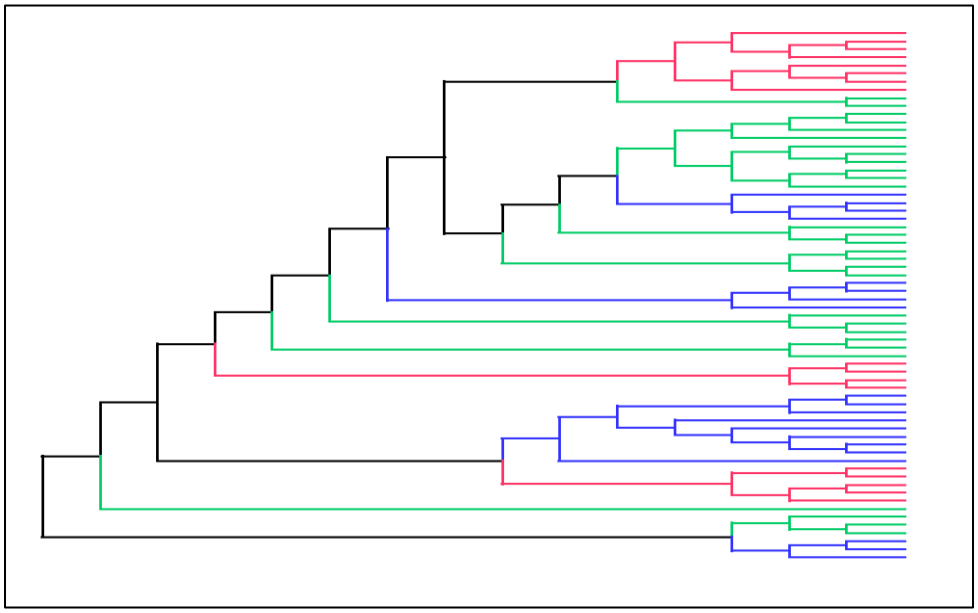


Figure 2-2. Illustrative phylogenetic tree of HIV-1 *Gag* sequences. Different colors along the tree show different selection regimes for the corresponding sequences. Black branches for between-patients, green for within-patients, red and blue for HLA-B57 and HLA-B35 categories, respectively.

## Priors

The topology ($\tau$) of the tree is fixed. The parameters of the model consist of branch lengths, $l_j$ ($1<j<2N-3$ where $N$ is the number of sequences), nucleotide exchangeabilities $\rho$ and nucleotide equilibrium frequencies $\pi$. The priors that we used are as follows: on branch lengths: a product of independent exponentials of mean $\lambda$; the hyperparameter $\lambda$ is from an exponential distribution of mean 0.1; on relative exchangeability rate: a product of exponentials of mean 1; on mutational equilibrium frequency: a uniform Dirichlet distribution. As mentioned above, the site-specific fitness profiles ($G$) and differential fitness effects ($D$) are random effects, integrated over Dirichlet and normal distributions, respectively.

## MCMC

We used Markov chain Monte Carlo (MCMC) to sample the parameters of the model from their joint posterior distribution. We used a graphical model environment previously introduced in (Lartillot 2006), heavily relying on data augmentation and parameter expansions methods, such as described in particular in (Lartillot and Poujol 2011). Briefly, the MCMC cycle consists of an alternation between two steps: first, a detailed substitution history at each coding site is Gibbs-sampled, from the posterior distribution conditional on the current parameter configuration. Second, conditional on these augmented data, the parameters and the random effects across sites

are updated through an extensive series of Metropolis-Hastings moves, cycling over all parameters or random variables of the model.

For the nucleotide equilibrium frequencies $\pi$ and the global fitness profiles *G,* which are under the constraint that they should sum to 1, we used constrained move as explained in (Lartillot 2006). Branch lengths $l$ and exchangeabilities $\rho$, which are positive real numbers, were updated using multiplicative moves (Lartillot 2006). Convergence of several key parameters and sufficient statistics was monitored first by plotting their summary statistics as a function of number of iterations (points) for two independent runs, and second by using the *tracecomp* program (from the Phylobayes suite (Lartillot, Lepage et al. 2009)) to compare the samples obtained under independent runs. *Tracecomp* gives an estimate of the discrepancy between the two runs, as well as the effective sample size, for several key parameters and statistics of interest. In the present case, the minimum effective size was higher than 300 and the discrepancy less than 0.2 for most statistics. After exclusion of the burn-in, posterior estimates were estimated by averaging over the remaining of the MCMC chain (approximately 1500 points for the empirical analyses, 1000 points for the simulations). As an additional control of the reproducibility of the MCMC analysis, we also checked that the posterior mean differential selection factors for all amino acids at all sites, as well as the associated posterior probabilities of a positive effect, were consistent between two independent runs (correlation coefficient $R^2 > 83\%$ in all cases, see Appendix C and D).

## Simulations

Simulations were conducted using a modified version of the posterior predictive formalism (Rubin 1984; Gelman, Meng et al. 1996). In all cases, parameter configurations were drawn from the posterior distribution under the 4-condition model fitted on the HIV dataset. Then, in the first

series of simulations, the differential selection effects across differential conditions were set to 0, while the global selection profiles were left unchanged, thus giving empirically calibrated simulation replicates under the null hypothesis of no differential effect across conditions. These simulations were conducted to estimate the rate of false positives.

In the second series of simulations, we implemented a sparse distribution of differential selection effects across sites, with various fractions (f = 0.5, 0.1 and 0.05) of sites with non-zero effects. Sites with non-zero effects were chosen uniformly at random, independently for conditions 2 (HLA B57+) and 3 (HLA B35+) and were endowed with differential condition effects independently drawn from a reflected gamma distribution of mean 1 and shape parameter 2. This second series of simulations were conducted to evaluate the precision and sensitivity of the method. In both cases, the phenomenological (M1) and the mechanistic (M2) models were investigated, and simulations were conducted based on ten parameter configurations sampled from the posterior distribution (10 points regularly spaced from the MCMC run), yielding a total of 10 replicates per condition.

For all simulations, the full model (with $K=4$ conditions) was then applied to these simulated data. For a given pair of condition (e.g., HLAB57+ versus HLAB57-), and for several $\alpha$ levels, the number of positions inferred to be under differential selection with posterior probability greater than $1-\alpha$ was determined. In the context of the first series of simulations (no differential selection simulated), dividing this number by the total number of positions times the number of amino acids gives the rate of false positives, which was tabulated for several values of $\alpha$. For the second series of simulations (with differential selection simulated), the discoveries made at a given threshold were compared with the true differential selection values, and the precision (fraction of true discoveries over all discoveries) and the sensitivity (fraction of true discoveries over all

differentially selected sites) were determined as a function of the significance threshold. A discovery is deemed true if the true differential selection effect is non-zero and of the same sign as the inferred differential selection effect.

## 2.5.4 Results

### Simulation analyses

The properties of the model were first investigated through simulations. Since the primary application of the model introduced here is to identify positions for which specific amino acids are under differential condition-dependent selection pressure, the simulation analyses were more specifically designed to evaluate the rate of false positives of the method, as well as its precision and sensitivity. In order to ensure that the conclusions of the simulations are relevant to the empirical situations considered here, simulations were calibrated against parameter estimates obtained from the empirical analyses on the HIV dataset. This was done using a modified version of the posterior predictive formalism (Rubin 1984; Gelman, Meng et al. 1996).

The first series of 10 replicates were produced under the null model assuming no differential selection effect across conditions — thus, considering a constant fitness landscape over the whole phylogenetic tree. The model with $K=4$ conditions was then applied to these simulated data. For a given pair of condition (e.g., HLAB57+ versus HLAB57-), and for different α levels, the number of positions inferred to be under differential selection with posterior probability greater than $1-\alpha$ was determined, giving us an estimate of the false positive rate as a function of the stringency of the selection. As can be seen from table 2-III, for reasonable posterior probability thresholds, the rate of false positive is low, reaching 5% for $1-\alpha = 0.6$, and lower than 1% for $1-\alpha > 0.8$.

This simulation experiment illustrates a point about the Bayesian approach used here: using Normal distribution centered on 0 enforces shrinkage of the differential fitness effects across positions towards 0 (i.e., the model is centered on the null hypothesis representing an absence of selective difference between conditions). One critical consequence of this choice is that, in the absence of a sufficiently strong empirical signal able to counteract this prior, the method will typically not infer high posterior probability support for differential selective effects. Note that these simulations, which have been calibrated against the empirical dataset of interest, can also be used to obtain a rough estimate of the fraction of false discoveries, by comparing, for a given threshold, the total number of discoveries (d) on the real dataset with the mean number of false positives (d0) under the simulations. An estimate of the fraction of false discoveries is then given by d0/d (see below).

Table 2-III. False Positive Rates (FPR) for different conditions as a function of the posterior probability thresholds, under model M1 and M2.

| | **M1** | | **M2** | |
|---|---|---|---|---|
| **condition 1 (within-patients)** | | | | |
| **threshold** | **mean number of FP** | **FPR** | **mean number of FP** | **FPR** |
| >0.55 | 1843.7 | 20.8 | 1845.8 | 20.8 |
| >0.60 | 1112.7 | 12.6 | 1166.8 | 13.2 |
| >0.65 | 684.9 | 7.7 | 737.8 | 8.3 |
| >0.70 | 316.9 | 3.6 | 334.5 | 3.8 |
| >0.75 | 173.1 | 2.0 | 181.8 | 2.1 |
| >0.80 | 81.8 | 0.9 | 86.5 | 1.0 |
| >0.85 | 26.0 | 0.3 | 31.7 | 0.4 |
| >0.90 | 7.1 | 0.1 | 4.6 | 0.1 |

| | | | | |
|---|---|---|---|---|
| >0.95 | 0.75 | 0.01 | 0.1 | 0.0 |
| **condition 2 (HLA-B57+)** | | | | |
| **threshold** | **mean number of FP** | **FPR** | **mean number of FP** | **FPR** |
| >0.55 | 1004.1 | 11.3 | 957 | 10.8 |
| >0.60 | 456.3 | 5.15 | 471 | 5.3 |
| >0.65 | 229 | 2.58 | 237.1 | 2.7 |
| >0.70 | 88.9 | 1 | 78.7 | 0.9 |
| >0.75 | 31.3 | 0.3 | 27.3 | 0.3 |
| >0.80 | 12.9 | 0.15 | 8.4 | 0.1 |
| >0.85 | 3.8 | 0.04 | 1.6 | 0.02 |
| >0.90 | 0.4 | 0 | 0.05 | 0 |
| >0.95 | 0 | 0 | 0 | 0 |
| **condition 3 (HLA-B35+)** | | | | |
| **threshold** | **mean number of FP** | **FPR** | **mean number of FP** | **FPR** |
| >0.55 | 1245.1 | 14 | 1226.5 | 13.8 |
| >0.60 | 632.4 | 7.1 | 683 | 7.7 |
| >0.65 | 345.6 | 3.9 | 385.3 | 4.3 |
| >0.70 | 141.4 | 1.6 | 148.1 | 1.7 |
| >0.75 | 58.6 | 0.7 | 64.4 | 0.73 |
| >0.80 | 25.3 | 0.3 | 23.1 | 0.3 |
| >0.85 | 7.7 | 0.1 | 6.5 | 0.07 |
| >0.90 | 1.2 | 0.01 | 0.9 | 0.01 |
| >0.95 | 0 | 0 | 0 | 0 |

The second series of simulations were conducted, assuming the presence of modulations of the fitness landscape across conditions, with various fractions of sites under non-zero differential selection effects. For a given pair of condition (e.g., HLAB57+ versus HLAB57-), and for a given $\alpha$ level, the set of discoveries at level $\alpha$ (i.e., the set of all positions/amino acid pairs such that the posterior probability of a differential selection effect between the two conditions is greater than $1-\alpha$) was determined. A discovery was then deemed to be false if the true selective effect of that amino acid at that position is either 0 or of the opposite direction. The precision and sensitivity were tabulated as a function of 1-α (table 2-IV and table 2-V, for condition 2 and 3, respectively). As we see in table 2-IV and 2-V, for a given posterior probability threshold, the precision decreases when the proportion of differentially selected sites (f) decreases. This reflects the fact that the number of true positives is directly proportional to the proportion of sites with differential selection effects, while the number of false positives remains stable. Overall, the power of the method is relatively low. Under a precision of 0.9 (10% of false discoveries), the sensitivity (or recall) is between 1% and 0.3%, depending on the exact simulation condition (i.e., less than 1% of the differentially selected positions are detected).

Table 2-IV. Precision (prec) and sensitivity (sens) as a function of the proportion of true (simulated) differentially selected sites (f) in condition B57+ hosts, under model M1 and M2.

| | **M1** | | | | | | **M2** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **threshold** | **f=0.5** | | **f=0.1** | | **f=0.05** | | **f=0.5** | | **f=0.1** | | **f=0.05** | |
| | prec | sens | prec | sens | prec | sens | prec | sens | prec | sens | prec | sens |
| **>0.50** | 26.7 | 53.5 | 4.7 | 47.0 | 2.8 | 56.4 | 2.5 | 49.7 | 5.3 | 53.0 | 25.5 | 51.1 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **>0.55** | 39.1 | 9.0 | 7.7 | 7.6 | 4.7 | 9.3 | 4.0 | 8.4 | 8.3 | 8.4 | 37.5 | 8.5 |
| **>0.60** | 47.0 | 6.1 | 10.0 | 4.9 | 6.3 | 6.0 | 5.6 | 5.4 | 10.9 | 5.3 | 45.2 | 5.2 |
| **>0.65** | 55.2 | 4.3 | 14.1 | 3.5 | 9.0 | 4.4 | 8.2 | 3.8 | 14.3 | 3.4 | 52.8 | 3.4 |
| **>0.70** | 67.0 | 2.9 | 21.9 | 2.4 | 15.8 | 3.2 | 12.1 | 2.1 | 21.3 | 1.8 | 63.8 | 1.9 |
| **>0.75** | 78.7 | 2.0 | 34.4 | 1.8 | 25.7 | 2.5 | 23.3 | 1.4 | 33.0 | 1.1 | 78.2 | 1.1 |
| **>0.80** | 84.1 | 1.5 | 50.8 | 1.4 | 35.2 | 1.8 | 33.0 | 0.7 | 44.8 | 0.6 | 86.5 | 0.7 |
| **>0.85** | 91.2 | 1.1 | 66.2 | 1.0 | 49.0 | 1.2 | 41.4 | 0.3 | 73.7 | 0.3 | 92.7 | 0.3 |
| **>0.90** | 93.5 | 0.8 | 81.7 | 0.7 | 68.3 | 0.6 | 33.3 | 0.1 | 90.9 | 0.1 | 100 | 0.1 |
| **>0.95** | 96.0 | 0.4 | 93.3 | 0.3 | 86.7 | 0.3 | 0.0 | 0.0 | 100 | 0.01 | 100 | 0.03 |

Table 2-V. Precision (prec) and sensitivity (sens) as a function of the true (simulated) proportion of differentially selected sites (f) in condition B35+ hosts, under model M1 and M2.

| | **M1** | | | | | | **M2** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **threshold** | **f=0.5** | | **f=0.1** | | **f=0.05** | | **f=0.5** | | **f=0.1** | | **f=0.05** | |
| | prec | sens | prec | sens | prec | sens | prec | sens | prec | sens | prec | sens |
| **>0.50** | 26.6 | 53.2 | 5.2 | 51.6 | 2.5 | 49.8 | 2.7 | 54.3 | 5.1 | 51.0 | 26.1 | 52.2 |
| **>0.55** | 40.2 | 12.0 | 8.3 | 10.7 | 4.0 | 10.2 | 4.4 | 11.7 | 8.3 | 10.8 | 37.2 | 10.9 |
| **>0.60** | 47.2 | 8.3 | 10.7 | 7.6 | 5.5 | 7.4 | 5.7 | 7.9 | 10.5 | 7.6 | 44.1 | 7.4 |
| **>0.65** | 54.1 | 6.2 | 13.6 | 5.5 | 7.5 | 5.6 | 6.7 | 5.1 | 13.5 | 5.4 | 49.5 | 5.0 |
| **>0.70** | 64.7 | 3.9 | 21.4 | 3.7 | 11.9 | 3.6 | 9.9 | 2.9 | 21.4 | 3.3 | 60.2 | 2.9 |
| **>0.75** | 75.6 | 3.0 | 32.1 | 2.9 | 18.1 | 2.6 | 14.3 | 1.8 | 34.4 | 2.5 | 70.7 | 1.9 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **>0.80** | 83.5 | 2.4 | 44.4 | 2.3 | 27.5 | 2.1 | 22.7 | 1.2 | 50.3 | 1.8 | 80.8 | 1.2 |
| **>0.85** | 90.7 | 1.8 | 61.6 | 1.8 | 46.4 | 1.8 | 38.2 | 0.8 | 67.1 | 1.1 | 88.7 | 0.7 |
| **>0.90** | 93.7 | 1.4 | 77.3 | 1.3 | 64.1 | 1.3 | 66.7 | 0.4 | 79.4 | 0.6 | 97.1 | 0.4 |
| **>0.95** | 96.8 | 1.0 | 89.2 | 0.8 | 89.1 | 0.9 | 100 | 0.1 | 100 | 0.2 | 98.3 | 0.1 |

## Analyses of HIV empirical data

We applied our DS model to a dataset of HIV coding sequences (encoding the *Gag* protein) obtained from 41 patients. We used this dataset for two reasons. First, it contains multiple sequences for each patient, thus providing empirical information about within-host evolution of viral genetic sequences. Second, the HLA type of the patients is known, and therefore, it is possible to correlate the amino acid patterns observed in viral sequences with the HLA type of the host.

Accordingly, in this study, we partitioned the phylogenetic tree relating the viral sequences into different categories. A global reference selection profile was estimated by our method. This reference fitness landscapes, which captures the baseline site-specific amino acid preferences in the form of site-specific vectors of 20 fitness factors (one for each amino acid), can be visualized using a graphical logo representation (Schneider and Stephens 1990) and compared with the reference HIV-1 sequence (HXB2, the first 60 coding positions are shown in figure 2-3). The selection profile inferred with our method is highly similar to the reference sequence (the fittest amino acid corresponds to the amino acid of the reference sequence at 86% of the coding positions). In some cases, compared to the reference sequence, the fitness profile suggests a distinct but biochemically similar dominant amino acid (e.g., position 15, K instead of R), or several

equally fit amino acids at one position (e.g., position 30, K and R). This corresponds to the actual sequence variation observed in our empirical alignment. Altogether, this global reference selection profile illustrates that HIV evolution occurs on a background characterized by strong purifying selection, allowing for a limited set of amino acid sequences for the viral protein.

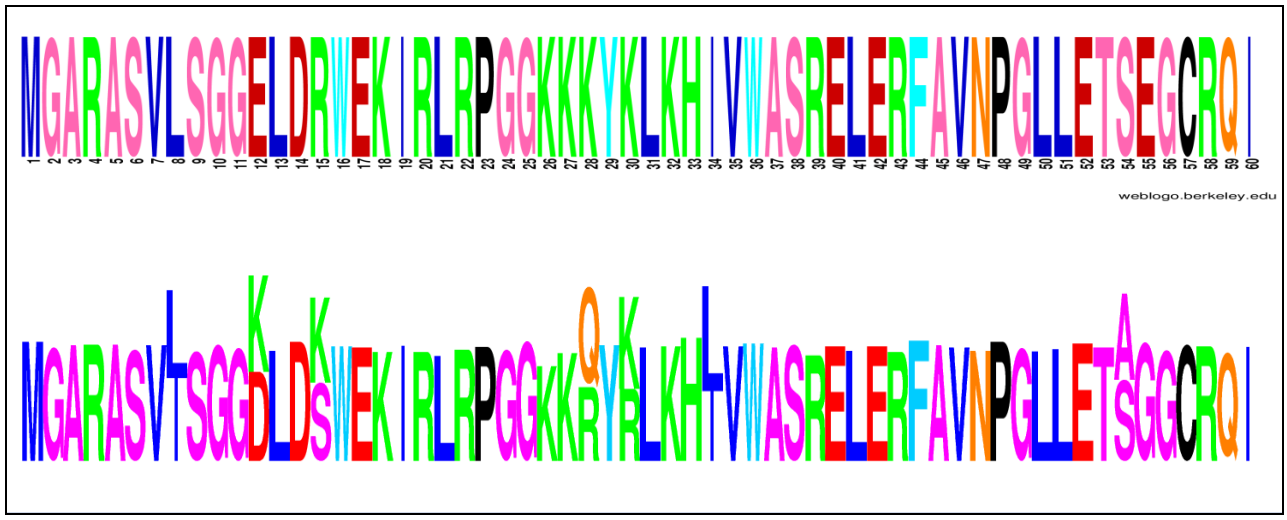



Figure 2-3. Comparison of global selection profile estimated by DS model (*b*) with HIV reference sequence HXB2 (*a*). The first 60 amino acids are shown. The reference logo was made using Weblogo (Crooks, Hon et al. 2004).

Against this background fitness landscape, our model then estimates differential selection profiles between each pair of conditions: first, between within-host and between-host (figure 2-4-*b* and 2-5-*b*), and second, among within-host sequences, between HLA-B57- and HLA-B57+ sequences (figure 2-4-*c*), or between HLA-B35- and HLA-B35+ sequences (figure 2-5-*c*). The

logos represented on figure 2-4 and 2-5 indicate whether the fitness of any particular amino acid is inferred to be increased (at the top) or decreased (at the bottom) with posterior probability >0.80, at a given position, between the two conditions being compared. These figures only give point estimates for the differential effects. In practice, the posterior probability support associated to these estimates is most often low, at about 0.5 (figure 2-6), except for a small subset of positions for which the model infers stronger evidence for a differential selection effect. These more clear-cut cases represent our findings, which are given in table 2-VI for the two model settings. In the following, we report the findings for two thresholds, at 0.80 and 0.90. We will refer to the corresponding discoveries as weakly and strongly supported findings, respectively.

Table 2-VI. Number of differentially selected amino acid-positions with posterior probability >0.80 and >0.90, in different conditions under model M1 and M2.

| threshold | model | within-patient | $B57^+$ patients | B35+ patients |
|---|---|---|---|---|
| **>0.80** | M1 | 281 | 15 | 48 |
| **>0.80** | M2 | 286 | 5 | 30 |
| **>0.90** | M1 | 54 | 2 | 13 |
| **>0.90** | M2 | 56 | 0 | 1 |

By far, we observe in table 2-VI that the largest number of differentially selected amino acid variants is found when comparing the within- and between-patient conditions, with more than 280 findings under both models. On the other hand, the corresponding profiles suggest that this is mostly due to a global difference intact the intensity of selection (or a global difference in statistical power), rather than to specific selective differences between the two conditions (see Discussion). The differences between alternative HLA backgrounds, on the contrary, seem to be more specific.

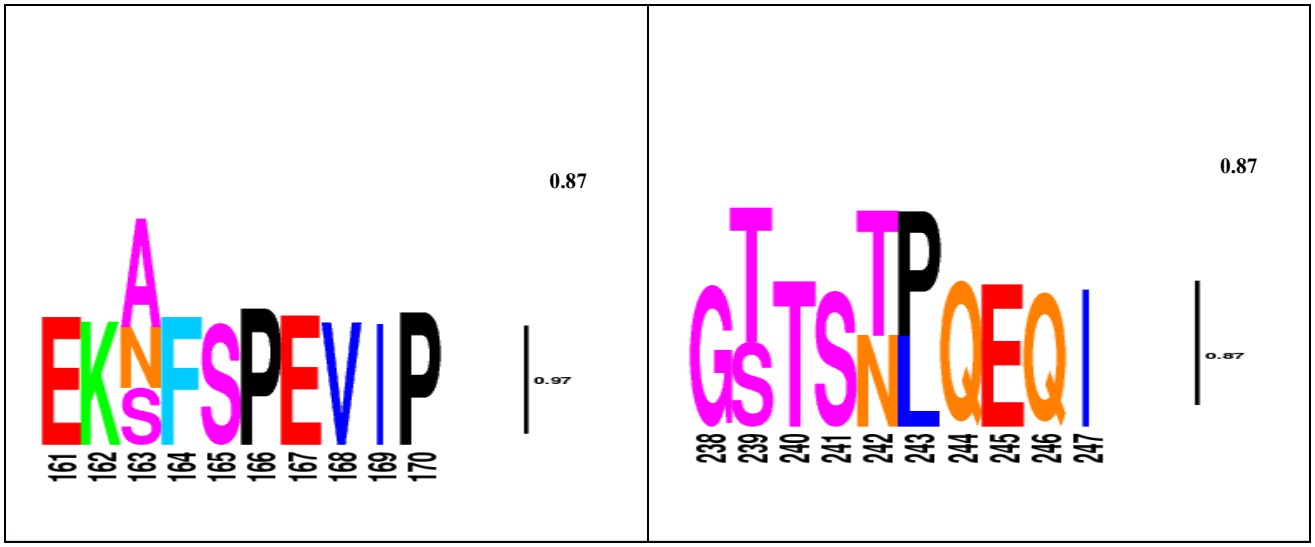
Comparing the number of findings reported in table 2-VI under conditions 2 and 3 with the mean number of false positives in simulation experiments under the null model with no differential selection and for the same threshold (table 2-III) gives a rough estimate of the fraction of false discoveries. Thus, for a threshold of 0.9, the fraction of false discoveries is approximately 20% in condition 1 and 9% in condition three under model M1, whereas model M2 does not seem to lead to a significant enrichment compared to the expected number of false positives. Therefore, in the following, we consider only model M1.

The findings under model M1 are listed in more details (position, amino acid, lower and upper 95% credible intervals and posterior probability support) in table 2-VII and table 2-VIII for B57+ and B35+ conditions, respectively. For each finding, the direction of the effect (whether the fitness is increased or decreased between the two conditions being tested) is indicated, together with the posterior probability that the effect is >0 or <0 (depending on the direction of the effect). Among our findings, there are some known mutations identified in association with specific HLAs. Two crucial HIV-1 escape mutations defined in B57+ patients are T242N and A163X in epitopes TW10 (Leslie, Pfafferott et al. 2004; Brockman, Schneidewind et al. 2007) and KF11 (Leslie, Kavanagh et al. 2005; Weber, Weberova et al. 2006), respectively. X at position 163 is mostly P and N. The logos of the corresponding regions are shown in figure 2-4. The selection factors estimated at these positions are in agreement with these previously known escape mutations.

Intriguingly, the T/N escape variant at position 242 (TW10 epitope) is not recovered by the mechanistic model (M2), suggesting that the phenomenological model is more capable to predict differential selection patterns. This confirms our simulation studies, proving that the phenomenological model has a higher detection power. Also, of interest, our method does not infer that T is preferred in a B57- environment, whereas N is favored in a B57+ background. Instead, it

suggests that both amino acids are acceptable in a B57- environment, but that N becomes the only one favored in B57+ patients. A similar pattern is observed for the A163X escape mutation, with posterior probability = 0.77. One known mutation for B35+ individuals is E260D in NY10 epitope (Matthews, Koyanagi et al. 2012). Our method detects this mutation to be under condition-specific selection with posterior probability of 0.81 (figure 2-5).

*a*



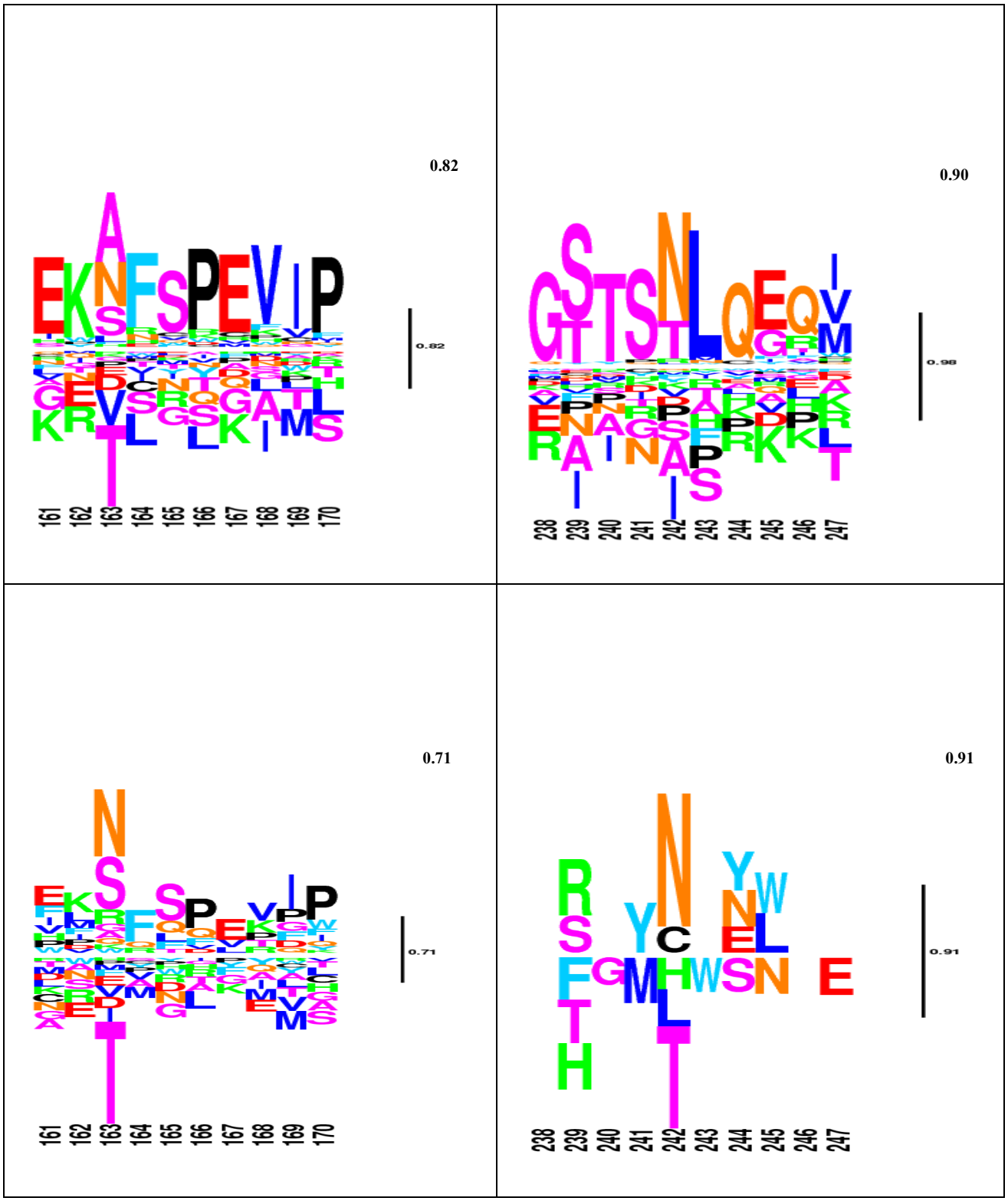


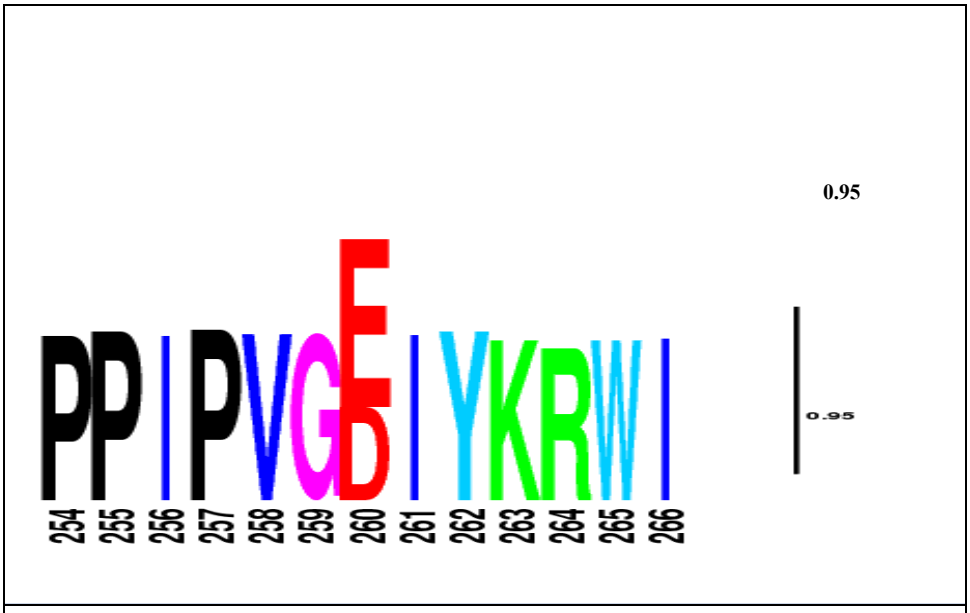

Figure 2-4. Global and differential selection profiles (for HLA-B57). *(a)* Global Selection profile (G). *(b)* Differential Selection profile contrasting between- and within-patient selection. *(c)* Differential Selection profile for HLA-B57+ versus HLA-B57-. The posterior probability (pp) of an increased fitness for N and a

decreased fitness for T at position 242 (TW10 epitope), in HLA-B57+ compared to HLA-B57-, is 0.93 and 0.87, respectively. At position 163 (KF11 epitope), the fitness of N is increased with pp of 0.77. The logos are filtered for pp below 0.05. Heights are proportional to posterior mean differential selective effects.

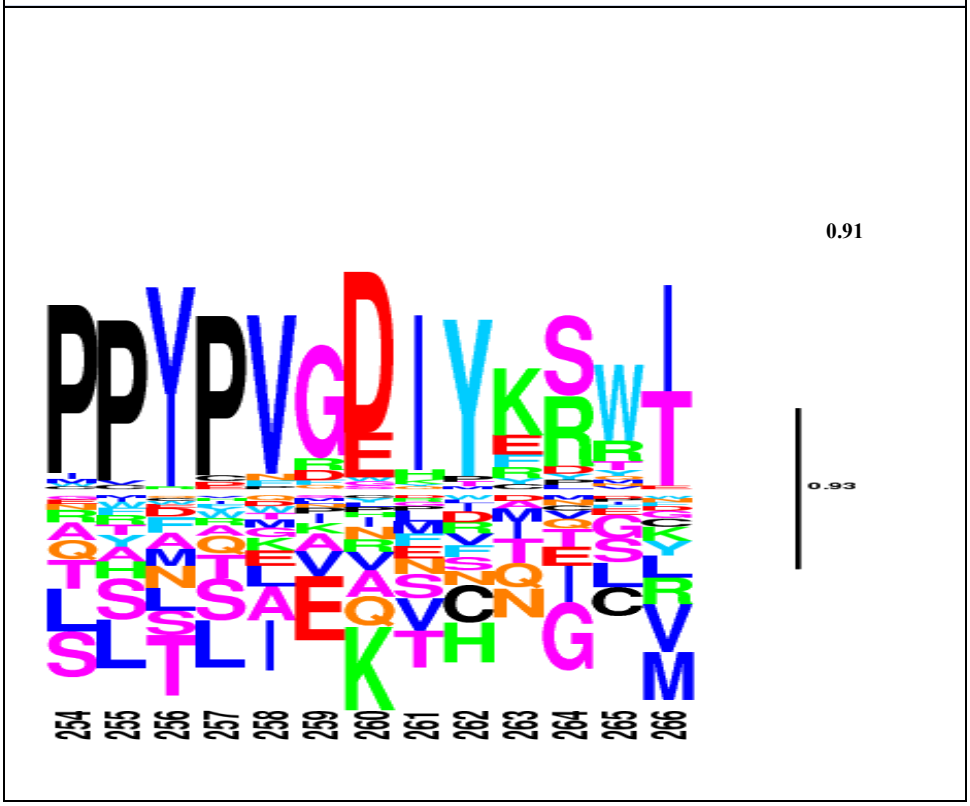
Table 2-VII. List of differentially selected amino acids for B57+ hosts with posterior probability> 0.80. The amino acid-positions are sorted according to the posterior probability score. Median, lower and upper 95% credible intervals and the direction of the effect on fitness (increased or decrease) are indicated.

| position | amino acid | posterior probability | median | lower | upper | fitness |
|---|---|---|---|---|---|---|
| **242** | N | 0.93 | 1.36 | -0.37 | 3.07 | increased |
| **248** | G | 0.91 | -1.20 | -2.82 | 0.45 | decreased |
| **30** | Q | 0.89 | 1.09 | -0.69 | 2.92 | increased |
| **242** | T | 0.87 | -0.95 | -2.55 | 0.78 | decreased |
| **30** | K | 0.87 | -0.96 | -2.49 | 0.69 | decreased |
| **357** | A | 0.86 | 0.94 | -0.73 | 2.86 | increased |
| **15** | R | 0.86 | 0.72 | -1.01 | 2.41 | increased |
| **118** | A | 0.85 | -0.93 | -2.69 | 0.79 | decreased |
| **239** | S | 0.85 | 1.02 | -0.95 | 2.64 | increased |
| **137** | L | 0.82 | -0.86 | -2.55 | 0.93 | decreased |
| **326** | S | 0.81 | 0.79 | -1.28 | 2.46 | increased |
| **357** | G | 0.81 | -0.78 | -2.55 | 0.97 | decreased |
| **280** | T | 0.80 | 0.83 | -0.79 | 2.43 | increased |
| **12** | E | 0.80 | 0.71 | -0.96 | 2.43 | increased |
| **248** | A | 0.80 | 0.66 | -0.97 | 2.42 | increased |
| **223** | I | 0.80 | -0.70 | -2.28 | 1.02 | decreased |

*a*


0.95
0.95
254
255
256
257
258
259
260
261
262
263
264
265
266

*b*


0.91
0.93
254
255
256
257
258
259
260
261
262
263
264
265
266

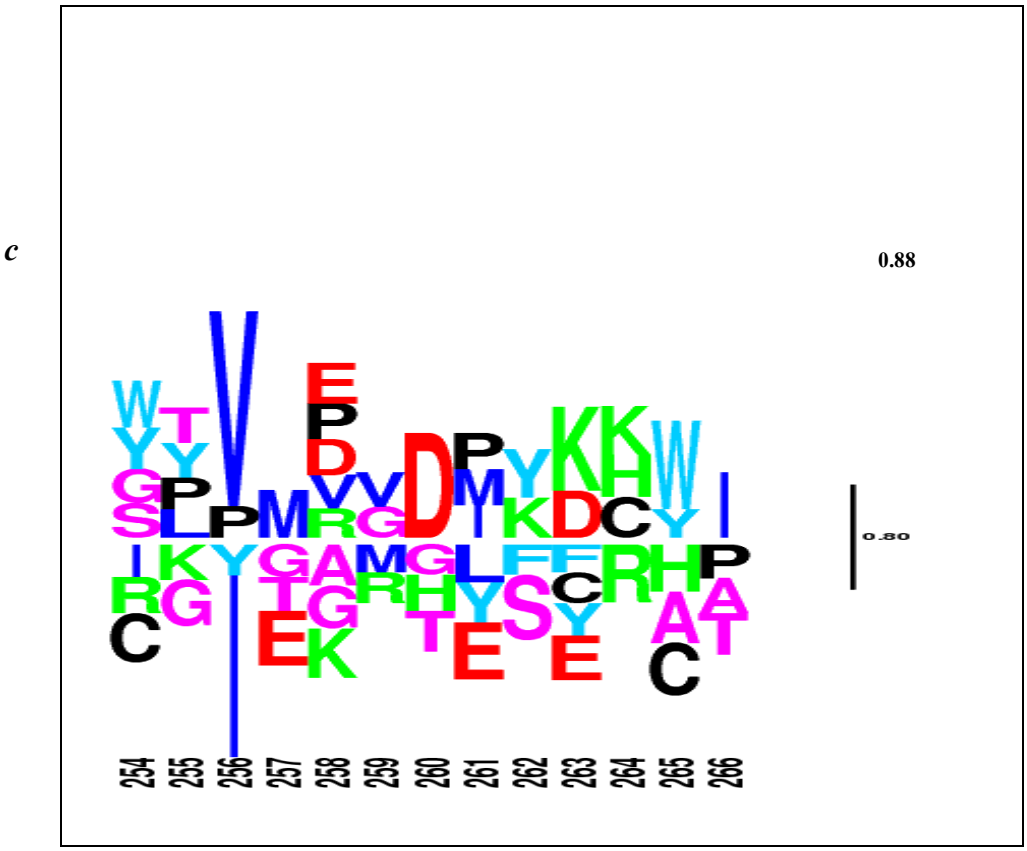



Figure 2-5. Global *(a)* and differential selection profiles, contrasting within and between patients *(b)* and HLA-B35+ versus HLA-B35- *(c)*. In *(c)*, the posterior probability (pp) of fitness shift from D to E at position 260 is 0.81. The logos are filtered for pp less than 0.05. Heights are proportional to posterior mean differential selective effects.

Table 2-VIII. List of differentially selected amino acids for B35+ hosts with posterior probability > 0.80. The amino acid-positions are sorted according to the posterior probability score. Median, lower and upper 95% credible intervals and the direction of the effect on fitness (increased or decrease) are indicated.

| position | amino acid | posterior probability | median | lower | upper | fitness |
|---|---|---|---|---|---|---|
| **46** | L | 0.97 | 1.69 | -0.05 | 3.44 | increased |
| **34** | L | 0.96 | 1.52 | -0.31 | 3.19 | increased |
| **252** | H | 0.96 | 1.59 | -0.18 | 3.28 | increased |
| **111** | S | 0.93 | -1.15 | -2.72 | 0.49 | decreased |
| **127** | Q | 0.93 | -1.11 | -2.74 | 0.48 | decreased |
| **376** | V | 0.93 | 1.16 | -0.49 | 2.68 | increased |
| **312** | D | 0.92 | 1.23 | -0.55 | 3.06 | increased |
| **137** | M | 0.92 | 1.26 | -0.47 | 3.22 | increased |
| **252** | N | 0.92 | -1.05 | -2.60 | 0.48 | decreased |
| **30** | K | 0.92 | -1.05 | -2.44 | 0.52 | decreased |
| **248** | A | 0.91 | 1.25 | -0.41 | 3.07 | increased |
| **310** | T | 0.91 | 1.25 | -0.54 | 2.97 | increased |
| **441** | H | 0.89 | 0.95 | -0.43 | 2.46 | increased |
| **46** | V | 0.89 | -1.06 | -2.74 | 0.52 | decreased |
| **67** | A | 0.89 | 1.09 | -0.66 | 2.82 | increased |
| **111** | C | 0.88 | 1.08 | -0.75 | 2.76 | increased |
| **375** | V | 0.88 | -0.85 | -2.48 | 0.72 | decreased |
| **255** | V | 0.88 | 1.08 | -0.79 | 2.61 | increased |
| **441** | Y | 0.87 | -0.92 | -2.37 | 0.53 | decreased |
| **405** | I | 0.86 | 0.94 | -0.72 | 2.51 | increased |
| **15** | Q | 0.86 | 0.94 | -0.77 | 2.84 | increased |
| **138** | L | 0.86 | -0.90 | -2.41 | 0.76 | decreased |
| **376** | I | 0.85 | -0.81 | -2.26 | 0.67 | decreased |
| **127** | T | 0.85 | 1.01 | -0.86 | 2.83 | increased |
| **69** | Q | 0.84 | -0.79 | -2.37 | 0.78 | decreased |

| | | | | | | |
|---|---|---|---|---|---|---|
| **81** | A | 0.84 | 0.94 | -0.74 | 2.65 | increased |
| **176** | A | 0.84 | 0.86 | -0.88 | 2.86 | increased |
| **280** | T | 0.83 | 0.96 | -0.86 | 2.40 | increased |
| **348** | S | 0.83 | 0.97 | -0.90 | 2.87 | increased |
| **61** | I | 0.83 | 0.77 | -1.14 | 2.61 | increased |
| **81** | T | 0.83 | -0.82 | -2.41 | 0.85 | decreased |
| **268** | M | 0.82 | 0.81 | -0.81 | 2.45 | increased |
| **280** | A | 0.82 | -0.82 | -2.41 | 0.85 | decreased |
| **388** | K | 0.82 | 0.74 | -0.90 | 2.37 | increased |
| **389** | P | 0.82 | 0.81 | -0.81 | 2.45 | increased |
| **397** | R | 0.82 | 0.72 | -1.00 | 2.53 | increased |
| **95** | R | 0.82 | 0.77 | -0.83 | 2.39 | increased |
| **68** | I | 0.81 | 0.87 | -1.14 | 2.67 | increased |
| **215** | L | 0.81 | -0.73 | -2.19 | 0.70 | decreased |
| **118** | T | 0.81 | 0.70 | -0.95 | 2.33 | increased |
| **260** | D | 0.81 | 0.75 | -1.00 | 2.48 | increased |
| **54** | A | 0.81 | 0.75 | -0.96 | 2.52 | increased |
| **93** | A | 0.80 | 0.73 | -1.04 | 2.44 | increased |
| **28** | K | 0.80 | -0.66 | -2.46 | 1.06 | decreased |
| **58** | K | 0.80 | 0.69 | -1.31 | 2.34 | increased |

## Robustness to the choice of the tree topology

The method relies on a fixed tree topology. However, in practice, the tree is reconstructed with errors. To test the robustness of the inference, we analyzed three alternative tree topologies, under the M1 model). We refer to these trees as tree T1, T2, and T3 (see methods). The set of

differentially selected positions were found to be very similar for all trees (table 2-IX), suggesting that the exact details of the tree topology not be so determining in the present context.

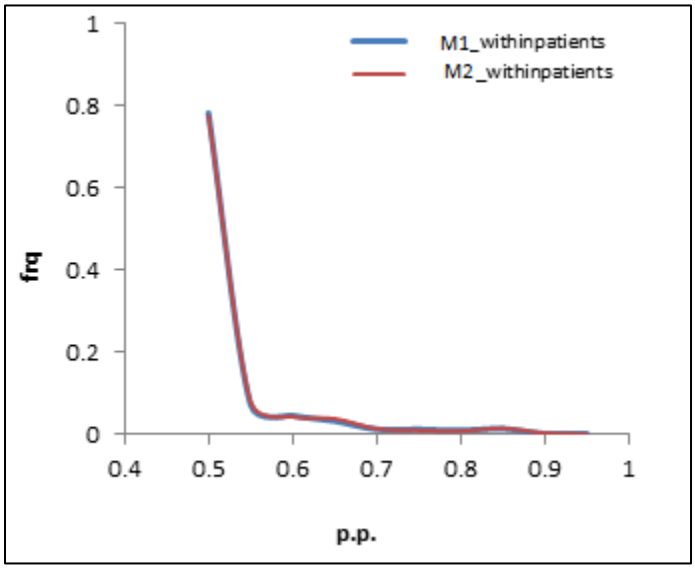



Figure 2-6. Posterior probability frequency plots of differential selection effects across all amino acid-positions; phenomenological (M1) vs. mechanistic (M2). Posterior probability of the majority of amino acid-position lies between 0.5-0.6.

By comparing the number of positions declared significant for each threshold (shown in table 2-IX), we see that for B57+ condition, the number of findings is very close in different tree topologies (15, 12 and 18 under posterior probability > 0.80, and 2, 2 and 3 under posterior probability > 0.90). We also summarized the common positions between the three topologies as a Venn diagram in figure 2-7. There is only one position in T1 which is not recovered by T2 or T3. The majority of positions (10) were found by all trees. None of the discrepancies between analyses under differing topologies belong to the positions previously known to correspond to viral escape

mutants. Altogether, the relatively small number of sequences that had to be removed, combined with the relative robustness of our result to the choice of the tree topologies despite their distances (specially between tree T1 and tree T2 and T3, see table 2-II), suggests that the problems of multiple infection patterns, or tree reconstruction errors, have a globally marginal impact on our analysis.

Table 2-IX. Number of differentially selected amino acid-positions with posterior probability >0.80 and >0.90 obtained by M1-DS model using tree T1, T2, and T3.

| **threshold** | **tree topology** | **B57+ patients** | **B35+ patients** |
|---|---|---|---|
| >0.80 | T1 | 15 | 48 |
| >0.80 | T2 | 12 | 51 |
| >0.80 | T3 | 18 | 48 |
| >0.90 | T1 | 2 | 13 |
| >0.90 | T2 | 2 | 10 |
| >0.90 | T3 | 3 | 12 |

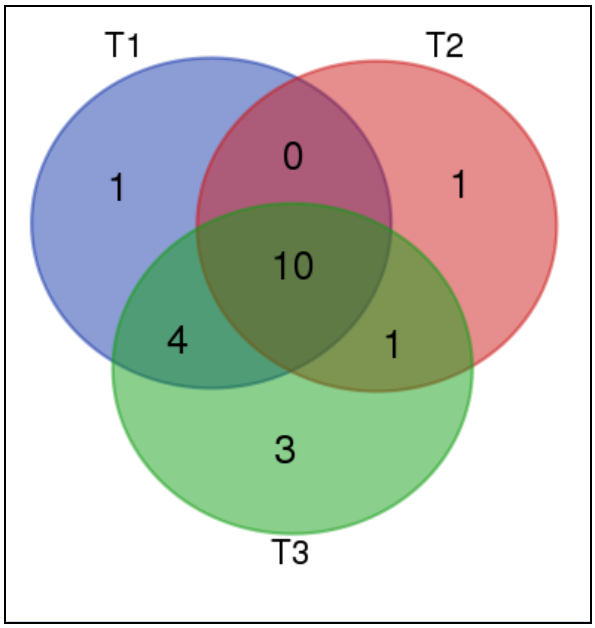


Figure 2-7. Venn diagram of positions found with posterior probability > 0.80, using tree T1 (NJ topology), T2 (MrBayes topology with constraint) and T3 (MrBayes topology without constraint). Three topologies share 10 positions.

## 2.5.5 Discussion

Here, we have introduced a hierarchical Bayesian method for detecting adaptive patterns in protein-coding sequences as a function of known selective backgrounds. Compared with previously introduced methods (Carlson, Brumme et al. 2008; Tamuri, dos Reis et al. 2009), our approach has several additional features. The approach of Carlson *et al.* (Carlson, Brumme et al. 2008), relying on a Bayesian network representation, is formulated at the codon level. Also, it can accommodate epistatic effects (see introduction). Nevertheless, it is focused on the terminal branches of the phylogeny and therefore ignores potentially relevant empirical information from the deeper parts of the phylogenetic tree. The approach of Tamuri *et al.* (Tamuri, dos Reis et al. 2009; Tamuri, dos Reis et al. 2012; Tamuri, Goldman et al. 2014), in contrast, entirely integrates

the empirical signal over the entire tree, and is thus much more similar, in spirit, to the present method. The main difference is in the statistical framework used to deal with site-specific effects (empirical Bayes versus maximum-likelihood estimation). The fact that our method integrates the empirical signal about more ancient codon substitutions opens interesting possibilities, in particular, for comparing short-term (within-host) and long-term (between-host) adaptive patterns. As it stands, however, the selection profiles obtained for between- and within-host are not yet so assuring: the within-host differential selection profiles obtained through our method (figure 2-4-*b* and 2-5-*b*) seem to partially reproduce the condition-independent amino acid fitness profiles (figure 2-4-*a* and 2-5-*a*). The reasons for such a redundant output are not clear. Deleterious mutations segregating within-host but purified away in the long-term (and therefore absent from the deeper branches of the phylogeny connecting host-specific clusters) are an essential difference between within- and between-host conditions. However, such segregating polymorphisms would be expected to result in an opposite pattern, leading to artefactual high selection coefficients in the within-host condition for unfit amino acids that are not observed in the between-host selection profiles. One alternative explanation for the observed redundancy would be that a Dirichlet distribution does not correctly capture the law of condition-independent selection profiles across sites. Possibly for that reason, the remaining part of the condition-independent selective effects may be captured by the differential selection profile of the within-host condition. Ultimately, more sophisticated hierarchical Bayesian settings could be used, such as non-parametric priors (Rodrigue, Philippe et al. 2010). The combination of condition- and site-specific effects is computationally challenging, and further algorithmic work is therefore needed in this direction to fully accommodate arbitrary distributions of random effects across positions and conditions.

The distribution of differential selective effects across sites and conditions may also need additional statistical and computational developments in the long term. Here, we have used Normal distributions centered on 0 to model differential selective effects. Doing this leads to efficient soft shrinkage toward 0. However, this approach does not implement sparsity. All amino acids, at all positions and under all conditions, have non-zero differential selective effects with a posterior probability of one. Ultimately, sparse differential selection profiles (with only a small number of positions and amino acids displaying significant non-null differential selective effects) could be obtained through the use a spike-and-slab mixture model (Lewin, Bochkina et al. 2007). In this context, estimating the proportion of non-null effects, as well as the effect size distribution directly on the empirical data, would have several advantages, including an increased power, more accurate quantification of the effect sizes, as well as a more direct control of the rate of false discovery. In addition, this hierarchical model would allow for testing the null hypothesis that the gene has no differentially selected positions, by purely comparing the full model with the one constrained to have a null proportion of differential effects.

As suggested by our simulation experiments, modeling differential selection effects as random variables, with a distribution centered on 0, ensures good regularity properties of the approach. On the other hand, the power of the approach appears to be slightly low. Further development of the current approach, along the lines, just suggested, combined with a more systematic comparison with the currently existing alternatives (Carlson, Brumme et al. 2008; Tamuri, dos Reis et al. 2009; Tamuri, dos Reis et al. 2012; Tamuri, Goldman et al. 2014), will have to be conducted, in order to establish whether this low power is a specific weakness of the present method (in particular because of the lack of sparsity of the model), or more fundamentally an inherent limitation of the

problem of detecting weak effects across a large number of coding sites and for all possible amino acids.

Two alternative models of the rate of change between codons were considered in this study: one purely phenomenological (Robinson, Jones et al. 2003; Rodrigue, Philippe et al. 2010), and another one that has a better mechanistic justification, based on first principles of population genetics. When applied to HIV sequences, the mechanistic model does not seem to lead to better results, compared to the phenomenological approach. In particular, it fails to detect known HLA-restricted escape mutations. The mechanistic model, however, makes several assumptions that are apparently not warranted in the present context: low-mutation approximation, and more fundamentally, a mutation-fixation paradigm (Halpern and Bruno 1998; Yang and Nielsen 2008), which amounts to ignoring clonal interference. In sharp contrast, viral sequences evolve under a very high mutation rate, leading to strong clonal interference. Another consequence of the very high mutation rate is that segregating deleterious polymorphisms are expected to be present at a substantial frequency, something which is not correctly captured by the mutation-selection model: fundamentally, this model is meant to be applied to inter-specific data. Here, in contrast, a meta-population model would be more adequate. The theoretical and computational developments in this direction still appear to be challenging.

Our method does not take into account epistatic interactions between positions. Those interactions seem to play a significant role in HIV evolution, in particular concerning escape mutations. Most escape mutations cause a viral fitness cost which leads to decreased replication of the virus (Brockman, Schneidewind et al. 2007). Position 242 is under the most substantial selection pressure from the immune system which corresponds to the ability of B57+ hosts to control the disease. Figure 2-8 shows the location of amino acid 242 in the 3D structure of *Gag*

protein, specifically p24 Capsid protein (amino acid 133-263). p24 Capsid lies in the N-terminal domain of *Gag* protein (CAN), consisting helix 6, N-terminal hairpin and the binding loop to the host protein cyclophilin A (CypA), which is necessary for HIV infectivity (Braaten, Franke et al. 1996). As figure 2-8 shows, amino acid 242 is situated where it caps the N terminus of helix 6. The mutation from T to N at this position destabilizes helix 6 and may disrupt the conformational coupling between three parts of CAN. This would disturb CypA binding and reduce the fitness of the virus (Martinez-Picado, Prado et al. 2006).

T242N mutation in B57+ individuals reverts in viruses transmitted to an HLA-mismatched host (Leslie, Pfafferott et al. 2004), which confirms that the mutation has a high fitness cost for the virus regarding replication capacity (Martinez-Picado, Prado et al. 2006). This fitness cost might be compensated for, to some extent, by mutations at other positions, mostly around the escape mutation. In sequences with T242N mutation, the compensatory mutations, H219Q, I223V, M228I/V, G248A, and N252H have been identified (Leslie, Pfafferott et al. 2004; Brockman, Schneidewind et al. 2007). It has been reported that these mutations are significantly more frequent in HLA-B57+ patients with a progressing disease compared to HLA-B57+ non-progressors (Brockman, Schneidewind et al. 2007). Here, we did not see significant differences for final amino acids (Q, V, I/V, A and H) between B57+ and B57- patients at those suppressing positions (their posterior probability is less than 0.70), although initial amino acids are strongly unfavored (posterior probability =0.80, 0.91, 0.77 for I, G and N at positions 223, 248 and 252, respectively). There may be two reasons for that; first, our model takes each site into account independently, and codon co-variation is not considered. Secondly, contrary to escape mutations which revert in the HLA mismatch host, compensatory mutations do not tend to revert after transmission to HLA mismatch individuals (Leslie, Pfafferott et al. 2004). For example, H219Q, the associated mutation

to T242N, is reported to be maintained after transmission from B57+ to B57- hosts. So, this mutation might be stable and spread in the population. As it stands, explicitly implementing epistatic effects in the context of the present modeling framework appears to be challenging, although not impossible (Kleinman, Rodrigue et al. 2010).

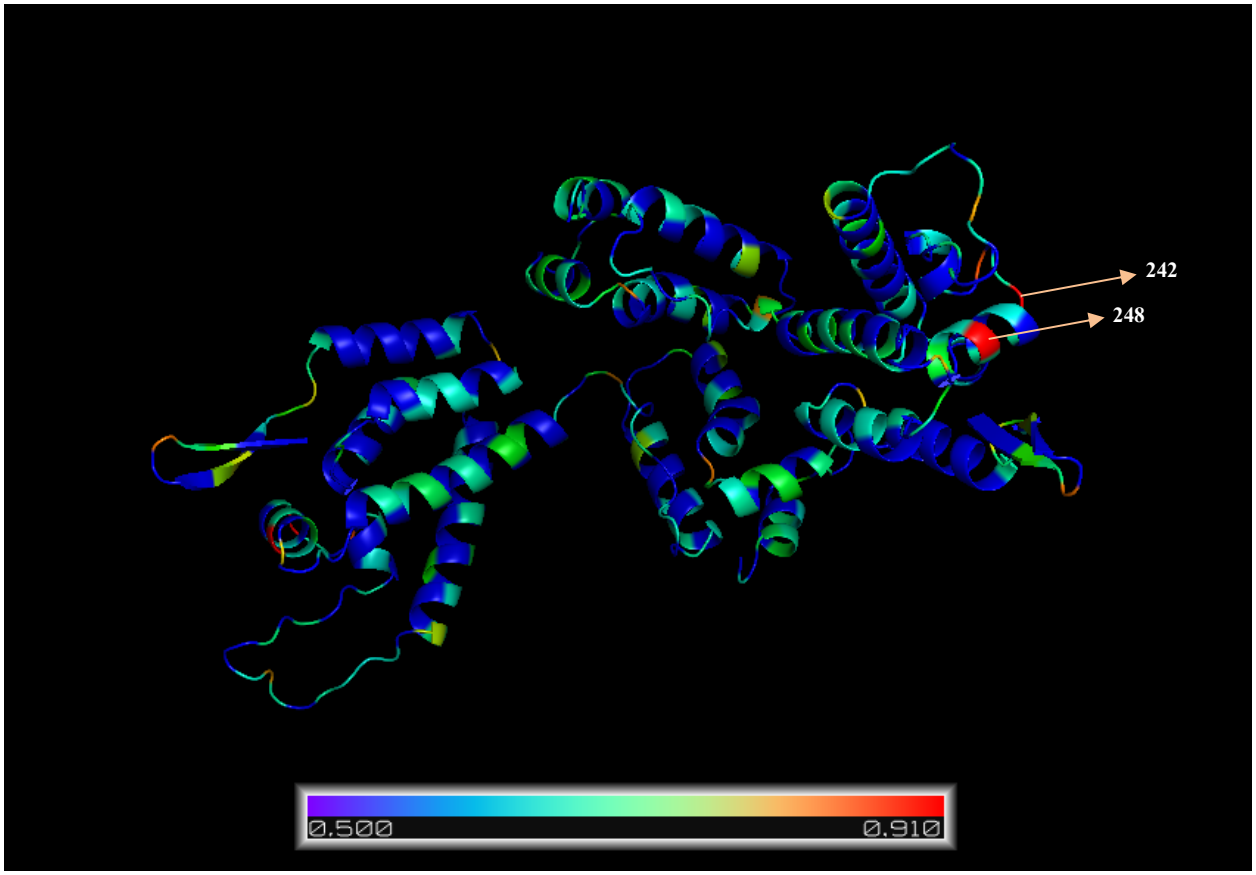



Figure 2-8. 3D structure of p24 Capsid protein. The amino acids are colored based on their posterior probabilities using PyMol (L DeLano 2002). The amino acids in red have the highest posterior probability for selection in B57+ individuals.

## 2.5.6 Conclusions

We proposed a phylogenetic differential selection model, which is able to find adaptive patterns in coding sequences influenced by selective environments. Applying the model to HIV-1 *Gag* sequences leads to the detection of a few amino acid-positions that are differentially selected under different host HLA types, as HIV escapes from immune system through its fast evolution. The model is thus able to find known HLA-restricted mutations, as well as some new mutations, to be under differential selection. The power of our model is that it is capable of detecting both positive and negative selection pressure on each amino acid at each position under each environmental condition.

This DS model can be used in other situations in which differential selective effects are suspected, as a function of known predictors, for viruses (e.g., finding adaptive patterns of HIV sequences under the selection pressure of immune system or antiviral therapy provides an insight of the direction of HIV-1 evolution in different hosts with different genetic characteristics), or in other species (e.g., convergent adaptations of multiple lineages of plants, or animals, to specific environmental conditions (Parto and Lartillot 2018).

# CHAPTER 3: MOLECULAR ADAPTATION IN RUBISCO

The Differential Selection model introduced in the last chapter was applied to HIV evolution as a response to HLA diversity. Nevertheless, in principle, this model can be applied to any case in which repeated (i.e., convergent) evolution between small numbers of known selective environments has occurred.

A good case in point is the evolution of photosynthesis in plants, and the difference between C3 and C4 photosynthetic regimes. C4 plants have repeatedly evolved from their C3 ancestors to overcome unfavorable climate situations, and C3/C4 photosynthesis is thus a great example of convergent evolution in variable environmental conditions. A very important enzyme in the process of photosynthesis is Rubisco. Both C3 and C4 plants use Rubisco for the fixation of $CO_2$ in their metabolic pathway, but the specificity and the efficiency of the enzyme are different in the two groups. Thus, we may expect Rubisco to show patterns of molecular convergent evolution associated with the transition between C3 to C4 photosynthetic regimes.

Rubisco molecular evolution has previously been analyzed using classical codon models, based on the estimation of site- and/or branch-specific non-synonymous to synonymous rate ratios ($d_N/d_S$ values). In this context, sites under positive, or diversifying, selection (i.e. with $d_N/d_S$ greater than 1) have been reported. This raises an interesting question: to what extent we should characterize the adaptive patterns in Rubisco regarding diversifying versus convergent evolution? How should we distinguish between these two alternative modes of adaptation, and to what extent the DS and the classical codon models differ in their aim and their predictions?

In this chapter, the DS model (presented in chapter 2) is applied to Rubisco to investigate the patterns of convergent evolution at the molecular level in this enzyme. We also developed a version of the classical codon models, allowing for a site- and condition-specific estimation of $d_N/d_S$. We then, compared the patterns of positive selection inferred by this model with the patterns of convergent adaptation inferred under the DS model. Finally, we used this comparison to illustrate and emphasize the conceptual differences between the two types of codon models.

In the following, a brief introduction about Rubisco and its role in C3 and C4 plants is given (section 3.1). Then, the comparative analysis of the results obtained using either the DS model or the codon model based on $d_N/d_S$ is presented. The content of this chapter, corresponding to the comparative analysis, is currently under minor revision for publication in PLOS ONE.

## 3.1 Rubisco and its evolution

Ribulose-1,5-biphosphate carboxylase/oxygenase, known as Rubisco, is the enzyme responsible for the first major step of carbon fixation, a process by which the inorganic carbon enters the metabolic pathways. This enzyme, which is possibly the most important enzyme for life on earth, catalyzes the carboxylation of ribulose 1,5 biphosphate (RuBP). It was discovered by Wildman and Bonner more than 70 years ago (Wildman and Bonner 1947).

Rubisco catalyzes the carboxylation and the oxygenation of RuBP within the same active site (Roy and Andrews 2000), so Rubisco is poorly able to distinguish between $CO_2$ and $O_2$. The product of the $CO_2$-demanding carboxylase reaction consists of two molecules of phospho-glycerate, which are the primary input to the Calvin cycle. On the other hand, the oxygenase activity produces one molecule of phosphoglycerate and one molecule of phosphoglycolate. The latter is the substrate for the photorespiration pathway, which is highly demanding in carbon and energy. Consequently, the efficiency of Rubisco can be decreased up to 40% in some severe conditions that are favorable to photorespiration. The slow catalytic rate of Rubisco and the competing photorespiration makes Rubisco inefficient, to some extent, for the first step of photosynthesis. Therefore, land plants assign up to 50% of their nitrogen to Rubisco, which makes it the most abundant enzyme on earth (Ellis 1979).

Rubisco is also one of the largest enzymes with the mass of 56 kDa. This multimeric enzyme consists of eight large and eight small subunits (figure 3-1). In land plants, the rbcL gene encodes the large, and rbcS the small subunits (C Dean, E Pichersky et al. 1989; Spreitzer 1993). To form

the holoenzyme, small subunits migrate to the chloroplast, where they are then assembled with large subunits (Roy and Andrews 2000).

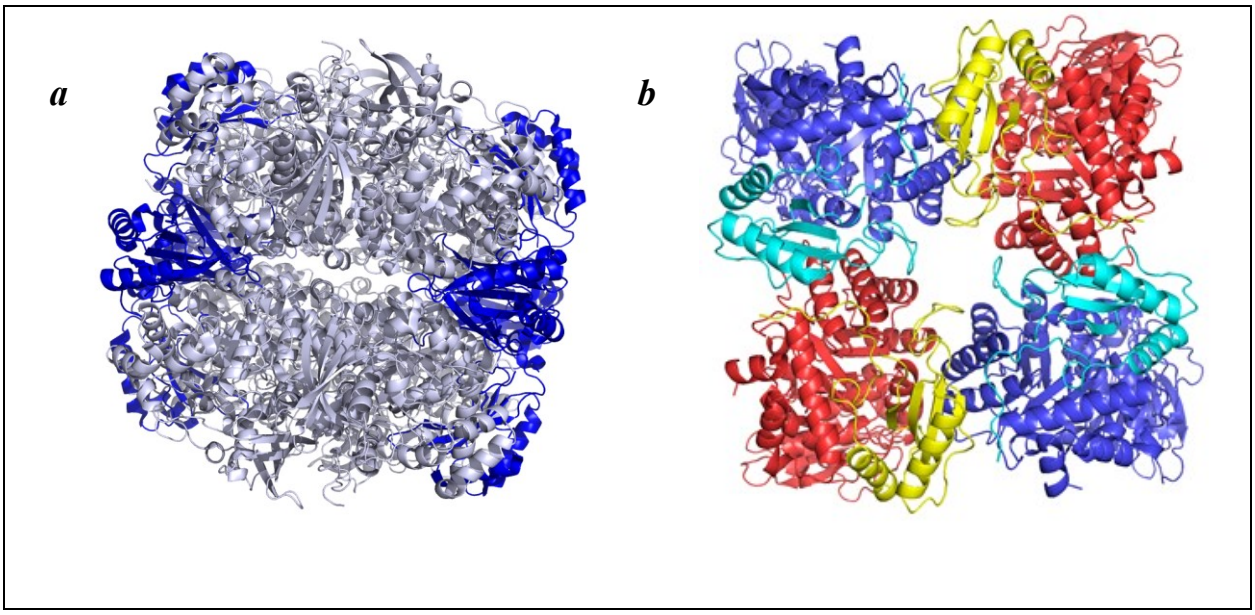



Figure 3-1. Rubisco holoenzyme. *(a)* The holoenzyme (L8S8) is composed of eight large subunits (pale blue) and eight small subunits (blue). *(b)* The L4S4 unit along the 4-fold symmetry axis. Large subunits are in blue and red and small subunits are in cyan and yellow. The images have been produced using PyMol (L DeLano 2002).

The evolutionary history of Rubisco goes back to approximately 3.5 billion years ago when the atmospheric $O_2$ concentration was low, and $CO_2$ was high (Lowe 1994). In this context, the poor discriminating ability of Rubisco between $O_2$ and $CO_2$ would have hardly restricted the performance of photosynthesis. After about 2 billion years, however, and probably because of the photosynthetic activity of cyanobacteria, the $O_2$ concentration began to increase. An unfortunate consequence of this change in the environment was the rise in Rubisco oxygenase activity, and thus, a decrease in the specificity of photosynthesis, through photorespiration. Photorespiration rate is further enhanced by warm temperature and aridity (Ehleringer, Sage et al. 1991), or by low atmospheric $CO_2$ concentration (Christin, Besnard et al. 2008). In such conditions, a considerable

selective pressure arose for C3 plants to increase the efficiency of $CO_2$ assimilation (Cowling 2001; Sage and Coleman 2001).

The C4 photosynthetic pathway, which emerged about 25-32 million years ago (Cerling 1999; Christin, Besnard et al. 2008), is an alternative solution to suppress photorespiration and increase $CO_2$ assimilation. It evolved as an adaptation to high temperature, intense light, and aridity (Gowik and Westhoff 2011), by concentrating $CO_2$ around the isolated Rubisco in the leaf. Hence, C4 plants dominate the grassland species in most tropical and subtropical regions (Edwards, Osborne et al. 2010). C4 eudicots, which appeared about 20 million years ago (Sage 2004), are primarily adapted to aridity more than to high temperature (Stowe and Teeri 1978), and they are not as frequent as C4 monocots. Molecular clock analysis shows that the evolution of C4 photosynthesis in eudicots has occurred more recently compared to monocots. The origin of C4 photosynthesis in *Amaranthaceae* family goes back to 8-11.5 million years ago (Kadereit, Borsch et al. 2003).

The evolution of C4 plants includes both anatomical and biochemical modifications in C3 plants. The kinetic of Rubisco has been altered in C4 plants, resulted in a higher efficiency (von Caemmerer and Quick 2000). As a result, less enzyme is needed within the leaf to assimilate the same amount of $CO_2$, which in turn leads to a higher PNUE (Photosynthetic nitrogen use efficiency) (Sage, Pearcy et al. 1987). Rubisco, in C4 plants, is located in bundle sheath cells (Monson and Rawsthorn 2000), In contrast, in C3 plants it is localized in mesophyll cells. The localization of Rubisco to bundle sheath cells is controlled by the rbcS gene (Nomura, Katayama et al. 2000). However, most of the changes in Rubisco kinetics from C3 to C4 are encoded by the large subunit (Hudson, Mahon et al. 1990). The main differences between C3 and C4 plants are listed in table 3-I.

The evolution of Rubisco rbcL sequence has attracted much attention in recent years (Christin, Salamin et al. 2008; Iida, Miyagi et al. 2009; Sen, Fares et al. 2011; Kapralov, Smith et al. 2012; Kapralov, Votintseva et al. 2013; Studer, Christin et al. 2014). Kapralov *et al.*(Kapralov and Filatov 2007) tried to identify positive selection in Rubisco using phylogenetic codon models (Yang 1997). They detected signatures of positive selection on rbcL in some photosynthetic organisms, especially in the main lineages of land plants, which they related to the transition between the C3 and C4 photosynthetic regimes (Kapralov and Filatov 2007). Studer *et al.* (Studer, Christin et al. 2014) investigated the effect of C3 and C4 Rubisco structural constraint on the evolution of the enzyme in monocots. They also identified a certain number of positions of the rbcL gene potentially affected by the C3/C4 transition and interpreted their findings in the context of a structural model of the enzyme.

Table 3-I. Differences between C3 and C4 plants.

| **C3 plants** | **C4 plants** |
|---|---|
| Leaves without Krans anatomy | Leaves with Krans anatomy |
| Chloroplast without peripheral reticulum | Chloroplast with peripheral reticulum |
| Chloroplast of one type (monomorphic), in mesophyll cells | Chloroplast of type dimorphic, one in mesophyll and one in bundle sheath cells |
| Complete photosynthesis in mesophyll cells | Initial fixation in mesophyll cells |
| Photosynthesis only when stomata are open | Photosynthesis at all time (stomata open or closed) |
| Less efficient in photosynthesis | More efficient in photosynthesis |

Interestingly, the analyses of Kapralov *et al.* and Studer *et al.* rely on fundamentally different types of models. Kapralov *et al. (*Kapralov, Smith et al. 2012) used classical codon models, formulated in terms of overall rates of synonymous and non-synonymous substitutions ($d_N/d_S$

based codon models, (Nielsen and Yang 1998; Yang and Nielsen 2002; Yang 2007)). In contrast, Studer *et al*. (Studer, Christin et al. 2014) used an amino acid replacement model (Tamuri, dos Reis et al. 2009), which has some common features with the Differential Selection model introduced here. These differences in the methodological approaches used in those previous articles prompted us to re-analyze the dataset of Kapralov *et al*. (Kapralov, Smith et al. 2012), under our model, but also under variants of the classical $d_N/d_S$ codon models which we re-implemented in the context of our Bayesian framework. This comparative analysis, presented in the next chapter, represents an interesting occasion to point out the conceptual differences between codon models and between adaptive regimes.

## 3.2 Molecular adaptation in Rubisco: discriminating between convergent evolution and positive selection using mechanistic and classical codon models

### 3.2.1 Abstract

Rubisco (Ribulose-1, 5-biphosphate carboxylase/oxygenase) is the most important enzyme on earth, catalyzing the first step of photosynthetic $CO_2$ fixation. So, without it, there would be no storing of the sun's energy in plants. Molecular adaptation of Rubisco to C4 photosynthetic pathway has drawn much attention. C4 plants, which comprise less than 5% of land plants, have evolved more efficient photosynthesis compared to C3 plants. Interestingly, a large number of independent transitions from C3 to C4 phenotype have occurred. Each time, the Rubisco enzyme has been subject to similar changes in selective pressure, thus providing an excellent model for

convergent evolution at the molecular level. Molecular adaptation is often identified with positive selection and is typically characterized by an elevated ratio of non-synonymous to synonymous substitution rate ($d_N/d_S$). However, convergent adaptation is expected to leave a different molecular signature, taking the form of repeated transitions toward identical or similar amino acids. Here, we used a previously introduced codon-based Differential Selection model to detect and quantify consistent patterns of convergent adaptation in Rubisco in eudicots. We further contrasted our results with those obtained by classical codon models based on the estimation of $d_N/d_S$. We found that the two classes of models tend to select distinct, although overlapping, sets of positions. This discrepancy in the results illustrates the conceptual difference between these models while emphasizing the need to better discriminate between qualitatively different selective regimes, by using a broader class of codon models than those currently considered in molecular evolutionary studies.

***Keywords:*** *Rubisco, C4 plants, convergent evolution, positive selection, codon models*

## 3.2.2 Introduction

Rubisco (Ribulose-1, 5-biphosphate carboxylase/oxygenase) is an enzyme that catalyzes the major step in carbon fixation in all photosynthetic organisms. It is the most abundant protein on earth (Ellis 1979), as it encompasses up to 50% of soluble proteins (Feller, Anders et al. 2008) and 20-30% of total nitrogen (Makino 2003) in C3 leaves. During carbon fixation, Rubisco reacts with both $CO_2$ and $O_2$ as its substrate, with poor distinguishing ability. The carboxylase activity results

in the incorporation of inorganic carbon into the metabolic C3 pathway, whereas the oxygenase activity boosts the photorespiration pathway. The latter prompts both energy consumption and $CO_2$ loss.

The evolution of C3 pathway goes back to 3 billion years ago when the atmosphere comprised high $CO_2$ and low $O_2$. In those conditions, photorespiration would have rarely happened. However, under present atmospheric conditions (lower $CO_2$ and higher $O_2$ concentration), photorespiration can represent a significant proportion of the enzymatic activity of Rubisco, such that the efficiency of photosynthesis can be dropped by 40% under unfavorable climates like hot and dry conditions (Ehleringer, Sage et al. 1991). As a result, some plants have developed an evolved improvement to C3 pathway called C4 photosynthesis as an adaptation to these changes in the environment (Sage, Christin et al. 2011; Liu, Sun et al. 2013).

About 85% of plants use C3 photosynthetic pathway, covering 78.4 million $km^2$ land area, whereas less than 5% are C4 plants, with global coverage of 18.8 million $km^2$ (Still, Berry et al. 2003; Simpson 2010). The rate of photosynthesis is different in these groups, being much more efficient in C4 plants than in C3 species. C4 photosynthesis mostly evolved as an adaptation to intense light, high temperature, and aridity (Gowik and Westhoff 2011). Hence, C4 plants dominate the grassland plants in harsh climates such as tropical, subtropical and warm regions (Edwards, Osborne et al. 2010).

The evolution of C4 plants from C3 ancestors consists of both anatomical and biochemical changes. These modifications allow C4 plants to concentrate more $CO_2$ around Rubisco, such that the oxygenase activity and the subsequent photorespiration are partially or entirely repressed. The kinetics of Rubisco has been altered in C4 plants, leading to lower specificity and higher efficiency (Jordan and Ogren 1981; Andrews 1987; von Caemmerer and Quick 2000).

Interestingly, a relatively large number of independent transitions from C3 to C4 phenotype have occurred across monocots and eudicots. Each time, the Rubisco enzyme has been subject to similar selective pressure for tuning the tradeoff between substrate specificity and yield. As a consequence, C4 photosynthesis is an excellent model for convergent evolution at the molecular level in response to environmental changes (Russell K. Monson 2003). Regarding applications, finding features of C4 plants and applying them to C3 plants such as rice, can be potentially used to increase crop yields (Taniguchi, Ohkawa et al. 2008; von Caemmerer and Evans 2010). Considering the above issues, understanding how selection acts on Rubisco in C4 plants compared to C3 ancestors can be very beneficial.

Based on these considerations, the evolution of Rubisco has attracted substantial attention in recent years (Kapralov and Filatov 2007; Christin, Salamin et al. 2008; Iida, Miyagi et al. 2009; Sen, Fares et al. 2011; Kapralov, Smith et al. 2012; Kapralov, Votintseva et al. 2013; Studer, Christin et al. 2014). Kapralov *et al.* (Kapralov and Filatov 2007) tried to identify positive selection in Rubisco using molecular phylogenetic analyses. Employing codon models that allow for varying selection among sites (implemented in codeML (Yang 1997)), they detected sites under positive selection in some photosynthetic organisms, especially in the main lineages of land plants. More recently, Kapralov *et al.* (Kapralov, Smith et al. 2012) used a similar method to investigate the evolution of Rubisco in C4 plants in a large group of C4 eudicots and found sites under positive selection. They observed that some of those positively selected sites appear to display consistent patterns of amino acid substitutions associated with the C3 to C4 transitions.

These empirical analyses raise an interesting question, concerning the use of codon models to characterize selective regimes in protein-coding sequences. Typically, elevated $d_N/d_S$ results from ongoing adaptive processes, by which a protein-coding gene is constantly challenged by ever-

changing selective forces. However, in its general form, this process of ongoing adaptation is not associated to repeated transitions toward the same amino acid at a given position, independently across multiple lineages, and could instead continually elicit new amino acids at positively selected sites. In contrast, the multiple transitions between C3 and C4 photosynthetic regimes represent a case of *convergent* evolution. At the molecular level, this is expected to result in recurrent *directional* selection, thus, potentially favoring the same amino acid(s) at the same site(s) upon each C3 to C4 transition. In addition, the overall $d_N/d_S$ induced by this process of recurrent directional selection is fundamentally determined by the rate of C3/C4 transitions across the phylogeny, which may not be sufficiently high to induce a $d_N/d_S$ greater than one at those positions that are susceptible to respond to this convergent evolutionary process. Thus, positive selection, like what is formalized by classical codon models (i.e., by an elevated $d_N/d_S$), may not be the most appropriate selective regime to test in the present case. A similar distinction between episodic diversifying and directional selection has been previously proposed by Murrel *et al.* (Murrell, de Oliveira et al. 2012). They demonstrated that modeling the episodic and directional selection explicitly enhance the accuracy to identify drug-resistant sites in HIV-1.

Convergent amino acid substitutions which potentially linked to adaptation to the C4 phenotype have been more directly investigated by Studer *et al.* (Studer, Christin et al. 2014). These authors used the TDG09 model, allowing for site- and condition-specific amino acid preferences (Tamuri, dos Reis et al. 2009), to identify sites under condition-dependent selection. Recently, we have developed an approach similar to the TDG09 model, called Differential Selection (DS) model (Parto and Lartillot 2017) using a Bayesian mechanistic derivation of the codon substitution process, under the so-called mutation-selection formalism. Here, we re-assessed the question of positive versus convergent selective patterns in the Rubisco gene in

eudicots, using two types of codon models: first, we applied our DS model to identify amino acids which are differentially selected at specific positions along the Rubisco sequence, as a function of the photosynthesis pathway. Second, we implemented Bayesian versions of the classical $d_N/d_S$ -based codon models, allowing for both site- and branch-specific modulations of the $d_N/d_S$ ratio (Yang and Nielsen 2002), and applied them to the Rubisco dataset. We found that the two classes of models tend to select distinct, although overlapping, sets of positions. Altogether, our analysis emphasizes the existence of qualitatively different adaptive regimes undergone by protein-coding genes, and the need to better discriminate between these distinct regimes by using a broader class of codon models than those currently considered in molecular evolutionary studies.

## 3.2.3 Materials and Methods

### Sequence data, phylogenetic tree, and partitioning scheme

We obtained the *Amaranthaceae* rbcL multiple sequence alignment and the original phylogenetic tree (figure 3-2) from Kapralov *et al.* (Kapralov, Smith et al. 2012). The dataset consists of 179 rbcL sequences of length 1341 base pairs, corresponding to amino acids 22-468 (the first 21 coding positions are missing). Out of 179 sequences, 84 and 95 sequences belong to C4 and C3 species, respectively. List of these species and their photosynthetic type is provided in Appendix E.

The phylogenetic tree was partitioned according to two alternative schemes, with $K=3$ or $K=2$ distinct conditions, based on the type of the photosynthetic pathway. In the three-condition scheme, the largest monophyletic clades exclusively composed of C3 or C4 species were first

identified and defined as conditions 1 and 2. The branches at the base of each C3 and C4 clades were also included in conditions 1 and 2, respectively. All other branches outside from these clades (reconstructed ancestral branches) were considered as belonging to condition 0. The model that employs this approach is called DS3, and its phylogenetic tree is illustrated in figure 3-2. The two-condition setup (model DS2) differs from the three-condition scheme (model DS3) by allocating all branches outside of the C4 monophyletic clades (together with their basal branches) by default to the C3 condition. Model DS2 amounts to assuming a maximum-parsimony reconstruction of the evolution of the photosynthetic regime, under the assumption that evolutionary transitions are exclusively from C3 to C4, with no reversion back to C3 (Christin, Freckleton et al. 2010). However, model DS2 statistically implies a comparison between two conditions that are unevenly represented along the phylogeny, both in terms of total number of branches (152 for C4 versus 203 for C3) and concerning the evolutionary depth (the DS3 condition is mostly represented by recent branches, while the DS2 condition encompasses both ancient and recent lineages). In this respect, the advantage of model DS3 is to balance the empirical signal between the two conditions of interest (C3 and C4, represented by 158 and 153 branches under the DS3 model), and to focus exclusively on recent branches of the phylogeny for both conditions.

## Differential Selection model

The principles of the Differential Selection model were introduced previously (Parto and Lartillot 2017), and we only recall the general structure here. We used mutation-selection formalism, as in Halpern and Bruno (Halpern and Bruno 1998) or Rodrigue *et al.* (Rodrigue, Philippe et al. 2010). According to this formalism, the substitution rates between codons were

derived from first principles of population genetics, regarding mutation rates and selective effects. The latter was explicitly modeled and assumed to operate exclusively at the level of the amino acid sequence.

More specifically, consider a sequence of $N$ coding positions ($3N$ nucleotide positions). The number of conditions across the phylogenetic tree is denoted as $K$ ($K=2$ or $K=3$, depending on the partition scheme). The mutation process is assumed to be time-reversible and homogeneous among sites and across lineages. It is thus entirely characterized by a general time-reversible 4×4 matrix $Q$. In contrast, the selective forces acting at the amino acid level are both condition- and position-specific. Accordingly, for each position $i \in [1, N]$ and each condition $k \in [1, K]$, we introduced an array of 20 non-negative fitness factors $F^{ik} = (F_a^{ik})_{a\in[1,20]}$, one for each amino acid. In the following, these 20-dimensional vectors will be referred to as amino acid *fitness profiles*. In the present version of the model, they are assumed to be random effects across sites and conditions, drawn *iid* from a uniform Dirichlet distribution.

Once these mutation rates and fitness factors are specified, the substitution process can be defined as follows. Consider the substitution rate between codon $c_1$ to $c_2$ (encoding amino acids $a_1$ and $a_2$) at site $i$, and condition $k$, where codons $c_1$ and $c_2$ are assumed to vary only at one nucleotide position, with respective nucleotide states $n_1$ and $n_2$ at that position. First, we defined a Darwinian scaled selection coefficient, associated with a mutation from wild-type codon $c_1$ to mutant codon $c_2$. Since selection is assumed to act only at the level of the amino acid sequence, this scaled selection coefficient is given by

$$S_{a_1 a_2}^{ik} = ln\left(\frac{F_{a_2}^{ik}}{F_{a_2}^{ik}}\right) \qquad (3\text{-}1)$$

Then, the rate of substitution between codon $c_1$ and $c_2$ is given by the product of the mutation rate and the relative fixation probability $P$ (i.e., relative to neutral). This fixation probability is itself dependent on the scaled selection coefficient just defined. Considering the classical diffusion approximation, we can express this relative fixation probability as

$$P_{fix} = \frac{S_{a_1 a_2}^{ik}}{1 - e^{-S_{a_1 a_2}^{ik}}} \qquad (3\text{-}2)$$

Thus, finally, the rate of substitution between codons $c_1$ and $c_2$ at position $i$ and under condition $k$ is given by

$$\boldsymbol{R_{c_1 c_2}^{ik}} = \begin{cases} \boldsymbol{Q_{n_1 n_2}} & \textbf{Synonymous} \\ \boldsymbol{Q_{n_1 n_2}} \times \dfrac{\boldsymbol{S_{a_1 a_2}^{ik}}}{\boldsymbol{1 - e^{-S_{a_1 a_2}^{ik}}}} & \textbf{Non} - \textbf{synonymous} \\ \boldsymbol{0} & \boldsymbol{c_1}\ \textbf{and}\ \boldsymbol{c_2}\ \textbf{differ at more than one site} \end{cases} \qquad (3\text{-}3)$$

## Omega-based codon model

As an alternative to mutation-selection models, one of the most well-known and widely used methods for characterizing the selective regimes, involved in the evolution of protein-coding genes, is to estimate the ratio of non-synonymous ($d_N$) to synonymous ($d_S$) substitution rate ($d_N/d_S$), denoted as ω. These omega-based models were first proposed by Goldman and Yang (Goldman and Yang 1994) and Muse and Gaut (Muse and Gaut 1994), and subsequently complexified to

account for site- and branch-specific modulations of the $d_N/d_S$ ratio (Nielsen and Yang 1998; Yang 1998; Anisimova, Bielawski et al. 2001; Yang and Nielsen 2002).

Here, we used the Muse and Gaut formalism, and proposed a Bayesian model allowing for site- and condition-specific modulations of $\omega = d_N/d_S$. According to this model, the instantaneous substitution rate from codon $c_1$ to $c_2$ at site $i$ and condition $k$ is specified as follows

$$R^{ik}_{c_1c_2} = \begin{cases} Q_{n_1n_2} & \textbf{Synonymous} \\ Q_{n_1n_2} \times \omega^{ik} & \textbf{Non} - \textbf{synonymous} \\ 0 & c_1 \textbf{ and } c_2 \textbf{ differ at more than one site} \end{cases} \quad (3\text{-}4)$$

Here, $\omega^{ik}$ is thus the $d_N/d_S$ ratio for site $i$ and under condition $k$. For each condition $k$, the $\omega^{ik}$ s, for $i \in [1, N]$ are modeled as random effects across sites, drawn *iid* from a gamma distribution of shape and scale parameters $\alpha^k$ and $\beta^k$.

We considered two alternative versions of this omega-based model: in model OM1, we assumed only one condition, thus defining a single (global) value of $\omega^i$ across the whole phylogenetic tree for site $i$; in model OM3, on the other hand, the tree is partitioned into three conditions according to the photosynthesis pathways, precisely as for model DS3 above, and a distinct value $\omega^{ik}$ is allowed for site $i$ and under condition $k \in [1, 3]$.

## Priors

In all analyses presented below, the topology ($\tau$) of the tree is fixed. For all models, the prior on branch lengths is a product of independent Exponentials of mean $\lambda$; the hyperparameter $\lambda$ is from an Exponential distribution of mean 0.1; the prior on relative exchangeabilities of the mutation process is a product of Exponentials of mean 1; the prior on the mutational equilibrium frequency vector is a uniform Dirichlet distribution. As mentioned above, under the DS2 and DS3

models, the site- and condition-specific fitness profiles, $F^{ik}_{a_1 a_2}$, are random effects integrated over a Dirichlet distribution. Concerning the OM1 and OM3 models, the site- and condition-specific $d_N/d_S$ values ($\omega^{ik}$) are random effects integrated over a gamma distribution of shape and scale parameters $\alpha^k$ and $\beta^k$, which are themselves drawn from an exponential prior of mean 1 for each $k \in [1, K]$.

## MCMC sampling

To sample the parameters from their joint posterior distribution, we used the general MCMC approach previously described in (Lartillot 2006; Lartillot and Poujol 2011; Parto and Lartillot 2017). This approach consists of an alternation between stochastic mapping of the detailed substitution history at each coding site, followed by a long series of Metropolis-Hastings updates of all parameters and all random effects across sites and across conditions, conditional on this stochastic mapping.

Two independent MCMC were run for each analysis. In all cases, burn-in was first estimated visually, and then convergence and mixing were quantified using the *tracecomp* program (from the Phylobayes suite (Lartillot, Lepage et al. 2009) to compare the samples obtained under independent runs. *Tracecomp* gives an estimate of the discrepancy between the two runs, as well as the effective sample size, for several key parameters and statistics of interest. In the present case, the minimum effective size was always higher than 3000 and the discrepancy less than 0.2 for most statistics. Finally, the reproducibility of the estimation of the posterior mean differential selection factors across all amino acids and all sites was verified by plotting the estimates for all amino acids and all sites across the two independent runs (Appendix F). After 400 points of burn-

in from a total of almost 6000 points have been removed, posterior estimates were obtained by averaging over the remaining of the MCMC run.

## Post-analysis

Under the DS models, for a given configuration of the model (typically drawn from the posterior distribution by MCMC), Differential Selection between two conditions C3 and C4 is simply calculated as the log-ratio between the amino acid fitness profiles ascribed to conditions 1 and 2

$$\boldsymbol{D^{i}_{a_1 a_2} = ln\left(\frac{F^{i2}_{a_1 a_2}}{F^{i1}_{a_1 a_2}}\right)} \qquad (3\text{-}5)$$

These arrays of 20 differential selection effects (for the 20 amino acids) at each position are then averaged over the posterior distribution by MCMC. A position is deemed to show strong statistical support for a differential effect in favor of amino acid $a_2$ (in condition C4) over amino acid $a_1$ (in condition C3) if the posterior probability that $D^{i}_{a_1 a_2} > 0$ is greater than 0.90. Conversely, strong support for a negative differential effect (i.e., a differential effect against $a_2$ in favor of $a_1$) is considered whenever the posterior probability that $D^{i}_{a_1 a_2} < 0$ is higher than 0.90.

Under the OM models, the posterior mean value of site- and condition-specific $d_N/d_S$ is reported. Position *i* is regarded to have a strong support for positive selection under condition *k* if the posterior probability that $\omega^{ik} > 1$ is greater than 0.90.

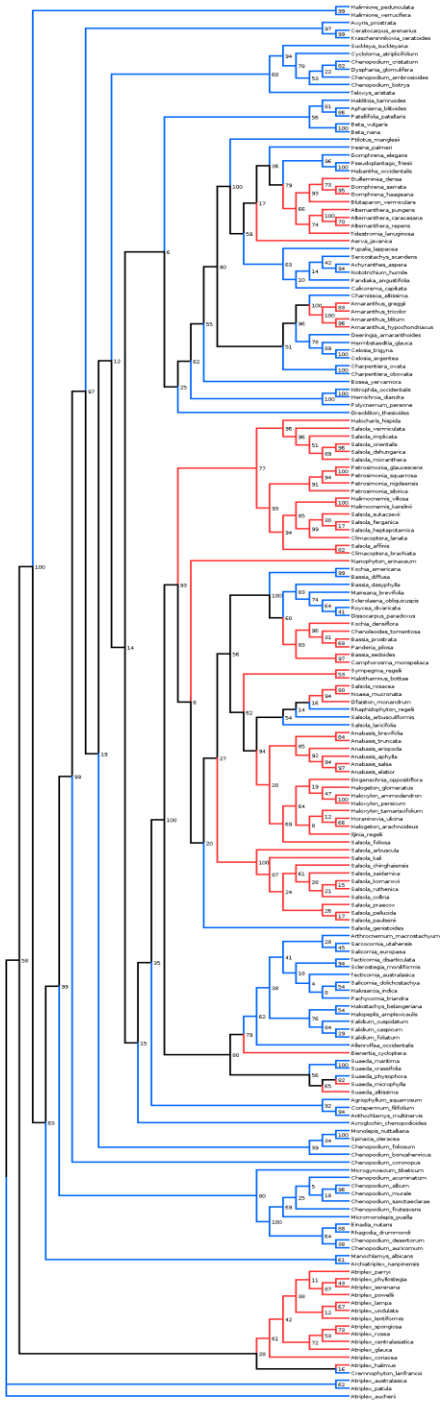


Figure 3-2. Phylogenetic tree of 179 rbcL sequences from *Amaranthaceae* family. The tree partitioned (according to model DS3) in C3 (blue), C4 (red) and interior branches (black). The number on each branch is the bootstrap support (provided by (Kapralov, Smith et al. 2012)). The tree is visualized using Dendroscope program (Huson and Scornavacca 2012).

## 3.2.4 Results and Discussion

*Amaranthaceae* is one of the plant families with the largest number of C4 species. This makes it a suitable case for Differential Selection (DS) analysis. Based on a multiple sequence alignment of rbcL genes and an annotated phylogenetic tree of *Amaranthaceae*, our DS model captures site-specific amino acid preferences as vectors of 20 fitness factors (for the 20 amino acids) under each condition. Then, contrasting for each position, the fitness factors estimated in the two conditions of interest (here, in the C3 and C4 regimes), allows us to identify positions for which the fitness of a specific amino acid has undergone a significant change, either upward or downward, associated with the transition between the C3 and the C4 photosynthetic regime (see Methods).

### DS2 versus DS3: finding an optimal contrast between C3 and C4

As described in the methods, the tree was divided into either two or three conditions, resulting in two version of the Differential Selection codon model, referred to as DS2 and DS3. In DS2, the interior branches (black branches in figure 3-2) which connect C3 and C4 clusters are defined as C3. This corresponds to a plausible reconstruction of ancestral photosynthetic regimes across the group, as no reversal from C4 to C3 is known (Christin, Freckleton et al. 2010).

The selection profiles at position 306-331 estimated by DS2 are shown in figure 3-3. In this figure, we use a graphical logo representation (Schneider and Stephens 1990) to display both absolute (global) and differential fitness distributions. Absolute logos for the reference condition represent the fitness of amino acids under a specific condition, with the height of the letter being proportional to the fitness of the corresponding amino acid. Differential logos, on the other hand, represent the difference in log-fitness between two conditions: letters above (resp. below) the

baseline corresponds to amino acids whose fitness is increased (resp. decreased) in each condition, compared to its parent condition.

The global selection profile (figure 3-3-a) captures the absolute amino acid fitness for C3 plants. This profile primarily reflects the strong conservation of the protein sequence, with one single amino acid overwhelmingly favored at most positions. The differential profile between C4 and C3 (figure 3-3-b) shows interesting patterns of opposite selective effects concerning pairs of amino acids, specifically at positions 309, 315 and 328. However, the differential profile between the C4 and C3 is also characterized by an inferred background of apparently non-specific differential selective effects concerning all primary amino acids represented in the absolute fitness profile under C3: essentially, the absolute profile under C3 displays the consensus sequence of the alignment, while the differential profile between C4 and C3 reproduces this consensus sequence, although now at the bottom. This is likely to be a statistical artifact, which might have two alternative explanations. The first one is the possible existence of non-fixed polymorphic states in the multiple sequence alignment. These mutations, whose fate is to be ultimately removed by purifying selection, are expected to be mapped specifically along the terminal branches of the phylogeny and may thus contribute to an apparent decrease in the inferred fitness of ancestral amino acids in the condition that is most enriched in terminal branches (here, C4). Another possible explanation is that the number of branches allocated to the C4 condition is smaller than that allocated to the C3 condition, potentially leading to a difference in statistical power between the two conditions. As a result, and in the presence of shrinkage mediated by the prior, the fitness of conserved amino acids is inferred to be higher in that condition that is endowed with the largest number of branches (here, C3).

One way to avoid this artifact is to balance the signal between condition C3 and C4, by allocating the interior branches of the tree to another baseline condition and by restricting the inference of C3-specific selection to the monophyletic groups of C3 species. In the present case, there are a comparable number of C3 and C4 branches (about 150 for each). This new setting (model DS3) is therefore expected to result in a much more balanced assessment of the Differential Selection effects between recent C3 and C4 lineages.

Indeed, and unlike the differential profile between C3 and C4 provided by the DS2 model (figure 3-3), the differential profile given by DS3 between recent C3 and C4 lineages (figure 3-4) appears to have more reasonable properties: sparse, selecting a small number of positions for which specific amino acids appear to be differentially selected between the two photosynthetic regimes, and balanced between positive and negative effects (at the top and bottom, respectively). For instance, at position 309, the fitness of Methionine is substantially decreased in C4 plants, compared to C3 species (pp =0.93). Correlatively, the fitness of Isoleucine is increased at that position (pp = 0.87). Similarly, residue 328 is identified by the DS3 model as a position of the rbcL gene under the highest differential selection effect between C3 and C4 *Amaranthaceae* species. At site 328, Alanine is globally preferred in *Amaranthaceae* eudicots, yet in C4 group, its fitness is significantly decreased (pp =0.99) in favor of Serine, whose fitness is increased compared to what prevails in C3 lineages (pp=0.96). Based on these observations, in the following, we conduct all Differential Selection analyses under the DS3 model. The complete C4/C3 differential logo, for the whole sequence alignment, is displayed in Appendix G.

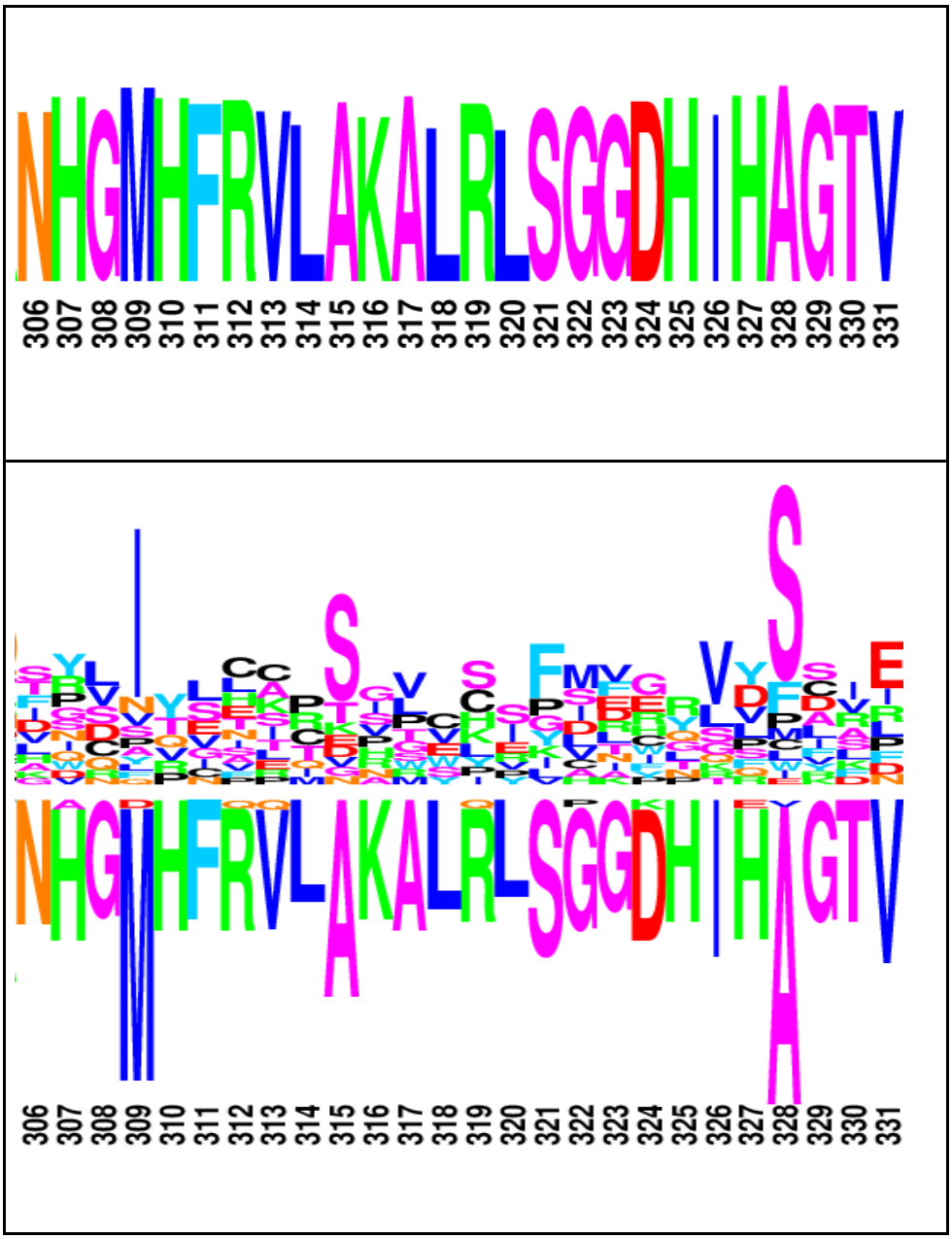



Figure 3-3. Global and Differential Selection profiles for position 306-331, by model DS2. *(a)* Amino acid fitness for C3 plants. *(b)* Differential amino acid fitness for C4 plants. Amino acids at the top (bottom) show an increase (decrease) in fitness compared to *(a)*.

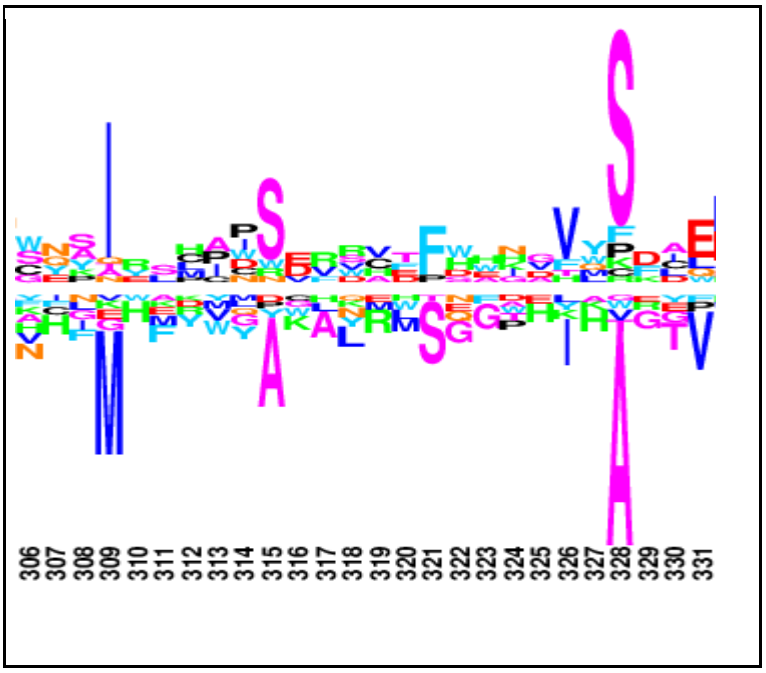


Figure 3-4. C4/C3 differential selection profile for position 309-328, under the DS3 model.

## Differential Selection (DS) versus omega-based (OM) codon models

In addition to DS3, which belongs to the family of mutation-selection codon models (Halpern and Bruno style (Halpern and Bruno 1998)), the *Amaranthaceae* dataset was also analyzed under two omega-based models (Muse and Gaut (Muse and Gaut 1994) style). The first of these models (OM1) is a site-specific model: each site has its own value for $\omega = d_N/d_S$, all of which are modeled as site-specific gamma-distributed random effects. By selecting sites with a high posterior probability of having a value of $d_N/d_S$ greater than 1, model OM1 allows for the detection of sites under positive selection globally across *Amaranthaceae* dataset. The second model (OM3) allows for independent values of $d_N/d_S$, simultaneously across sites and conditions. Conditions are defined as in DS3 model (internal branches, as well as terminal C3 and C4 clades). Thus, OM3, unlike OM1, allows for the detection of sites under positive selection specifically in C3 or C4 species. To

facilitate the comparison, all three models, DS3, OM1 and OM3, were implemented in a Bayesian framework, using similar strategies for designing the models (both $d_N/d_S$ and differential selective effects modeled as either global or condition-dependent *iid* random effects across sites) and for detecting significant effects (based on the posterior probability for a site to have a value of $\omega > 1$ or a differential selective effect greater and smaller than 0, globally or in a given condition). The results of these analyses are summarized in table 3-II. In this table, all sites for which a firm support ($pp > 0.90$) was found under at least one of the three models are reported.

Under model OM1, 6 positions (32, 43, 145, 225, 262 and 279) were found to have a $d_N/d_S > 1$ with a posterior probability greater than 0.90, and 2 positions (439 and 443) with $pp < 0.90$. These eight positions are exactly those reported by Kapralov *et al.* (Kapralov, Smith et al. 2012), found using the BEB approach implemented in CodeML. Note that the approach used here and the one implemented in CodeML are rather different in their statistical strategy for detecting sites under positive selection. The approach of CodeML relies on a mixture model, whereas the present approach explicitly assumes independent values of $d_N/d_S$ across sites. The results obtained here, therefore, suggest that at least in the present context, the details of the overall statistical strategy do not have a strong influence on the outcome of the model.

Kapralov *et al.* (Kapralov, Smith et al. 2012) also reported an additional two sites, 281 and 309, detected by the *branch-site* model and thus inferred to be under positive selection specifically in C4 condition. Here, using the OM3 model, which allows for condition- and site-specific values for $d_N/d_S$, we found statistical support for positive selection in C4 only for position 281. For position 309, the posterior mean value of omega in condition C4 is indeed greater than 1 (1.08), although only with a weak posterior probability support ($pp = 0.47$). Conversely, it is worth noting

that several sites (such as 32, 43, 225, and 26 2) are inferred by model OM3 to be under positive selection only under C3, but not under C4.

Finally, the DS3 model uncovers a series of 11 sites under Differential Selection between C3 and C4 with pp > 0.90. These 11 sites include 4 of the sites discovered by model OM1, thus inferred to be under global positive selection (32, 225, 262 and 443), as well as sites 281 and 309 (inferred to be under positive selection specifically in C4, either by model OM3 or by *branch-site* models of CodeML). Conversely, and importantly, half of the discoveries made by the DS3 model (6 sites out of 11, including site 309) do not show any signal of positive selection under either OM1 or OM3.

## Differential Selection patterns in *Amaranthaceae* family

Here we studied the molecular adaptations associated with the C3 to C4 transitions in *Amaranthaceae* eudicots. Using a mechanistic codon model for detecting differential selection patterns associated with these adaptations, we found 11 positions to be under Differential Selection pressure between C3 and C4 eudicots. Some of the amino acid substitutions undergone by these positions have a conformational or catalytic role in Rubisco enzyme in C4 plants, leading to its higher efficiency (van Lun, van der Spoel et al. 2011; Studer, Christin et al. 2014). Alternatively, they might be a compensatory mutation selected to maintain its optimized function.

Table 3-II. Findings of OM1, OM3 and DS3 model. Only positions with posterior probability>0.9 in any of the above models are reported here. Positions specified with an asterisk are those found previously by

Kapralov *et al.* (Kapralov, Smith et al. 2012) (one for C3 and two for C4). $\omega^1$, $\omega^2$, and $\omega^3$ represent ω values for condition 1, 2 (C3) and 3 (C4).

| Position | OM1 model | | OM3 model | | | | | | DS3 model | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ω | pp(ω>1) | $\omega^3$ | pp($\omega^3$>1) | $\omega^2$ | pp($\omega^2$>1) | $\omega^1$ | pp($\omega^1$>1) | Amino acid | pp |
| 32* | 3.2 | 0.99 | 1.32 | 0.61 | 3.68 | 0.99 | 3.76 | 0.97 | Q, L, K | 0.93, 0.91, 0.81 |
| 43* | 2.13 | 0.99 | 1.03 | 0.48 | 2.32 | 0.98 | 2.93 | 0.93 | - | - |
| 86 | 0.71 | 0.15 | 0.02 | 0 | 1.23 | 0.65 | 0.48 | 0.1 | H, N | 0.92, 0.89 |
| 143 | 0.55 | 0.05 | 0.01 | 0 | 0.9 | 0.34 | 0.43 | 0.14 | S, A | 0.94, 0.77 |
| 145* | 2.65 | 0.99 | 3.4 | 0.99 | 2.14 | 0.96 | 0.06 | 0.01 | L | 0.75 |
| 225* | 2.53 | 0.99 | 1.27 | 0.58 | 2.7 | 0.99 | 3.55 | 0.98 | L, I | 0.96, 0.88 |
| 262* | 2.25 | 0.99 | 0.33 | 0.05 | 3.61 | 0.99 | 1.69 | 0.64 | V, A | 0.99, 0.75 |
| 279* | 2.19 | 0.99 | 2.21 | 0.98 | 2.13 | 0.98 | 1.28 | 0.51 | - | - |
| 281** | 1.11 | 0.62 | 2.44 | 0.99 | 0.33 | 0.02 | 0.28 | 0 | A | 0.96 |
| 309** | 0.4 | 0.006 | 1.08 | 0.47 | 0.01 | 0 | 0.02 | 0 | M, I | 0.94, 0.87 |
| 328 | 1.35 | 0.8 | 1.77 | 0.89 | 0.87 | 0.3 | 0.56 | 0.2 | A, S | 0.99, 0.98 |
| 354 | 0.77 | 0.21 | 0.02 | 0 | 1.43 | 0.72 | 0.03 | 0 | T, I | 0.92, 0.89 |
| 439* | 1.15 | 0.63 | 0.15 | 0.01 | 2.05 | 0.98 | 0.26 | 0.08 | T, R | 0.89, 0.84 |
| 443* | 1.5 | 0.88 | 0.01 | 0 | 2.28 | 0.99 | 1.79 | 0.72 | T, A | 0.97, 0.77 |
| 461 | 1.36 | 0.79 | 0.09 | 0.01 | 1.65 | 0.85 | 2.93 | 0.93 | V, I | 0.98, 0.86 |

For instance, residue 328, with the highest differential effect, locates in the active loop 6 of the enzyme. Replacement of hydrophobic A with polar S destabilizes the active site, which leads to more flexibility of its opening and closing (Andersson and Backlund 2008) and might explain the higher efficiency of C4 plants. Site 281 lies in the core of C-terminal domain, and it may have a long-range effect on active loop 6 (Studer, Christin et al. 2014). Position 309 is in the interface of C-terminal domains of two subunits within a dimer (Studer, Christin et al. 2014), which might affect flexibility. Although residues 86, 354 and 461 are found to be under strong Differential Selection pressure between C3 and C4 *Amaranthaceae*; their exact role has not been specified. Position 461 locates near a large subunit residue (residue 466) which might account for the interaction with Rubisco activase.

## Comparing Differential Selection and omega-based codon models

Previously, some positions have been found by other phylogenetic methods to be under specific selective regimes, potentially associated with the C3 to C4 transitions. In particular, Kapralov *et al.* (Kapralov, Smith et al. 2012) used the concept of $d_N/d_S$ as selection strength along the coding sequence. Using classic $d_N/d_S$ codon models, they uncovered a set of 10 positions putatively under positive selection, either globally over the tree (8 positions) or explicitly in the C4 groups (2 positions). To further explore this point, we implemented new $d_N/d_S$ codon models, allowing for site- and condition-specific $d_N/d_S$, in our Bayesian framework. Selecting sites based on the posterior probability support for $d_N/d_S > 1$, we essentially recovered the same set of positions as that reported by Kapralov *et al.* (Kapralov, Smith et al. 2012) (except for one position). On the other hand, if we compare the set of findings under $d_N/d_S$ models and the Differential Selection

model, we observe a partial overlap. Specifically, only half of the positions inferred to be under Differential Selection between C3 and C4 were also found by $d_N/d_S$ models. Conversely, 4 of the 10 findings under both classes of $d_N/d_S$ models showed differential selection effects.

The partial overlap between the findings of omega-based and Differential Selection models illustrates the conceptual difference between these models and the fact that they are meant to capture fundamentally different selective patterns. Classic omega-based codon models are meant to detect an overall *acceleration* of the rate of non-synonymous substitution. Such accelerations are typically caused by *ongoing adaptation*, due to diversifying selection, ecological red-queens, or fluctuating selection caused by environmental changes. In contrast, Differential Selection models are intended to capture convergent patterns of *directional selection* associated with a specific change in the environment, having occurred several times independently across the phylogeny.

These two classes of selective patterns are not quite mutually exclusive. In principle, recurrent substitution events due to directional selection caused by repeated transitions from C3 to C4 photosynthesis across the *Amaranthaceae* family could result in an overall increase in the $d_N/d_S$ observed at the corresponding sites. However, if the rate of C3 to C4 transitions is not sufficiently high, the resulting increase in $d_N/d_S$ may not be enough to lead to a situation where $d_N/d_S > 1$. As a result, some of the critical condition-specific adaptations might be missed by $d_N/d_S$ codon models. For instance, as illustrated here, positions 86, 143 or 354, which show a strong Differential Selection effect, yet have a $d_N/d_S$ not exceeding 1.

In addition, this phenomenon of recurrent directional selection linked to repeated C3 to C4 transitions cannot explain that most of the sites inferred to be under positive selection have a $d_N/d_S > 1$ globally over the tree, and often within the C3 terminal clades (e.g. positions 32, 43 and 279),

in which no such substitution event induced by C3 to C4 transition is supposed to have occurred. Concerning positions 43 and 279, for instance, no differential selection effect is detected by the DS3 model, while the $d_N/d_S$ is inferred to be of the order of 2, including within terminal C3 clades. Thus, the most likely explanation for the pattern of Darwinian evolution at those sites is solely the presence of ongoing adaptation that would not be directly related to the repeated transitions between C3 and C4 photosynthetic regimes.

Conversely, we observed some positions (in particular 262 and 461) that show a differential selection effect between C3 and C4, combined with a pattern of positive selection over the tree, *except* in the C4 condition, in which the $d_N/d_S$ is specifically and markedly decreased (posterior mean $d_N/d_S = 0.33$ and 0.09, respectively). A possible explanation for this pattern is that, in C3 species, those positions are available for ongoing adaptation to a constantly fluctuating environment, but the transition to C4 photosynthesis essentially locks those positions into more specific adaptive amino acid states, thereby stopping the flux of adaptive substitutions at those sites. Of note, this concurrence of positive selection and differential selection effects (e.g., positions 32, 225) is not so easily explained in the context of the mutation-selection modeling framework used here. Mutation-selection models predict that the $d_N/d_S$ should always be less than one at mutation-selection balance (Spielman and Wilke 2015). In the present case, this means that the DS model does not predict $d_N/d_S$ greater than 1, except possibly during the transient phases following a change between the C3 and the C4 regimes -- thus, at any rate, not within the C3 clades.

### 3.2.5 Conclusions

Rubisco has long been known to be under positive selection (Kapralov and Filatov 2007). In addition, it has been shown that Rubisco has been evolved in different structural forms and functions (Tabita, Hanson et al. 2007). It exemplifies a convergent evolution of enzyme properties in its phylogenetic pathway. One example of this convergent evolution happens between C3 and C4 plants through crossing the fitness landscape (Sage 2002; Kapralov, Smith et al. 2012; Studer, Christin et al. 2014). Therefore, the complex molecular evolutionary patterns displayed by the Rubisco gene in eudicots represent an exciting case-study for assessing and comparing current codon modeling strategies (Kapralov, Smith et al. 2012). In this respect, our comparative analysis, by making an inventory of the amino acid-positions in rbcL sequences that are positively or differentially selected in C3 and C4 *Amaranthaceae* family, emphasizes the fundamental difference, in scope and meaning, between the two main classes of models currently considered in the literature: on one side, classic codon models based on the measure of the overall $d_N/d_S$, whose focus is primarily on positive selection; and on the other side, Differential Selection models, whose aim is instead, to detect convergent patterns of directional selection associated with repeated transitions between known evolutionary regimes. Our analysis also emphasizes that none of the models considered here, either omega-based models or mutation-selection approaches, offers an absolute satisfactory explanation of the complex patterns of molecular evolution observed in *Amaranthaceae*, and probably also present in other species groups -- thus suggesting that further developments are still needed on the front of phylogenetic codon-models.

# CHAPTER 4: CONCLUSIONS AND FUTURE DIRECTIONS

The main objective of this thesis has been to develop a phylogenetic model for detecting and characterizing patterns of convergent evolution at the molecular level. This codon-based model, which we called Differential Selection (DS) model, is formulated in a Bayesian MCMC framework in a C++ object-oriented environment. Convergent evolution can be formalized in terms of repeated events of directional selection, induced by recurrent switches between small numbers of known environments. The DS model formalizes this idea. It also accounts for both mutational and selective effects, the latter being allowed to differ along the coding sequence and between different environmental conditions. Given the capability and versatility provided by the model, site- and condition-specific selection profiles are estimated for the whole sequence. Based on two empirical analyses as well as simulation experiments, I have shown that this model can find consistent patterns of convergent evolution at specific positions along the coding sequence.

A well-known example of where robust empirical knowledge about fitness fluctuation and convergent adaptation is available is the case of HLA-restricted HIV-1 evolution. HIV-1 is a virus that has been widely used in phylogenetic evolutionary studies, due to its error-prone replication and its subsequent fast mutation and adaptation to different immune responses, through a mechanism called escape mutation.

We explored two alternative approaches to model the codon substitution process of HIV-1; a phenomenological and a mechanistic approach. This part is presented in chapter 2 in the format of an article. We found that the phenomenological approach fits better to HIV-1 sequence data. This

result is not so surprising, given that the mechanistic approach (such as first formalized by Halpern and Bruno (Halpern and Bruno 1998)) relies on several fundamental assumptions (low mutation rates, no clonal interference, and no selection on synonymous variants) which are likely to be violated in the case of HIV-1 populations. The phenomenological model, although less directly interpretable in terms of population genetics mechanisms, appears to be more robust in this context.

We considered two HLA types in our analysis; HLA-B57 and HLA-B35 which are associated with HIV-1 control and disease progression, respectively. Most known mutations related to these two HLA types were recovered by our model, as well as new mutations. For HLA-B57+ condition, 16 amino acid-positions (12 positions) were recognized as differentially selected with posterior probability greater than 0.80.

This Differential Selection model can be used in other situations which bear Differential Selection as a function of known predictors, for viruses or other species. For example, HIV-1 positive patients who are exposed to different drug therapies are good candidates to study the evolution of the virus in response to each therapy.

We also applied the DS model to the evolution of Rubisco, an important enzyme participating in the first step of photosynthesis. The Rubisco gene has been known to be specially adapted to different modes of photosynthesis (C3 and C4). Chapter 3 presented our mechanistic attempt to reveal the amino acid-positions in rbcL sequences that are differentially selected in C3 and C4 *Amaranthaceae* family. Our Differential Selection model was successfully able to capture consistent patterns of molecular convergence in the sequence of the large subunit of Rubisco. We found 20 positions to be under Differential Selection in C4 plants compared to their C3

counterparts with posterior probability greater than 0.80. These amino acid substitutions have a conformational role in Rubisco enzyme in C4 plants, leading to its higher efficiency or be a compensatory mutation selected to maintain its optimized function. Apart from the DS model, we used a classical omega-based codon model, which is built on the approach of Muse and Gaut (Muse and Gaut 1994). As illustrated by our comparative analysis, the two models differ in the type of selective regime that they try to detect. The DS model is a mutation-selection model based on population genetics principles, which is able to detect convergent directional selection occurred along the C3 to C4 transitions in plants. On the other hand, the omega-based codon model focuses on finding positive selection based on the $d_N/d_S$ measurement of ongoing adaptations.

The model developed during this Ph.D. has been successfully employed the above applications and presented in the form of two articles. However, in each of these two domains, there are some extensions to this research that can be proposed as future directions

## 4.1 Improving statistical power of DS model

In the DS model, such as introduced in chapter 2, site-specific fitness differentials are modeled as random effects across sites. Parts of these random effects are assumed to be normally distributed, of mean 0 and variance 1. This model could certainly be improved, in several directions.

First, the variance of the Normal distribution should be a hyperparameter of the model and estimated from the dataset. Our current implementation allows for this to be done; however, this

is computationally challenging: still, some MCMC developments are needed here (combined with optimization/parallelization strategies).

Second, the DS model is not sparse: all positions have non-zero differential fitness effects for all amino acids with probability 1. The model, therefore, implements *soft* shrinkage toward zero. In most cases, differential effects are small and are such that their posterior distribution clearly overlaps 0. Our method for calling significant differential fitness focuses merely on those few cases for which the distribution is almost entirely on the positive or on the negative side. In practice, selection of significant positions is implemented based on the two-sided tail posterior distribution (i.e., posterior probability that fitness differential is greater than 0 is either more than $1- \alpha$ or less than $\alpha$).

An alternative to this soft shrinkage approach would be to employ a mixture model, in which not all amino acids at all positions are under differential-selection. For instance, the differential fitness effect for a given position and a given amino acid could be strictly positive with probability $\varepsilon_1$, or strictly negative with probability $\varepsilon_2$, or equal to 0 with probability $1- \varepsilon_1 - \varepsilon_2$. The $\varepsilon_i$ parameters, as well as the mean and the variance of the distribution of positive and negative effects, would then be estimated from the dataset. Under this model, a differential fitness effect would be called significant if the posterior probability of a non-zero effect is greater than $1-\alpha$, which would thus represent a more natural formalization of the problem. This type of spike-and-slab mixture model, which implements hard shrinkage, has already been implemented in other contexts in bioinformatics (Yang, Nielsen et al. 2000; Lewin, Bochkina et al. 2007). Ultimately, sparse Differential Selection profiles (with only a small number of positions and amino acids displaying significant non-null differential selective effects with high posterior probability) could be obtained

through the use this spike-and-slab mixture model. Also, a property of this model is to lead to a more principled control of the rate of false discovery (Lewin, Bochkina et al. 2007).

Another issue is the use of a simple parametric approach to model site-specific amino acid fitness profiles (which are *iid* from a Dirichlet distribution). In contrast, previous Bayesian implementations of mutation-selection models have relied on a non-parametric estimation of the distribution of fitness across sites (Rodrigue, Philippe et al. 2010). In this non-parametric approach, the distribution of site-specific selection is not specified a priori. Instead, it is inferred from the data. Ultimately, a general non-parametric DS model should be developed, in which, the unknown law of site- and condition-specific amino acid fitness profiles would be estimated.

Another limitation of the model is that it does not account for the uncertainty about the location of the shifts between alternative selective regimes. In practice, branches are allocated a priori, based on a maximum parsimony reconstruction. A better approach would be to explicitly model the evolution of the discrete character (e.g., the C3/C4 mode of photosynthesis) conditional on available data about extinct species and integrate this component of the model into the MCMC sampler.

## 4.2 Between- and within-host Differential Selection of HIV-1

The evolution of HIV is the result of the interplay between short-term within host and long-term between host evolutionary processes. The short-term evolution is mainly exerted by within-host selection pressure. Adaptation of the virus in this context is geared toward more efficient replication within the host body, as well as immune escape. On the other hand, long-term evolution

entails consecutive transmission between hosts. Therefore, between-host selection favors a high infectivity. The interplay between these two levels of selection is complex and has been the subject of recent interest (Lemey, Rambaut et al. 2006; Lythgoe and Fraser 2012; Alizon and Fraser 2013). In particular, it raises the question of whether there are conflicts, at the molecular level, between these two modes of adaptation (i.e., pleiotropy).

Our model can enable us to distinguish between these two sources of selection. Here, in contrast with other studies, within- and between-host level of selection would directly be captured based on one single dataset including multiple sequences from multiple hosts. In practice, we would need to take the whole genome of HIV-1 from patients with more than five different copies. Addressing these questions would indeed be challenging and might require many additional developments. However, the current interest in those problems makes it certainly worth a try.

## 4.3 Calibration and quantification of Rubisco evolution

Our analysis of Rubisco was on eudicots. We briefly compared our findings with those presented by Studer *et al.* (Studer, Christin et al. 2014) on monocots rbcL sequences, which were obtained based on a model also accounting for explicit condition-dependent amino acid patterns, although not at the codon level. However, the comparison was made complicated by the fact that the two analyses were not conducted on the same dataset. A more extended analysis, jointly, of monocot and eudicot Rubisco sequences, using our model, would be useful to better understand the relative merits and properties of our approach, compared to that of (Tamuri, dos Reis et al. 2009).

In contrast to HIV-1, Rubisco evolves at a much slower rate. In part for that reason, the mechanistic model, based on the Halpern and Bruno mutation-selection approach (Halpern and Bruno 1998), is more appropriate in this context than in the case of HIV-1. In turn, the use of a mechanistic model offers new opportunities. In particular, it is possible to quantify the changes in actual Darwinian fitness contributed by the evolution of Rubisco, during C3 to C4 transitions. This can be done by investigating the changes in relative fitness of the reconstructed sequences through time along the phylogenetic tree. A simple prediction is that the fitness of a sequence abruptly decreases at the point of the C3 to C4 transition, and then progressively recovers in subsequent branches. Addressing this question would be of great interest.

Finally, our comparison between classical omega-based and DS modeling approaches did not include *branch-site* codon models, whether in their fixed effects (Kosakovsky Pond and Frost 2005), random effects (Kosakovsky Pond, Murrell et al. 2011) or mixed effects versions (Murrell, Wertheim et al. 2012) (see Introduction). *Branch-site* models aim at detecting episodes of diversifying selection occurring on specific branches for specific sites, but it is challenging to identify specific branches under episodic selection. Also, *branch-site* models tend to be sensitive to violations of the model's assumptions. In particular, multinucleotide mutations produce false support for positive selection as they are most apt to be nonsynonymous (Venkat, Hahn et al. 2017). Eventually, a more thorough comparative analysis of the merit of *branch-site* models, compared to DS models, would be warranted. In particular, it would be interesting to investigate whether these models are able to detect those branches on which the transitions between C3 and C4 have occurred.

# BIBLIOGRAPHY

(12/12/2013). "www.hiv.lanl.gov."

Adachi, J. and M. Hasegawa (1996). "Model of amino acid substitution in proteins encoded by mitochondrial DNA." Journal of molecular evolution **42**(4): 459-468.

Alizon, S. and C. Fraser (2013). "Within-host and between-host evolutionary rates across the HIV-1 genome." Retrovirology **10**(1): 1-10.

Altfeld, M., M. M. Addo, E. S. Rosenberg, F. M. Hecht, P. K. Lee, M. Vogel*, et al.* (2003). "Influence of HLA-B57 on clinical presentation and viral control during acute HIV-1 infection." AIDS **17**(18): 2581-2591.

Altfeld, M. and T. M. Allen (2006). "Hitting HIV where it hurts: an alternative approach to HIV vaccine design." Trends in Immunology **27**(11): 504-510.

Andersson, I. and A. Backlund (2008). "Structure and function of Rubisco." Plant Physiol Biochem **46**(3): 275-291.

Andrews, T. J., and Lorimer, G.H. (1987). Rubisco: Structure, mechanisms and prospect for improvement. The Biochemistry of Plants, New York: Academic Press. **10:** 131-218.

Anisimova, M., J. P. Bielawski and Z. Yang (2001). "Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution." Molecular biology and evolution **18**(8): 1585-1592.

Baeten, J. M., B. Chohan, L. Lavreys, V. Chohan, R. S. McClelland, L. Certain*, et al.* (2007). "HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads." The Journal of infectious diseases **195**(8): 1177-1180.

Blanquart, S. and N. Lartillot (2006). "A Bayesian Compound Stochastic Process for Modeling Nonstationary and Nonhomogeneous Sequence Evolution." Molecular biology and evolution **23**(11): 2058-2071.

Blanquart, S. and N. Lartillot (2008). "A site- and time-heterogeneous model of amino acid replacement." Molecular biology and evolution **25**(5): 842-858.

Boc, A., A. B. Diallo and V. Makarenkov (2012). "T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks." Nucleic acids research **40**(W1): W573-W579.

Boutwell, C. L., C. F. Rowley and M. Essex (2009). "Reduced Viral Replication Capacity of Human Immunodeficiency Virus Type 1 Subtype C Caused by Cytotoxic-T-Lymphocyte Escape Mutations in HLA-B57 Epitopes of Capsid Protein." Journal of virology **83**(6): 2460-2468.

Braaten, D., E. K. Franke and J. Luban (1996). "Cyclophilin A is required for an early step in the life cycle of human immunodeficiency virus type 1 before the initiation of reverse transcription." Journal of virology **70**(6): 3551-3560.

Brockman, M. A., A. Schneidewind, M. Lahaie, A. Schmidt, T. Miura, I. Desouza*, et al.* (2007). "Escape and compensation from early HLA-B57-mediated cytotoxic T-lymphocyte pressure on human immunodeficiency virus type 1 Gag alter capsid interactions with cyclophilin A." Journal of virology **81**(22): 12608-12618.

Brooks, J. I., H. Niznick, M. Ofner, H. Merks and J. B. Angel (2013). "Local phylogenetic analysis identifies distinct trends in transmitted HIV drug resistance: implications for public health interventions." BMC infectious diseases **13**(1): 1-8.

Brumme, Z. L., C. J. Brumme, D. Heckerman, B. T. Korber, M. Daniels, J. Carlson*, et al.* (2007). "Evidence of differential HLA class I-mediated viral evolution in functional and accessory/regulatory genes of HIV-1." PLoS Pathog **3**(7): e94-e94.

Brumme, Z. L., I. Tao, S. Szeto, C. J. Brumme, J. M. Carlson, D. Chan*, et al.* (2008). "Human leukocyte antigen-specific polymorphisms in HIV-1 Gag and their association with viral load in chronic untreated infection." AIDS **22**(11): 1277-1286.

C Dean, a. E Pichersky and P. Dunsmuir (1989). "Structure, Evolution, and Regulation of RbcS Genes in Higher Plants." Annual Review of Plant Physiology and Plant Molecular Biology **40**(1): 415-439.

Carlson, J. M. and Z. L. Brumme (2008). "HIV evolution in response to HLA-restricted CTL selection pressures: a population-based perspective." Microbes and Infection **10**(5): 455-461.

Carlson, J. M., Z. L. Brumme, C. M. Rousseau, C. J. Brumme, P. Matthews, C. Kadie*, et al.* (2008). "Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag." PLoS computational biology **4**(11): e1000225.

Carrington, M., G. W. Nelson, M. P. Martin, T. Kissner, D. Vlahov, J. J. Goedert*, et al.* (1999). "HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage." Science **283**(5408): 1748-1752.

Castro-Nallar, E., M. Pérez-Losada, G. F. Burton and K. A. Crandall (2012). "The evolution of HIV: Inferences using phylogenetics." Molecular phylogenetics and evolution **62**(2): 777-792.

Cavalli-Sforza, L. L. and A. W. F. Edwards (1967). "Phylogenetic analysis. Models and estimation procedures." American journal of human genetics **19**(3 Pt 1): 233-257.

Cerling, T. E. (1999). Paleorecords of C4 Plants and Ecosystems. C4 Plant Biology. R. S. a. R. Monson. Saan Diego, CA, Academic Press**:** 445-469.

Chopera, D. R., Z. Woodman, K. Mlisana, M. Mlotshwa, D. P. Martin, C. Seoighe*, et al.* (2008). "Transmission of HIV-1 CTL Escape Variants Provides HLA-Mismatched Recipients with a Survival Advantage." PLoS Pathogens **4**(3): e1000033.

Christin, P.-A., R. P. Freckleton and C. P. Osborne (2010). "Can phylogenetics identify C4 origins and reversals?" Trends in Ecology & Evolution **25**(7): 403-409.

Christin, P.-A., N. Salamin, A. M. Muasya, E. H. Roalson, F. Russier and G. Besnard (2008). "Evolutionary Switch and Genetic Convergence on rbcL following the Evolution of C4 Photosynthesis." Molecular biology and evolution **25**(11): 2361-2368.

Christin, P. A., G. Besnard, E. Samaritani, M. R. Duvall, T. R. Hodkinson, V. Savolainen*, et al.* (2008). "Oligocene CO2 decline promoted C4 photosynthesis in grasses." Curr Biol **18**(1): 37-43.

Cowling, S. A. (2001). "Plant carbon balance, evolutionary innovation and extinction in land plants." Global Change Biology **7**(3): 231-239.

Crooks, G. E., G. Hon, J. M. Chandonia and S. E. Brenner (2004). "WebLogo: a sequence logo generator." Genome research **14**(6): 1188-1190.

Dayhoff, M. O., Schwartz, R. M., Orcutt, B. C. (1978). A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure. Washington, DC., Natl. Biomed. Res. Found.**:** 345-352.

Dobzhansky, T. (1973). "Nothing in Biology Makes Sense except in the Light of Evolution." The American Biology Teacher **35**(3): 125-129.

Edwards, C. T. T., E. C. Holmes, O. G. Pybus, D. J. Wilson, R. P. Viscidi, E. J. Abrams*, et al.* (2006). "Evolution of the Human Immunodeficiency Virus Envelope Gene Is Dominated by Purifying Selection." Genetics **174**(3): 1441-1453.

Edwards, E. J., C. P. Osborne, C. A. E. Strömberg, S. A. Smith and C. G. Consortium (2010). "The Origins of C4 Grasslands: Integrating Evolutionary and Ecosystem Science." Science **328**(5978): 587-591.

Ehleringer, J. R., R. F. Sage, L. B. Flanagan and R. W. Pearcy (1991). "Climate change and the evolution of C(4) photosynthesis." Trends Ecol Evol **6**(3): 95-99.

Elena, S. F., C. O. Wilke, C. Ofria and R. E. Lenski (2007). "Effects of population size and mutation rate on the evolution of mutational robustness." Evolution **61**(3): 666-674.

Ellis, R. J. (1979). "The most abundant protein in the world." Trends in Biochemical Sciences **4**(11): 241-244.

Feller, U., I. Anders and T. Mae (2008). "Rubiscolytics: fate of Rubisco after its enzymatic function in a cell is terminated." J Exp Bot **59**(7): 1615-1624.

Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." Journal of molecular evolution **17**(6): 368-376.

Flores-Villanueva, P. O., H. Hendel, S. Caillat-Zucman, J. Rappaport, A. Burgos-Tiburcio, S. Bertin-Maghit*, et al.* (2003). "Associations of MHC Ancestral Haplotypes with Resistance/Susceptibility to AIDS Disease Development." The Journal of Immunology **170**(4): 1925-1929.

Foster, P. G. (2004). "Modeling Compositional Heterogeneity." Systematic biology **53**(3): 485-495.

Galtier, N. and M. Gouy (1998). "Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis." Molecular biology and evolution **15**(7): 871-879.

Gascuel, O. and S. Guindon (2007). Modelling the Variability of Evolutionary Processes. Reconstructing Evolution: New Mathematical and Computational Advances. G. Olivier and M. Steel. **II Models of sequence evolution:** 65-99.

Gelman, A., X.-L. Meng and H. Stern (1996). "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies." Statistica Sinica **6**(4): 733-760.

Gelman, A. and D. B. Rubin (1992). "Inference from Iterative Simulation Using Multiple Sequences." (4): 457-472.

Goepfert, P. A., W. Lumm, P. Farmer, P. Matthews, A. Prendergast, J. M. Carlson*, et al.* (2008). "Transmission of HIV-1 Gag immune escape mutations is associated with reduced viral load in linked recipients." The Journal of experimental medicine **205**(5): 1009-1017.

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

Goulder, P. J., M. Bunce, P. Krausa, K. McIntyre, S. Crowley, B. Morgan*, et al.* (1996). "Novel, cross-restricted, conserved, and immunodominant cytotoxic T lymphocyte epitopes in slow progressors in HIV type 1 infection." AIDS research and human retroviruses **12**(18): 1691-1698.

Goulder, P. J. and D. I. Watkins (2004). "HIV and SIV CTL escape: implications for vaccine design." Nature reviews. Immunology **4**(8): 630-640.

Gowik, U. and P. Westhoff (2011). "The Path from C3 to C4 Photosynthesis." Plant Physiology **155**(1): 56-63.

Groussin, M., B. Boussau and M. Gouy (2013). "A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences." Systematic biology **62**(4): 523-538.

Halpern, A. L. and W. J. Bruno (1998). "Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies." Molecular biology and evolution **15**(7): 910-917.

Hasegawa, M., H. Kishino and T. Yano (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA." Journal of molecular evolution **22**(2): 160-174.

Hastings, W. K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." Biometrika **57**(1): 97-109.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proceedings of the National Academy of Sciences of the United States of America **89**(22): 10915-10919.

Hillis, D. M. (1999). Phylogenetics and the study of HIV. The Evolution of HIV. K. A. Crandall. Baltimore, MD, Johns Hopkins University Press.

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore*, et al.* (2016). "RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language." Systematic biology **65**(4): 726-736.

Hudson, G. S., J. D. Mahon, P. A. Anderson, M. J. Gibbs, M. R. Badger, T. J. Andrews*, et al.* (1990). "Comparisons of rbcL genes for the large subunit of ribulose-bisphosphate carboxylase from closely related C3 and C4 plant species." J Biol Chem **265**(2): 808-814.

Huelsenbeck, J. P., B. Larget, R. E. Miller and F. Ronquist (2002). "Potential applications and pitfalls of Bayesian inference of phylogeny." Systematic biology **51**(5): 673-688.

Huelsenbeck, J. P., B. Rannala and J. P. Masly (2000). "Accommodating phylogenetic uncertainty in evolutionary studies." Science **288**(5475): 2349-2350.

Huelsenbeck, J. P. and F. Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees." Bioinformatics **17**(8): 754-755.

Huson, D. H. and C. Scornavacca (2012). "Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks." Systematic Biology **61**(6): 1061-1067.

Iida, S., A. Miyagi, S. Aoki, M. Ito, Y. Kadono and K. Kosuge (2009). "Molecular Adaptation of rbcL in the Heterophyllous Aquatic Plant Potamogeton." PloS one **4**(2): e4633.

Itescu, S., U. Mathur-Wagh, M. L. Skovron, L. J. Brancato, M. Marmor, A. Zeleniuch-Jacquotte*, et al.* (1992). "HLA-B35 is associated with accelerated progression to AIDS." Journal of acquired immune deficiency syndromes **5**(1): 37-45.

Jenkins, G. M., A. Rambaut, O. G. Pybus and E. C. Holmes (2002). "Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis." Journal of molecular evolution **54**(2): 156-165.

Jiang, X., B. Mu, Z. Huang, M. Zhang, X. Wang and S. Tao (2010). "Impacts of mutation effects and population size on mutation rate in asexual populations: a simulation study." BMC evolutionary biology **10**: 298-298.

Jones, D. T., W. R. Taylor and J. M. Thornton (1994). "A mutation data matrix for transmembrane proteins." FEBS letters **339**(3): 269-275.

Jordan, D. B. and W. L. Ogren (1981). "Species variation in the specificity of ribulose biphosphate carboxylase/oxygenase." Nature **291**(5815): 513-515.

Jukes, T. H. and C. R. Cantor (1969). Evolution of Protein Molecules, Academy Press.

Jukes, T. H. and J. L. King (1979). "Evolutionary nucleotide replacements in DNA." Nature **281**(5732): 605-606.

Kadereit, G., T. Borsch, K. Weising and H. Freitag (2003). "Phylogeny of Amaranthaceae and Chenopodiaceae and the evolution of C-4 photosynthesis." International Journal of Plant Sciences **164**(6): 959-986.

Kapralov, M. and D. Filatov (2007). "Widespread positive selection in the photosynthetic Rubisco enzyme." BMC evolutionary biology **7**(1): 1-10.

Kapralov, M. V., J. A. C. Smith and D. A. Filatov (2012). "Rubisco Evolution in C(4) Eudicots: An Analysis of Amaranthaceae Sensu Lato." PloS one **7**(12): e52974.

Kapralov, M. V., A. A. Votintseva and D. A. Filatov (2013). "Molecular Adaptation during a Rapid Adaptive Radiation." Molecular biology and evolution **30**(5): 1051-1059.

Kaslow, R. A., M. Carrington, R. Apple, L. Park, A. Munoz, A. J. Saah*, et al.* (1996). "Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection." Nature medicine **2**(4): 405-411.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." Journal of molecular evolution **16**(2): 111-120.

Kleinman, C. L., N. Rodrigue, N. Lartillot and H. Philippe (2010). "Statistical Potentials for Improved Structurally Constrained Evolutionary Models." Molecular biology and evolution **27**(7): 1546-1560.

Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes*, et al.* (2000). "Timing the ancestor of the HIV-1 pandemic strains." Science **288**(5472): 1789-1796.

Kosakovsky Pond, S. L. and S. D. Frost (2005). "Not so different after all: a comparison of methods for detecting amino acid sites under selection." Molecular biology and evolution **22**(5): 1208-1222.

Kosakovsky Pond, S. L., B. Murrell, M. Fourment, S. D. Frost, W. Delport and K. Scheffler (2011). "A random effects branch-site model for detecting episodic diversifying selection." Molecular biology and evolution **28**(11): 3033-3043.

Kosiol, C., T. Vinař, R. R. da Fonseca, M. J. Hubisz, C. D. Bustamante, R. Nielsen*, et al.* (2008). "Patterns of Positive Selection in Six Mammalian Genomes." PLoS Genet **4**(8): e1000144.

Kuiken C., L. T., et al (2008). HIV Sequence Compendium

L DeLano, W. (2002). The PyMOL Molecular Graphics System (2002) DeLano Scientific, Palo Alto, CA, USA. http://www.pymol.org.

Lam, T. T.-Y., C.-C. Hon and J. W. Tang (2010). "Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections." Critical Reviews in Clinical Laboratory Sciences **47**(1): 5-49.

Larget, B. and D. Simon (1999). "Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees." Molecular biology and evolution **16**(6): 750.

Lartillot, N. (2006). "Conjugate Gibbs Sampling for Bayesian Phylogenetic Models." Journal of Computational Biology **13**(10): 1701-1722.

Lartillot, N. (2015). "Probabilistic models of eukaryotic evolution: time for integration." Philosophical Transactions of the Royal Society B: Biological Sciences **370**(1678).

Lartillot, N., T. Lepage and S. Blanquart (2009). "PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating." Bioinformatics **25**.

Lartillot, N. and H. Philippe (2004). "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process." Molecular biology and evolution **21**(6): 1095-1109.

Lartillot, N. and R. Poujol (2011). "A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters." Molecular biology and evolution **28**(1): 729-744.

Lemey, P., A. Rambaut and O. G. Pybus (2006). "HIV evolutionary dynamics within and among hosts." AIDS Rev **8**(3): 125-140.

Leslie, A., D. Kavanagh, I. Honeyborne, K. Pfafferott, C. Edwards, T. Pillay*, et al.* (2005). "Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA." The Journal of experimental medicine **201**(6): 891-902.

Leslie, A. J., K. J. Pfafferott, P. Chetty, R. Draenert, M. M. Addo, M. Feeney*, et al.* (2004). "HIV evolution: CTL escape mutation and reversion after transmission." Nature medicine **10**(3): 282-289.

Levy, J. A. (1993). "HIV pathogenesis and long-term survival." AIDS **7**(11): 1401-1410.

Lewin, A., N. Bochkina and S. Richardson (2007). "Fully Bayesian mixture model for differential gene expression: simulations and model checks." Stat Appl Genet Mol Biol **6**: Article36.

Lio, P., N. Goldman, J. L. Thorne and D. T. Jones (1998). "PASSML: combining evolutionary inference and protein secondary structure prediction." Bioinformatics **14**(8): 726-733.

Liu, Z., N. Sun, S. Yang, Y. Zhao, X. Wang, X. Hao*, et al.* (2013). "Evolutionary transition from C3 to C4 photosynthesis and the route to C4 rice." Biologia **68**(4): 577-586.

Lowe, D. R. (1994). Early environments: constraints and opportunities for early evolution Early Life on Earth. S. Bengston. New York, Columbia University Press**:** 24-35.

Lythgoe, K. A. and C. Fraser (2012). "New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels." Proceedings of the Royal Society B: Biological Sciences **279**(1741): 3367-3375.

Makino, A. (2003). "Rubisco and nitrogen relationships in rice: Leaf photosynthesis and plant growth." Soil Science and Plant Nutrition **49**(3): 319-327.

Martinez-Picado, J., J. G. Prado, E. E. Fry, K. Pfafferott, A. Leslie, S. Chetty*, et al.* (2006). "Fitness Cost of Escape Mutations in p24 Gag in Association with Control of Human Immunodeficiency Virus Type 1." Journal of virology **80**(7): 3617-3623.

Mateiu, L. and B. Rannala (2006). "Inferring Complex DNA Substitution Processes on Phylogenies Using Uniformization and Data Augmentation." Systematic biology **55**(2): 259-269.

Matthews, P. C., M. Koyanagi, H. N. Kloverpris, M. Harndahl, A. Stryhn, T. Akahoshi*, et al.* (2012). "Differential clade-specific HLA-B*3501 association with HIV-1 disease outcome is linked to immunogenicity of a single Gag epitope." Journal of virology **86**(23): 12643-12654.

Matthews, P. C., A. Prendergast, A. Leslie, H. Crawford, R. Payne, C. Rousseau*, et al.* (2008). "Central role of reverting mutations in HLA associations with human immunodeficiency virus set point." Journal of virology **82**(17): 8548-8559.

Mayrose, I., A. Stern, E. O. Burdelova, Y. Sabo, N. Laham-Karam, R. Zamostiano*, et al.* (2013). "Synonymous site conservation in the HIV-1 genome." BMC evolutionary biology **13**(1): 1-11.

McCloskey, R. M., R. H. Liang, P. R. Harrigan, Z. L. Brumme and A. F. Poon (2014). "An evaluation of phylogenetic methods for reconstructing transmitted HIV variants using longitudinal clonal HIV sequence data." Journal of virology **88**(11): 6181-6194.

McMichael, A. J. and S. L. Rowland-Jones (2001). "Cellular immune responses to HIV." Nature **410**(6831): 980-987.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller (1953). "Equation of State Calculations by Fast Computing Machines." The Journal of Chemical Physics **21**(6): 1087-1092.

Migueles, S. A., M. S. Sabbaghian, W. L. Shupert, M. P. Bettinotti, F. M. Marincola, L. Martino*, et al.* (2000). "HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors." Proceedings of the National Academy of Sciences **97**(6): 2709-2714.

Miura, T., C. J. Brumme, M. A. Brockman, Z. L. Brumme, F. Pereyra, B. L. Block*, et al.* (2009). "HLA-Associated Viral Mutations Are Common in Human Immunodeficiency Virus Type 1 Elite Controllers." Journal of virology **83**(7): 3407-3412.

Monson, R. K. and S. Rawsthorn (2000). Carbon dioxide assimilation in C3-C4 intermediate plants. Photosynthesis: Physiology and Metabolism. Advances in Photosynthesis. T. S. R Leegood, S von Caemmerer. Dordrecht, NL, Kluwer Academic Press**:** 533-550.

Moore, C. B., M. John, I. R. James, F. T. Christiansen, C. S. Witt and S. A. Mallal (2002). "Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level." Science **296**(5572): 1439-1443.

Murrell, B., T. de Oliveira, C. Seebregts, S. L. Kosakovsky Pond, K. Scheffler, T. on behalf of the Southern African*, et al.* (2012). "Modeling HIV-1 Drug Resistance as Episodic Directional Selection." PLOS Computational Biology **8**(5): e1002507.

Murrell, B., J. O. Wertheim, S. Moola, T. Weighill, K. Scheffler and S. L. Kosakovsky Pond (2012). "Detecting Individual Sites Subject to Episodic Diversifying Selection." PLOS Genetics **8**(7): e1002764.

Muse, S. V. and B. S. Gaut (1994). "A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome." Molecular biology and evolution **11**(5): 715-724.

Mustonen, V. and M. Lassig (2009). "From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation." Trends Genet **25**(3): 111-119.

Mustonen, V. and M. Lässig (2008). "Molecular Evolution under Fitness Fluctuations." Physical review letters **100**(10): 108101.

Nielsen, R. and Z. Yang (1998). "Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene." Genetics **148**(3): 929-936.

Nomura, M., K. Katayama, A. Nishimura, Y. Ishida, S. Ohta, T. Komari*, et al.* (2000). "The promoter of rbcS in a C3 plant (rice) directs organ-specific, light-dependent expression in a C4 plant (maize), but does not confer bundle sheath cell-specific expression." Plant Mol Biol **44**(1): 99-106.

Novitsky, V., U. R. Smith, P. Gilbert, M. F. McLane, P. Chigwedere, C. Williamson*, et al.* (2002). "Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design?" Journal of virology **76**(11): 5435-5451.

Pagel, M. and A. Meade (2004). "A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data." Systematic biology **53**(4): 571-581.

Parto, S. and N. Lartillot (2017). "Detecting consistent patterns of directional adaptation using differential selection codon models." BMC evolutionary biology **17**: 147.

Parto, S. and N. Lartillot (2018). "Molecular adaptation in Rubisco: Discriminating between convergent evolution and positive selection using mechanistic and classical codon models." PloS one **13**(2): e0192697.

Peretz, Y., G. Alter, M.-P. Boisvert, G. Hatzakis, C. M. Tsoukas and N. F. Bernard (2005). "Human Immunodeficiency Virus (HIV)-Specific Gamma Interferon Secretion Directed against All Expressed HIV Genes: Relationship to Rate of CD4 Decline." J. Virol. **79**(8): 4908-4917.

Pereyra, F., M. M. Addo, D. E. Kaufmann, Y. Liu, T. Miura, A. Rathod*, et al.* (2008). "Genetic and immunologic heterogeneity among persons who control HIV infection in the absence of therapy." Journal of Infectious Diseases **197**(4): 563-571.

Pollock, D. D., W. R. Taylor and N. Goldman (1999). "Coevolving protein residues: maximum likelihood identification and relationship to structure." Journal of molecular biology **287**(1): 187-198.

Pond, S. L., S. D. Frost, Z. Grossman, M. B. Gravenor, D. D. Richman and A. J. Brown (2006). "Adaptation to different human populations by HIV-1 revealed by codon-based analyses." PLoS computational biology **2**(6): e62.

Pybus, O. G. and A. Rambaut (2009). "Evolutionary analysis of the dynamics of viral infectious disease." Nature reviews. Genetics **10**(8): 540-550.

Rambaut, A., D. Posada, K. A. Crandall and E. C. Holmes (2004). "The causes and consequences of HIV evolution." Nature reviews. Genetics **5**(1): 52-61.

Rannala, B. and Z. Yang (1996). "Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference." Journal of molecular evolution **43**(3): 304-311.

Rios, M., E. Delgado, L. Perez-Alvarez, J. Fernandez, P. Galvez, E. V. de Parga*, et al.* (2007). "Antiretroviral drug resistance and phylogenetic diversity of HIV-1 in Chile." Journal of medical virology **79**(6): 647-656.

Robertson, D. L., B. H. Hahn and P. M. Sharp (1995). "Recombination in AIDS viruses." Journal of molecular evolution **40**(3): 249-259.

Robinson, D. F. and L. R. Foulds (1981). "Comparison of phylogenetic trees." Mathematical Biosciences **53**(1): 131-147.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman and J. L. Thorne (2003). "Protein evolution with dependence among codons due to tertiary structure." Molecular biology and evolution **20**(10): 1692-1704.

Rodrigue, N. (2013). "On the statistical interpretation of site-specific variables in phylogeny-based substitution models." Genetics **193**(2): 557-564.

Rodrigue, N., H. Philippe and N. Lartillot (2010). "Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles." Proceedings of the National Academy of Sciences of the United States of America **107**(10): 4629-4634.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna*, et al.* (2012). "MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space." Systematic biology **61**(3): 539-542.

Rousseau, C. M., M. G. Daniels, J. M. Carlson, C. Kadie, H. Crawford, A. Prendergast*, et al.* (2008). "HLA Class I-Driven Evolution of Human Immunodeficiency Virus Type 1 Subtype C Proteome: Immune Escape and Viral Load." Journal of virology **82**(13): 6434-6446.

Roy, H. and T. Andrews (2000). Rubisco: assembly and mechanism. Photosynthesis: Physiology and Metabolism. . Springer, Dordrecht, Kluwer Academic Publishers. **9:** 53-83.

Rubin, D. B. (1984). "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician." (4): 1151-1172.

Russell K. Monson (2003). "Gene Duplication, Neofunctionalization, and the Evolution of C4 Photosynthesis." International Journal of Plant Sciences **164**(S3): S43-S54.

Sage, R. F. (2002). "Variation in the k(cat) of Rubisco in C(3) and C(4) plants and some implications for photosynthetic performance at high and low temperature." J Exp Bot **53**(369): 609-620.

Sage, R. F. (2004). "The evolution of C4 photosynthesis." New Phytologist **161**(2): 341-370.

Sage, R. F., P.-A. Christin and E. J. Edwards (2011). "The C4 plant lineages of planet Earth." Journal of Experimental Botany **62**(9): 3155-3169.

Sage, R. F. and J. R. Coleman (2001). "Effects of low atmospheric CO2 on plants: more than a thing of the past." Trends in Plant Science **6**(1): 18-24.

Sage, R. F., R. W. Pearcy and J. R. Seemann (1987). "The Nitrogen Use Efficiency of C(3) and C(4) Plants : III. Leaf Nitrogen Effects on the Activity of Carboxylating Enzymes in Chenopodium album (L.) and Amaranthus retroflexus (L.)." Plant Physiology **85**(2): 355-359.

Saitou, N. and M. Nei (1987). "The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees." Molecular biology and evolution **4**(4): 406-425.

Salemi, M., B. R. Burkhardt, R. R. Gray, G. Ghaffari, J. W. Sleasman and M. M. Goodenow (2007). "Phylodynamics of HIV-1 in Lymphoid and Non-Lymphoid Tissues Reveals a Central Role for the Thymus in Emergence of CXCR4-Using Quasispecies." PloS one **2**(9): e950.

Sawyer, S. L., M. Emerman and H. S. Malik (2004). "Ancient Adaptive Evolution of the Primate Antiviral DNA-Editing Enzyme APOBEC3G." PLoS Biol **2**(9): e275.

Schneider, T. D. and R. M. Stephens (1990). "Sequence logos: a new way to display consensus sequences." Nucleic acids research **18**(20): 6097-6100.

Schweighardt, B., T. Wrin, D. A. Meiklejohn, G. Spotts, C. J. Petropoulos, D. F. Nixon*, et al.* (2010). "Immune escape mutations detected within HIV-1 epitopes associated with viral control during treatment interruption." Journal of acquired immune deficiency syndromes **53**(1): 36-46.

Sen, L., M. A. Fares, B. Liang, L. Gao, B. Wang, T. Wang*, et al.* (2011). "Molecular evolution of rbcL in three gymnosperm families: identifying adaptive and coevolutionary patterns." Biology Direct **6**: 29-29.

Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan*, et al.* (1999). "Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection." Journal of virology **73**(12): 10489-10502.

Sharp, P. M. and B. H. Hahn (2010). "The evolution of HIV-1 and the origin of AIDS." Philosophical Transactions of the Royal Society B: Biological Sciences **365**(1552): 2487-2494.

Simpson, M. G. (2010). Plant Systematics. San Diego, CA, USA, Academic Press.

Spielman, S. J. and C. O. Wilke (2015). "The Relationship between dN/dS and Scaled Selection Coefficients." Molecular biology and evolution.

Spreitzer, R. J. (1993). "Genetic Dissection of Rubisco Structure and Function." Annual Review of Plant Physiology and Plant Molecular Biology **44**(1): 411-434.

Ssemwanga, D., R. N. Nsubuga, B. N. Mayanja, F. Lyagoba, B. Magambo, D. Yirrell*, et al.* (2013). "Effect of HIV-1 Subtypes on Disease Progression in Rural Uganda: A Prospective Clinical Cohort Study." PloS one **8**(8): e71768.

Still, C. J., J. A. Berry, G. J. Collatz and R. S. DeFries (2003). "Global distribution of C3 and C4 vegetation: Carbon cycle implications." Global Biogeochemical Cycles **17**(1): 6-1-6-14.

Stowe, L. G. and J. A. Teeri (1978). "The Geographic Distribution of C4 Species of the Dicotyledonae in Relation to Climate." The American Naturalist **112**(985): 609-623.

Studer, R. A., P.-A. Christin, M. A. Williams and C. A. Orengo (2014). "Stability-activity tradeoffs constrain the adaptive evolution of RubisCO." Proceedings of the National Academy of Sciences **111**(6): 2223-2228.

Tabita, F. R., T. E. Hanson, H. Li, S. Satagopan, J. Singh and S. Chan (2007). "Function, Structure, and Evolution of the RubisCO-Like Proteins and Their RubisCO Homologs." Microbiology and Molecular Biology Reviews **71**(4): 576-599.

Tamura, K. and M. Nei (1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees." Molecular biology and evolution **10**(3): 512-526.

Tamuri, A. U., M. dos Reis and R. A. Goldstein (2012). "Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation-Selection Models." Genetics **190**(3): 1101-1115.

Tamuri, A. U., M. dos Reis, A. J. Hay and R. A. Goldstein (2009). "Identifying Changes in Selective Constraints: Host Shifts in Influenza." PLoS computational biology **5**(11): e1000564.

Tamuri, A. U., N. Goldman and M. dos Reis (2014). "A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data." Genetics **197**(1): 257-271.

Taniguchi, Y., H. Ohkawa, C. Masumoto, T. Fukuda, T. Tamai, K. Lee*, et al.* (2008). "Overproduction of C4 photosynthetic enzymes in transgenic rice plants: an approach to introduce the C4-like photosynthetic pathway into rice." J Exp Bot **59**(7): 1799-1809.

Tavaré, S. (1986). Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. American Mathematical Society: Lectures on Mathematics in the Life Sciences, Amer Mathematical Society. **17:** 57-86.

Templeton, A. R., R. A. Reichert, A. E. Weisstein, X. F. Yu and R. B. Markham (2004). "Selection in context: patterns of natural selection in the glycoprotein 120 region of human immunodeficiency virus 1 within infected individuals." Genetics **167**(4): 1547-1561.

Thorne, J. L., S. C. Choi, J. Yu, P. G. Higgs and H. Kishino (2007). "Population Genetics Without Intraspecific Data." Molecular biology and evolution **24**(8): 1667-1677.

Thorne, J. L., N. Goldman and D. T. Jones (1996). "Combining protein evolution and secondary structure." Molecular biology and evolution **13**(5): 666-673.

van Lun, M., D. van der Spoel and I. Andersson (2011). "Subunit Interface Dynamics in Hexadecameric Rubisco." Journal of Molecular Biology **411**(5): 1083-1098.

Venkat, A., M. W. Hahn and J. W. Thornton (2017). "Multinucleotide mutations cause false inferences of positive selection." bioRxiv.

von Caemmerer, S. and J. R. Evans (2010). "Enhancing C3 Photosynthesis." Plant Physiology **154**(2): 589-592.

von Caemmerer, S. and W. P. Quick (2000). Rubisco: Physiology in Vivo. Photosynthesis. R. Leegood, T. Sharkey and S. von Caemmerer, Springer Netherlands. **9:** 85-113.

Voronin, Y., S. Holte, J. Overbaugh and M. Emerman (2009). "Genetic Drift of HIV Populations in Culture." PLOS Genetics **5**(3): e1000431.

Weber, J., J. Weberova, M. Carobene, M. Mirza, J. Martinez-Picado, P. Kazanjian*, et al.* (2006). "Use of a novel assay based on intact recombinant viruses expressing green (EGFP) or red (DsRed2) fluorescent proteins to examine the contribution of pol and env genes to overall HIV-1 replicative fitness." J Virol Methods **136**(1-2): 102-117.

Wei, M., H. Xing, Y. Feng, J. H. Hsi, P. Liu and Y. Shao (2015). "Estimating HIV-1 Transmission Routes for Patients With Unknown Risk Histories by Viral Sequence Phylogenetic Analyses." Journal of acquired immune deficiency syndromes **70**(2): 195-203.

Wildman, S. G. and J. Bonner (1947). "The proteins of green leaves; isolation, enzymatic properties and auxin content of spinach cytoplasmic proteins." Arch Biochem **14**(3): 381-413.

Yang, Z. (1993). "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites." Molecular biology and evolution **10**(6): 1396-1401.

Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." Comput Appl Biosci **13**(5): 555-556.

Yang, Z. (1998). "Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution." Molecular biology and evolution **15**(5): 568-573.

Yang, Z. (2007). "PAML 4: Phylogenetic Analysis by Maximum Likelihood." Molecular biology and evolution **24**(8): 1586-1591.

Yang, Z. and J. P. Bielawski (2000). "Statistical methods for detecting molecular adaptation." Trends Ecol Evol **15**(12): 496-503.

Yang, Z. and R. Nielsen (1998). "Synonymous and nonsynonymous rate variation in nuclear genes of mammals." Journal of molecular evolution **46**(4): 409-418.

Yang, Z. and R. Nielsen (2002). "Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages." Molecular biology and evolution **19**(6): 908-917.

Yang, Z. and R. Nielsen (2008). "Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage." Molecular biology and evolution **25**(3): 568-579.

Yang, Z., R. Nielsen, N. Goldman and A. M. Pedersen (2000). "Codon-substitution models for heterogeneous selection pressure at amino acid sites." Genetics **155**(1): 431-449.

Yang, Z. and B. Rannala (1997). "Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method." Molecular biology and evolution **14**(7): 717-724.

Yang, Z. and D. Roberts (1995). "On the use of nucleic acid sequences to infer early branchings in the tree of life." Molecular biology and evolution **12**(3): 451-458.

Yang, Z., W. S. Wong and R. Nielsen (2005). "Bayes empirical bayes inference of amino acid sites under positive selection." Molecular biology and evolution **22**(4): 1107-1118.

Zanini, F. and R. A. Neher (2013). "Quantifying Selection against Synonymous Mutations in HIV-1 env Evolution." Journal of virology **87**(21): 11843-11850.

Zhang, J., R. Nielsen and Z. Yang (2005). "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level." Molecular biology and evolution **22**(12): 2472-2479.

Zuckerkandl, E. and L. Pauling (1965). Evolutionary divergence and convergence in proteins. Evolving Genes and Proteins, Academic Press, New York**:** 97-166.

Zuckerkandl, E. and L. B. Pauling (1962). "Molecular disease, evolution, and genetic heterogeneity." Horizons in Biochemistry, Academic Press, New York: 189-225.

# Appendix A

## HIV dataset

For each sequence, the HLA type and the country of the patient is given. Sequences belonging to the same patient have common patient IDs.

| Accession number | Patient ID | Subtype | HLA type | Country |
|---|---|---|---|---|
| **AY331282** | 27177 | B | A*0202 A*0301 B*0702 B*1516 | US |
| **AY331283** | 27177 | B | A*0202 A*0301 B*0702 B*1516 | US |
| **AY331284** | 4574 | B | A*2601 A*6802 B*1510 B*1510 | US |
| **AY331285** | 13621 | B | A*6601 A*6802 B*4201 B*5301 | US |
| **AY331286** | 13621 | B | A*6601 A*6802 B*4201 B*5301 | US |
| **AY331287** | 4575 | B | A*2301 A*3001 B*4201 B*5301 | US |
| **AY331289** | 13509 | B | A*6801 A*7401 B*1503 B*5802 | US |
| **AY331290** | 13509 | B | A*6801 A*7401 B*1503 B*5802 | US |
| **AY331293** | 13227 | B | A*0301 A*2301 B*0702 B*1503 | US |
| **AY331296** | 27178 | B | A*0308 A*3001 B*0705 B*4501 | US |
| **AY331297** | 27178 | B | A*0308 A*3001 B*0705 B*4501 | US |
| **AY332236** | 4574 | B | A*2601 A*6802 B*1510 B*1510 | US |
| **AY423381** | 12939 | B | A3 A36 B15 B51 Cw3 Cw6 DR4 DR8 DQ7 | NL |
| **AY423382** | 12939 | B | A3 A36 B15 B51 Cw3 Cw6 DR4 DR8 DQ7 | NL |
| **AY423384** | 12939 | B | A3 A36 B15 B51 Cw3 Cw6 DR4 DR8 DQ7 | NL |
| **AY423385** | 12939 | B | A3 A36 B15 B51 Cw3 Cw6 DR4 DR8 DQ7 | NL |
| **AY423386** | 12939 | B | A3 A36 B15 B51 Cw3 Cw6 DR4 DR8 DQ7 | NL |
| **AY779550** | 9869 | B | A2 A3 B57 B65 | CA |
| **AY779551** | 9869 | B | A2 A3 B57 B65 | CA |
| **AY779552** | 9869 | B | A2 A3 B57 B65 | CA |
| **AY779553** | 6944 | B | A2 A11 B56 B62 Cw1 | CA |
| **AY779554** | 6944 | B | A2 A11 B B62 Cw1 | CA |
| **AY779555** | 6944 | B | A2 A11 B56 B62 Cw1 | CA |
| **AY779557** | 13458 | B | A2 A24 B7 B13 | CA |
| **AY779558** | 13458 | B | A2 A24 B7 B13 | CA |
| **AY779559** | 13458 | B | A2 A24 B7 B13 | CA |
| **AY779560** | 13458 | B | A2 A24 B7 B13 | CA |
| **AY779561** | 13458 | B | A2 A24 B7 B13 | CA |
| **AY779562** | 13458 | B | A2 A24 B7 B13 | CA |
| **AY779563** | 13458 | B | A2 A24 B7 B13 | CA |

| | | | | |
|---|---|---|---|---|
| **AY779564** | 9869 | B | A2 A3 B57 B65 | CA |
| **AY786790** | 9751 | B | A*03 A*31 B*08 B*15 Cw*04 Cw*07 | US |
| **AY786791** | 9751 | B | A*03 A*31 B*08 B*15 Cw*04 Cw*07 | US |
| **AY786792** | 9751 | B | A*03 A*31 B*08 B*15 Cw*04 Cw*07 | US |
| **AY786793** | 9751 | B | A*03 A*31 B*08 B*15 Cw*04 Cw*07 | US |
| **AY786794** | 9751 | B | A*03 A*31 B*08 B*15 Cw*04 Cw*07 | US |
| **AY786795** | 9751 | B | A*03 A*31 B*08 B*15 Cw*04 Cw*07 | US |
| **AY786796** | 9751 | B | A*03 A*31 B*08 B*15 Cw*04 Cw*07 | US |
| **AY786797** | 9751 | B | A*03 A*31 B*08 B*15 Cw*04 Cw*07 | US |
| **AY786798** | 9751 | B | A*03 A*31 B*08 B*15 Cw*04 Cw*07 | US |
| **AY786799** | 9751 | B | A*03 A*31 B*08 B*15 Cw*04 Cw*07 | US |
| **AY786800** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786801** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786802** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786803** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786804** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786805** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786806** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786807** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786808** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786809** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786810** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786811** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786812** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786813** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786814** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786815** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786816** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786817** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786818** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786819** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786820** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786821** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786822** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786823** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786824** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786825** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786826** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786827** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786828** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |
| **AY786829** | 9752 | B | A*24 A*31 B*47 B*15 Cw*04 Cw*07 | US |

| | | | | |
|---|---|---|---|---|
| **EF363125** | 13403 | B | A*02 A*03 B*5703 B*4402/4419 | US |
| **EF363126** | 13403 | B | A*02 A*03 B*5703 B*4402/4419 | US |
| **EU807832** | 28160 | B | A*02 A*30 B*5703 B*2703 | US |
| **EU807833** | 28161 | B | A*330301 A*3402 B*440301 B*5703 | US |
| **EU807834** | 28161 | B | A*330301 A*3402 B*440301 B*5703 | US |
| **EU807835** | 28161 | B | A*330301 A*3402 B*440301 B*5703 | US |
| **EU807836** | 28161 | B | A*330301 A*3402 B*440301 B*5703 | US |
| **EU807837** | 28161 | B | A*330301 A*3402 B*440301 B*5703 | US |
| **EU807838** | 28161 | B | A*330301 A*3402 B*440301 B*5703 | US |
| **FJ495937** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495939** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495940** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495941** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495942** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495943** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495957** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495958** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495961** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495962** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495963** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495973** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495974** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495975** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495976** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495977** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495978** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495979** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495980** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495981** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495991** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495992** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495993** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495994** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495995** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495996** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495997** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495998** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ495999** | 28466 | B | A*0101 A*2301 B*1402 B*5701 Cw*0701 Cw*0802 | US |
| **FJ496000** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496001** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496002** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |

| | | | | |
|---|---|---|---|---|
| **FJ496003** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496004** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496005** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496006** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496007** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496024** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496025** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496026** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496027** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496033** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496034** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496035** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496036** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496037** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496038** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496039** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496040** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496041** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496058** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496059** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496068** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496069** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496070** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496071** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496123** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496128** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496130** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496131** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496132** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496133** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496134** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496135** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ496136** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ919955** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ919956** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ919957** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ919958** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ919959** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ919960** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ919961** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |
| **FJ919962** | 28467 | B | A*0205 A*0205 B*5301 B*5701 Cw*0401 Cw*1801 | US |

| | | | | |
|---|---|---|---|---|
| **JF320363** | 39593 | B | B*5701 CW*0702 A*0301 CW*0602 A*0101 B*0702 | US |
| **JF320364** | 39593 | B | B*5701 CW*0702 A*0301 CW*0602 A*0101 B*0702 | US |
| **JF320365** | 39593 | B | B*5701 CW*0702 A*0301 CW*0602 A*0101 B*0702 | US |
| **JF320366** | 39593 | B | B*5701 CW*0702 A*0301 CW*0602 A*0101 B*0702 | US |
| **JF320369** | 39593 | B | B*5701 CW*0702 A*0301 CW*0602 A*0101 B*0702 | US |
| **JF320373** | 39593 | B | B*5701 CW*0702 A*0301 CW*0602 A*0101 B*0702 | US |
| **JF320374** | 39593 | B | B*5701 CW*0702 A*0301 CW*0602 A*0101 B*0702 | US |
| **JF320514** | 39593 | B | B*5701 CW*0702 A*0301 CW*0602 A*0101 B*0702 | US |
| **JF320515** | 39593 | B | B*5701 CW*0702 A*0301 CW*0602 A*0101 B*0702 | US |
| **JF320516** | 39593 | B | B*5701 CW*0702 A*0301 CW*0602 A*0101 B*0702 | US |
| **JF320518** | 39593 | B | B*5701 CW*0702 A*0301 CW*0602 A*0101 B*0702 | US |
| **JF320519** | 39593 | B | B*5701 CW*0702 A*0301 CW*0602 A*0101 B*0702 | US |
| **JF320559** | 39645 | B | B*5701 CW*0802 A*2910 CW*0602 A*0101 B*1402 | US |
| **JF320561** | 39645 | B | B*5701 CW*0802 A*2910 CW*0602 A*0101 B*1402 | US |
| **JF320562** | 39645 | B | B*5701 CW*0802 A*2910 CW*0602 A*0101 B*1402 | US |
| **JF320563** | 39645 | B | B*5701 CW*0802 A*2910 CW*0602 A*0101 B*1402 | US |
| **AY786830** | 9753 | B | A*30 B*18 B*40 Cw*02 Cw*05 | US |
| **AY786831** | 9753 | B | A*30 B*18 B*40 Cw*02 Cw*05 | US |
| **AY786832** | 9753 | B | A*30 B*18 B*40 Cw*02 Cw*05 | US |
| **AY786833** | 9753 | B | A*30 B*18 B*40 Cw*02 Cw*05 | US |
| **AY786834** | 9753 | B | A*30 B*18 B*40 Cw*02 Cw*05 | US |
| **AY786835** | 9753 | B | A*30 B*18 B*40 Cw*02 Cw*05 | US |
| **AY786836** | 9753 | B | A*30 B*18 B*40 Cw*02 Cw*05 | US |
| **AY786837** | 9753 | B | A*30 B*18 B*40 Cw*02 Cw*05 | US |
| **AY786838** | 9753 | B | A*30 B*18 B*40 Cw*02 Cw*05 | US |
| **AY786839** | 9753 | B | A*30 B*18 B*40 Cw*02 Cw*05 | US |
| **AY786840** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786841** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786842** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786843** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786844** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786845** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786846** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786847** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786848** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786849** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786850** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786851** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786852** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786853** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786854** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |

| | | | | |
|---|---|---|---|---|
| **AY786855** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786856** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786857** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786858** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786859** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786860** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786861** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786862** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786863** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786864** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786865** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786866** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786867** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786868** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786869** | 9754 | B | A*02 A*30 B*18 B*13 Cw*01 Cw*05 | US |
| **AY786870** | 9755 | B | A*24 A*30 B*39 B*47 Cw*12 Cw*17 | US |
| **AY786871** | 9755 | B | A*24 A*30 B*39 B*47 Cw*12 Cw*17 | US |
| **AY786872** | 9755 | B | A*24 A*30 B*39 B*47 Cw*12 Cw*17 | US |
| **AY786873** | 9755 | B | A*24 A*30 B*39 B*47 Cw*12 Cw*17 | US |
| **AY786874** | 9755 | B | A*24 A*30 B*39 B*47 Cw*12 Cw*17 | US |
| **AY786875** | 9755 | B | A*24 A*30 B*39 B*47 Cw*12 Cw*17 | US |
| **AY786876** | 9755 | B | A*24 A*30 B*39 B*47 Cw*12 Cw*17 | US |
| **AY786877** | 9755 | B | A*24 A*30 B*39 B*47 Cw*12 Cw*17 | US |
| **AY786878** | 9755 | B | A*24 A*30 B*39 B*47 Cw*12 Cw*17 | US |
| **AY786879** | 9755 | B | A*24 A*30 B*39 B*47 Cw*12 Cw*17 | US |
| **AY786880** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786881** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786882** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786883** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786884** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786885** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786886** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786887** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786888** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786889** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786890** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786891** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786892** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786893** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786894** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786895** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |

| | | | | |
|---|---|---|---|---|
| **AY786896** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786897** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786898** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786899** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786900** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786901** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786902** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786903** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **AY786904** | 15306 | B | A*24 A*23 B*39 B*07 Cw*12 Cw*17 | US |
| **DQ487190** | 21249 | B | A1 A19 B*3501 B44 Cw7 Cw16 | US |
| **DQ487191** | 21249 | B | A1 A19 B*3501 B44 Cw7 Cw16 | US |
| **HM208363** | 36113 | B | B*27, B*35, A*24, A*30 | US |
| **HM586191** | 36106 | B | A*02 A*03 B*15 B*35 Cw*09 Cw*04 | GB |
| **HM586193** | 36106 | B | A*02 A*03 B*15 B*35 Cw*09 Cw*04 | GB |
| **HM586194** | 36106 | B | A*02 A*03 B*15 B*35 Cw*09 Cw*04 | GB |
| **HM586196** | 36106 | B | A*02 A*03 B*15 B*35 Cw*09 Cw*04 | GB |
| **JF320028** | 39682 | B | B*4064 CW*0401 A*2402 CW*0304 A*0201 B*3520 | PE |
| **JF320029** | 39682 | B | B*4064 CW*0401 A*2402 CW*0304 A*0201 B*3520 | PE |
| **JF320031** | 39682 | B | B*4064 CW*0401 A*2402 CW*0304 A*0201 B*3520 | PE |
| **JF320032** | 39682 | B | B*4064 CW*0401 A*2402 CW*0304 A*0201 B*3520 | PE |
| **JF320044** | 39591 | B | B*3501 CW*0401 A*2402 CW*0303 A*0206 B*2705 | US |
| **JF320047** | 39591 | B | B*3501 CW*0401 A*2402 CW*0303 A*0206 B*2705 | US |
| **JF320048** | 39733 | B | B*4901 CW*0701 A*2402 CW*0102 A*2301 B*3543 | US |
| **JF320049** | 39591 | B | B*3501 CW*0401 A*2402 CW*0303 A*0206 B*2705 | US |
| **JF320051** | 39591 | B | B*3501 CW*0401 A*2402 CW*0303 A*0206 B*2705 | US |
| **JF320053** | 39591 | B | B*3501 CW*0401 A*2402 CW*0303 A*0206 B*2705 | US |
| **JF320055** | 39591 | B | B*3501 CW*0401 A*2402 CW*0303 A*0206 B*2705 | US |
| **JF320056** | 39733 | B | B*4901 CW*0701 A*2402 CW*0102 A*2301 B*3543 | US |
| **JF320058** | 39591 | B | B*3501 CW*0401 A*2402 CW*0303 A*0206 B*2705 | US |
| **JF320059** | 39731 | B | B*3501 CW*0401 A*2402 CW*0202 A*0201 B*2705 | US |
| **JF320060** | 39591 | B | B*3501 CW*0401 A*2402 CW*0303 A*0206 B*2705 | US |
| **JF320062** | 39591 | B | B*3501 CW*0401 A*2402 CW*0303 A*0206 B*2705 | US |
| **JF320064** | 39591 | B | B*3501 CW*0401 A*2402 CW*0303 A*0206 B*2705 | US |
| **JF320065** | 39733 | B | B*4901 CW*0701 A*2402 CW*0102 A*2301 B*3543 | US |
| **JF320068** | 39733 | B | B*4901 CW*0701 A*2402 CW*0102 A*2301 B*3543 | US |
| **JF320071** | 39733 | B | B*4901 CW*0701 A*2402 CW*0102 A*2301 B*3543 | US |
| **JF320145** | 39631 | B | B*5305 A*6802 CW*0401 A*0301 B*3527 | US |
| **JF320147** | 39631 | B | B*5305 A*6802 CW*0401 A*0301 B*3527 | US |
| **JF320152** | 39631 | B | B*5305 A*6802 CW*0401 A*0301 B*3527 | US |
| **JF320153** | 39631 | B | B*5305 A*6802 CW*0401 A*0301 B*3527 | US |
| **JF320154** | 39631 | B | B*5305 A*6802 CW*0401 A*0301 B*3527 | US |

| | | | | |
|---|---|---|---|---|
| **JF320179** | 39693 | B | B*3501 CW*0701 A*1101 CW*0401 A*0101 B*0801 | US |
| **JF320181** | 39693 | B | B*3501 CW*0701 A*1101 CW*0401 A*0101 B*0801 | US |
| **JF320182** | 39693 | B | B*3501 CW*0701 A*1101 CW*0401 A*0101 B*0801 | US |
| **JF320183** | 39708 | B | B*5701 CW*0602 A*9205 CW*0304 A*0101 B*3501 | PE |
| **JF320185** | 39730 | B | B*3501 CW*0602 CW*0202 A*6802 B*1801 | US |
| **JF320186** | 39620 | B | B*3543 CW*0401 A*3101 CW*0102 A*0211 B*3520 | PE |
| **JF320187** | 39708 | B | B*5701 CW*0602 A*9205 CW*0304 A*0101 B*3501 | PE |
| **JF320188** | 39730 | B | B*3501 CW*0602 CW*0202 A*6802 B*1801 | US |
| **JF320190** | 39730 | B | B*3501 CW*0602 CW*0202 A*6802 B*1801 | US |
| **JF320192** | 39730 | B | B*3501 CW*0602 CW*0202 A*6802 B*1801 | US |
| **JF320193** | 39620 | B | B*3543 CW*0401 A*3101 CW*0102 A*0211 B*3520 | PE |
| **JF320194** | 39730 | B | B*3501 CW*0602 CW*0202 A*6802 B*1801 | US |
| **JF320197** | 39615 | B | B*3517 CW*0702 A*2601 CW*0401 A*0201 B*0702 | US |
| **JF320200** | 39615 | B | B*3517 CW*0702 A*2601 CW*0401 A*0201 B*0702 | US |
| **JF320201** | 39620 | B | B*3543 CW*0401 A*3101 CW*0102 A*0211 B*3520 | PE |
| **JF320202** | 39615 | B | B*3517 CW*0702 A*2601 CW*0401 A*0201 B*0702 | US |
| **JF320205** | 39615 | B | B*3517 CW*0702 A*2601 CW*0401 A*0201 B*0702 | US |
| **JF320207** | 39615 | B | B*3517 CW*0702 A*2601 CW*0401 A*0201 B*0702 | US |
| **JF320209** | 39708 | B | B*5701 CW*0602 A*9205 CW*0304 A*0101 B*3501 | PE |
| **JF320212** | 39708 | B | B*5701 CW*0602 A*9205 CW*0304 A*0101 B*3501 | PE |
| **JF320215** | 39620 | B | B*3543 CW*0401 A*3101 CW*0102 A*0211 B*3520 | PE |
| **JF320307** | 39669 | B | B*3501 A*2402 CW*0303 A*0201 B*1501 | US |
| **JF320309** | 39669 | B | B*3501 A*2402 CW*0303 A*0201 B*1501 | US |
| **JF320311** | 39669 | B | B*3501 A*2402 CW*0303 A*0201 B*1501 | US |
| **JF320381** | 39719 | B | B*4901 CW*0701 A*6801 CW*0401 A*0205 B*3501 | US |
| **JF320384** | 39719 | B | B*4901 CW*0701 A*6801 CW*0401 A*0205 B*3501 | US |
| **JF320409** | 39664 | B | B*5101 CW*1502 A*3101 CW*0401 A*1101 B*3503 | US |
| **JF320460** | 39608 | B | B*5101 CW*1502 A*2402 C*0404 A*2301 B*3502 | US |
| **JF320461** | 39608 | B | B*5101 CW*1502 A*2402 C*0404 A*2301 B*3502 | US |
| **JF320462** | 39608 | B | B*5101 CW*1502 A*2402 C*0404 A*2301 B*3502 | US |
| **JF320463** | 39608 | B | B*5101 CW*1502 A*2402 C*0404 A*2301 B*3502 | US |
| **JF320464** | 39608 | B | B*5101 CW*1502 A*2402 C*0404 A*2301 B*3502 | US |
| **JF320465** | 39608 | B | B*5101 CW*1502 A*2402 C*0404 A*2301 B*3502 | US |
| **JF320466** | 39608 | B | B*5101 CW*1502 A*2402 C*0404 A*2301 B*3502 | US |
| **JF320467** | 39608 | B | B*5101 CW*1502 A*2402 C*0404 A*2301 B*3502 | US |
| **JF320468** | 39608 | B | B*5101 CW*1502 A*2402 C*0404 A*2301 B*3502 | US |
| **JF320469** | 39608 | B | B*5101 CW*1502 A*2402 C*0404 A*2301 B*3502 | US |
| **JF320470** | 39608 | B | B*5101 CW*1502 A*2402 C*0404 A*2301 B*3502 | US |
| **JF320569** | 39664 | B | B*5101 CW*1502 A*3101 CW*0401 A*1101 B*3503 | US |
| **JF320571** | 39664 | B | B*5101 CW*1502 A*3101 CW*0401 A*1101 B*3503 | US |
| **JF320572** | 39664 | B | B*5101 CW*1502 A*3101 CW*0401 A*1101 B*3503 | US |

| | | | | |
|---|---|---|---|---|
| **JF320573** | 39664 | B | B*5101 CW*1502 A*3101 CW*0401 A*1101 B*3503 | US |
| **JF320574** | 39664 | B | B*5101 CW*1502 A*3101 CW*0401 A*1101 B*3503 | US |
| **JF320575** | 39664 | B | B*5101 CW*1502 A*3101 CW*0401 A*1101 B*3503 | US |
| **JF320576** | 39664 | B | B*5101 CW*1502 A*3101 CW*0401 A*1101 B*3503 | US |
| **JF320577** | 39664 | B | B*5101 CW*1502 A*3101 CW*0401 A*1101 B*3503 | US |
| **JF320582** | 39664 | B | B*5101 CW*1502 A*3101 CW*0401 A*1101 B*3503 | US |
| **JF320591** | 39664 | B | B*5101 CW*1502 A*3101 CW*0401 A*1101 B*3503 | US |
| **JF320615** | 39734 | B | B*4501 CW*0602 A*2902 CW*0401 A*1101 B*3501 | US |
| **JF320617** | 39734 | B | B*4501 CW*0602 A*2902 CW*0401 A*1101 B*3501 | US |
| **JF320620** | 39734 | B | B*4501 CW*0602 A*2902 CW*0401 A*1101 B*3501 | US |
| **JF320621** | 39734 | B | B*4501 CW*0602 A*2902 CW*0401 A*1101 B*3501 | US |
| **JF320623** | 39734 | B | B*4501 CW*0602 A*2902 CW*0401 A*1101 B*3501 | US |
| **JF320624** | 39734 | B | B*4501 CW*0602 A*2902 CW*0401 A*1101 B*3501 | US |
| **JF320625** | 39734 | B | B*4501 CW*0602 A*2902 CW*0401 A*1101 B*3501 | US |
| **JF320626** | 39734 | B | B*4501 CW*0602 A*2902 CW*0401 A*1101 B*3501 | US |
| **JF320627** | 39734 | B | B*4501 CW*0602 A*2902 CW*0401 A*1101 B*3501 | US |

# Appendix B

## Newick format of phylogenetic trees T1, T2, and T3.

**T1**

((((((FJ496035:2,(((((FJ496040:2,(FJ496041:2,FJ496033:2):2):2,FJ496036:2):2,FJ496038:2):2,FJ496034:2):2,FJ496037:2):2):2,((((FJ919955:2,(FJ919958:2,FJ919959:2):2):2,FJ919960:2):2,FJ919956:2):2,((FJ919961:2,FJ919962:2):2,FJ919957:2):2):2):2,(((FJ496025:2,FJ496027:2):2,FJ496039:2):2,(FJ496134:2,(FJ496058:2,(FJ496069:2,(FJ496128:2,(FJ496123:2,(FJ496071:2,(FJ496135:2,(FJ496131:2,(FJ496059:2,(FJ496132:2,(FJ496005:2,(FJ496003:2,(FJ496004:2,(FJ496133:2,(FJ496000:2,(((((FJ496068:2,FJ496136:2):2,(FJ496002:2,FJ496006:2):2):2,(FJ496007:2,FJ496001:2):2):2,(FJ496024:2,FJ496026:2):2):2,(FJ496070:2,FJ496130:2):2):2):2):2):2):2):2):2):2):2):2):2):2):2):2):2):2):2):2,(AY331290:1,AY331289:1):1):0,((((((((((AY786889:1,AY786883:1):1,(((AY786885:1,AY786880:1):1,AY786888:1):1,(((((((AY786895:1,(AY786893:1,AY786896:1):1):1,AY786894:1):1,AY786891:1):1,AY786899:1):1,AY786890:1):1,AY786898:1):1,AY786897:1):1,AY786892:1):1):1):1,AY786904:1):1,(((AY786881:1,(AY786886:1,AY786887:1):1):1,AY786884:1):1,(AY786900:1,AY786901:1):1):1):1,(((((AY779562:1,AY779557:1):1,AY779558:1):1,AY779563:1):1,(AY779560:1,(AY779561:1,AY779559:1):1):1):1,(EF363126:2,EF363125:2):2):0):0,((((JF320200:3,JF320202:3):3,JF320197:3):3,JF320207:3):3,JF320205:3):3):0,((((((AY423382:1,AY423384:1):1,AY423386:1):1,AY423385:1):1,AY423381:1):1,((((((((((AY786831:1,AY786837:1):1,AY786839:1):1,AY786832:1):1,AY786830:1):1,AY786835:1):1,AY786833:1):1,AY786834:1):1,AY786838:1):1,AY786836:1):1,((((((((((AY786843:1,(AY786842:1,AY786847:1):1):1,AY786841:1):1,AY786840:1):1,(AY786844:1,AY786848:1):1):1,AY786845:1):1,AY786849:1):1,(AY786852:1,AY786854:1):1):1,AY786846:1):1,AY786851:1):1,((AY786869:1,(AY786856:1,AY786853:1):1):1,AY786866:1):1):1,(((AY786868:1,AY786867:1):1,AY786863:1):1,(AY786861:1,AY786864:1):1):1):1,AY786860:1):1):0):0,(((AY331282:1,AY331283:1):1,((DQ487191:3,DQ487190:3):3,((((((EU807836:2,EU807838:2):2,EU807835:2):2,EU807834:2):2,EU807833:2):2,EU807837:2):2,EU807832:2):0):0):0,((JF320307:3,JF320309:3):3,JF320311:3):3):0):0):0,(((((JF320627:3,(JF320626:3,(JF320615:3,(JF320617:3,(JF320620:3,(JF320623:3,(JF320624:3,JF320625:3):3):3):3):3):3):3):3):3,JF320621:3):3,(JF320059:3,(((HM586193:3,HM586191:3):3,HM586194:3):3,HM586196:3):3):0):0,((JF320062:3,((((JF320051:3,JF320064:3):3,JF320047:3):3,JF320044:3):3,JF320049:3):3):3,((JF320058:3,(JF320060:3,JF320053:3):3):3,JF320055:3):3):3):0,((((JF320179:3,JF320182:3):3,JF320181:3):3,((JF320187:2,(JF320209:2,JF320212:2):2):2,JF320183:2):2):0,(((JF320193:3,JF320186:3):3,JF320215:3):3,JF320201:3):3):0):0):0,(((((JF320563:2,JF320562:2):2,JF320559:2):2,JF320561:2):2,(((JF320031:3,JF320028:3):3,(JF320029:3,JF320032:3):3):3,(((FJ495963:2,FJ495961:2):2,FJ495962:2):2,((((((((((((FJ495937:2,(FJ495978:2,(FJ495994:2,(FJ495979:2,FJ495975:2):2):2):2):2,FJ495974:2):2,FJ495981:2):2,FJ4

95941:2):2,FJ495939:2):2,FJ495943:2):2,FJ495973:2):2,FJ495940:2):2,FJ495942:2):2,(((((FJ49
5977:2,FJ495992:2):2,FJ495995:2):2,FJ495997:2):2,FJ495998:2):2,FJ495996:2):2):2,((FJ49599
1:2,(FJ495958:2,FJ495993:2):2):2,FJ495976:2):2):2,FJ495980:2):2,(FJ495999:2,FJ495957:2):2)
:2):2):0):0,((AY331293:1,HM208363:3):0,((((((JF320463:3,JF320466:3):3,(((JF320461:3,JF320
468:3):3,(((JF320467:3,JF320460:3):3,JF320469:3):3,JF320470:3):3):3,JF320464:3):3):3,(JF320
465:3,JF320462:3):3):3,(AY331284:1,AY332236:1):1):0,(JF320071:3,(((JF320048:3,JF320056:
3):3,JF320068:3):3,JF320065:3):3):3):0,(((((JF320518:2,(JF320519:2,(JF320364:2,(JF320366:2,
(JF320365:2,JF320363:2):2):2):2):2):2,JF320373:2):2,JF320369:2):2,(((JF320516:2,JF320514:2
):2,JF320515:2):2,JF320374:2):2):2,(JF320384:3,JF320381:3):3):0):0):0):0,(AY331286:1,AY
331285:1):1):0,((((AY779553:1,AY779554:1):1,AY779555:1):1,((AY779551:1,(AY779564:1,A
Y779550:1):1):1,AY779552:1):1):0,((JF320154:3,(JF320147:3,(JF320145:3,JF320153:3):3):3):3
,JF320152:3):3):0):0,(((JF320185:3,(((JF320188:3,JF320190:3):3,JF320194:3):3,JF320192:3):3)
:3,AY331287:1):0,((((((JF320569:3,JF320582:3):3,JF320409:3):3,JF320591:3):3,(JF320572:3,J
F320571:3):3):3,(((JF320576:3,JF320573:3):3,JF320574:3):3,JF320575:3):3):3,JF320577:3):3):
0):0):0,((((((((((((AY786800:1,AY786802:1):1,AY786803:1):1,AY786804:1):1,AY786808:1):1,
AY786801:1):1,AY786805:1):1,(AY786809:1,AY786807:1):1):1,AY786806:1):1,(((((((((AY78
6813:1,AY786817:1):1,AY786810:1):1,AY786811:1):1,AY786818:1):1,AY786819:1):1,AY786
814:1):1,AY786815:1):1,AY786812:1):1,AY786816:1):1):1,((((((AY786823:1,((AY786829:1,A
Y786824:1):1,AY786821:1):1):1,AY786828:1):1,(AY786825:1,AY786820:1):1):1,AY786826:1
):1,AY786827:1):1,AY786822:1):1):1,((AY786792:1,((((AY786793:1,AY786795:1):1,AY7867
96:1):1,AY786798:1):1,AY786799:1):1):1,(((AY786794:1,AY786790:1):1,AY786791:1):1,AY7
86797:1):1):1):0,(AY331296:1,AY331297:1):1):0);

**T2**

(AY331282:1.0,AY331283:1.0,(((AY331284:1.0,AY332236:1.0):1.0,((AY331287:1.0,(HM5861
91:3.0,HM586193:3.0,HM586194:3.0,HM586196:3.0):3.0):0.0,(JF320185:3.0,JF320188:3.0,JF3
20190:3.0,JF320192:3.0,JF320194:3.0):3.0):0.0):0.0,(AY331285:1.0,AY331286:1.0):1.0,((AY3
31289:1.0,AY331290:1.0):1.0,(((((AY786880:1.0,(AY786882:1.0,AY786885:1.0):1.0,AY78688
8:1.0):1.0,AY786902:1.0):1.0,(((AY786881:1.0,(AY786886:1.0,AY786887:1.0):1.0):1.0,AY786
884:1.0):1.0,(AY786890:1.0,AY786891:1.0,(AY786892:1.0,AY786893:1.0,AY786895:1.0,AY7
86896:1.0,AY786899:1.0):1.0,AY786894:1.0,AY786897:1.0,AY786898:1.0):1.0):1.0,(AY7868
83:1.0,AY786889:1.0):1.0):1.0,(AY786900:1.0,AY786901:1.0):1.0,(AY786903:1.0,AY786904:
1.0):1.0):1.0,(AY786879:1.0,AY786871:1.0,AY786870:1.0,AY786875:1.0,AY786872:1.0,AY7
86877:1.0,AY786878:1.0,AY786873:1.0,AY786874:1.0,AY786876:1.0):1.0):0.0):0.0,(AY3312
93:1.0,(AY331296:1.0,AY331297:1.0):1.0):0.0,(AY423381:1.0,AY423382:1.0,(AY423384:1.0,(
AY423385:1.0,AY423386:1.0):1.0):1.0):1.0,(((((AY779550:1.0,AY779551:1.0):1.0,AY779564:

1.0):1.0,AY779552:1.0):1.0,((AY779553:1.0,AY779555:1.0):1.0,AY779554:1.0):1.0):0.0,((((((
AY779557:1.0,AY779562:1.0):1.0,AY779558:1.0):1.0,AY779563:1.0):1.0,AY779560:1.0):1.0,
AY779561:1.0):1.0,AY779559:1.0):1.0):0.0,(((((AY786790:1.0,AY786791:1.0,AY786794:1.0,
AY786797:1.0):1.0,(AY786792:1.0,(AY786793:1.0,AY786795:1.0,AY786796:1.0,AY786798:1
.0,AY786799:1.0):1.0):1.0):1.0,((AY786800:1.0,AY786801:1.0,AY786802:1.0,AY786803:1.0,
AY786804:1.0,AY786805:1.0,AY786806:1.0,(AY786807:1.0,AY786809:1.0):1.0,AY786808:1.
0):1.0,(AY786810:1.0,AY786811:1.0,AY786812:1.0,AY786813:1.0,AY786814:1.0,AY786815:
1.0,AY786816:1.0,AY786817:1.0,AY786818:1.0,AY786819:1.0,((AY786820:1.0,AY786822:1.
0):1.0,((AY786821:1.0,AY786827:1.0,AY786829:1.0):1.0,AY786823:1.0,AY786824:1.0,AY78
6826:1.0,AY786828:1.0):1.0,AY786825:1.0):1.0):1.0):1.0):0.0,(JF320615:3.0,JF320617:3.0,JF3
20620:3.0,JF320621:3.0,JF320623:3.0,JF320624:3.0,JF320625:3.0,JF320626:3.0,JF320627:3.0):
3.0):0.0,((AY786830:1.0,AY786831:1.0,AY786832:1.0,AY786833:1.0,AY786834:1.0,(AY7868
35:1.0,AY786837:1.0):1.0,AY786836:1.0,AY786838:1.0,AY786839:1.0):1.0,((((AY786840:1.0,
AY786841:1.0,(AY786842:1.0,AY786847:1.0):1.0,AY786843:1.0,(AY786844:1.0,AY786848:1
.0):1.0,AY786845:1.0,AY786849:1.0):1.0,AY786852:1.0):1.0,AY786854:1.0):1.0,AY786846:1.
0,AY786851:1.0,((AY786853:1.0,AY786856:1.0):1.0,AY786869:1.0):1.0,AY786860:1.0,((AY7
86861:1.0,AY786864:1.0,AY786868:1.0):1.0,AY786867:1.0):1.0,(AY786863:1.0,AY786866:1.
0):1.0,(AY786865:1.0,AY786858:1.0,AY786850:1.0):1.0,((AY786859:1.0,AY786857:1.0):1.0,
AY786855:1.0,AY786862:1.0):1.0):1.0):0.0):0.0,(EF363125:2.0,EF363126:2.0):2.0,((EU807832
:2.0,((EU807833:2.0,EU807834:2.0,EU807835:2.0,EU807836:2.0,EU807838:2.0):2.0,EU80783
7:2.0):2.0):0.0,(DQ487190:3.0,DQ487191:3.0):3.0):0.0,((FJ495937:2.0,FJ495939:2.0,FJ495940:
2.0,FJ495941:2.0,FJ495942:2.0,FJ495943:2.0,((FJ495957:2.0,FJ495999:2.0):2.0,FJ495977:2.0,F
J495992:2.0,FJ495995:2.0,FJ495996:2.0,FJ495997:2.0,FJ495998:2.0):2.0,(FJ495958:2.0,((FJ49
5961:2.0,FJ495963:2.0):2.0,FJ495962:2.0):2.0,FJ495976:2.0,FJ495991:2.0,FJ495993:2.0):2.0,FJ
495973:2.0,FJ495974:2.0,FJ495975:2.0,FJ495978:2.0,FJ495979:2.0,FJ495980:2.0,FJ495981:2.0
,FJ495994:2.0):2.0,(FJ496000:2.0,FJ496001:2.0,FJ496002:2.0,FJ496003:2.0,FJ496004:2.0,FJ49
6005:2.0,(FJ496006:2.0,FJ496058:2.0):2.0,FJ496007:2.0,FJ496024:2.0,((FJ496025:2.0,(FJ4960
33:2.0,(FJ496034:2.0,FJ496036:2.0):2.0,FJ496035:2.0,FJ496037:2.0,FJ496038:2.0,FJ496040:2.
0,FJ496041:2.0,(FJ919955:2.0,FJ919956:2.0,(((FJ919957:2.0,FJ919962:2.0):2.0,FJ919961:2.0):
2.0,FJ919960:2.0):2.0,FJ919958:2.0,FJ919959:2.0):2.0):2.0):2.0,FJ496027:2.0,FJ496039:2.0):2.
0,FJ496026:2.0,FJ496059:2.0,FJ496068:2.0,FJ496069:2.0,FJ496070:2.0,FJ496071:2.0,FJ49612
3:2.0,FJ496128:2.0,FJ496130:2.0,FJ496131:2.0,FJ496132:2.0,FJ496133:2.0,FJ496134:2.0,FJ49
6135:2.0,FJ496136:2.0):2.0):0.0,(((JF320363:2.0,JF320364:2.0,JF320366:2.0,JF320519:2.0):2.0,
JF320365:2.0,JF320369:2.0,JF320373:2.0,JF320374:2.0,JF320514:2.0,JF320515:2.0,JF320516:
2.0,JF320518:2.0):2.0,(HM208363:3.0,(JF320381:3.0,JF320384:3.0):3.0):0.0):0.0,(((JF320559:2
.0,JF320561:2.0,JF320562:2.0,JF320563:2.0):2.0,((((JF320460:3.0,JF320464:3.0):3.0,((JF32046
1:3.0,JF320468:3.0):3.0,JF320462:3.0,JF320463:3.0,JF320465:3.0,JF320466:3.0):3.0,JF320467:
3.0):3.0,JF320469:3.0):3.0,JF320470:3.0):3.0):0.0,(JF320028:3.0,JF320029:3.0,JF320031:3.0,JF
320032:3.0):3.0):0.0,(((((JF320044:3.0,JF320047:3.0,JF320049:3.0,JF320051:3.0,JF320062:3.0)

:3.0,JF320064:3.0):3.0,((JF320053:3.0,JF320058:3.0,JF320060:3.0):3.0,JF320055:3.0):3.0):3.0,(
JF320145:3.0,JF320147:3.0,JF320152:3.0,JF320153:3.0,JF320154:3.0):3.0):0.0,((JF320048:3.0,
JF320056:3.0,JF320065:3.0,JF320068:3.0,JF320071:3.0):3.0,(JF320179:3.0,JF320181:3.0,JF320
182:3.0):3.0):0.0,((JF320059:3.0,(JF320307:3.0,JF320309:3.0,JF320311:3.0):3.0):0.0,((JF32019
7:3.0,JF320200:3.0,JF320202:3.0,JF320207:3.0):3.0,JF320205:3.0):3.0):0.0):0.0,((JF320183:2.0
,(JF320187:2.0,JF320209:2.0,JF320212:2.0):2.0):2.0,(JF320186:3.0,JF320193:3.0,JF320201:3.0,
JF320215:3.0):3.0):0.0,(((((JF320409:3.0,JF320569:3.0,JF320582:3.0):3.0,JF320591:3.0):3.0,(J
F320571:3.0,JF320577:3.0):3.0,JF320572:3.0):3.0,JF320575:3.0):3.0,(JF320573:3.0,JF320574:3
.0,JF320576:3.0):3.0):3.0):0.0);

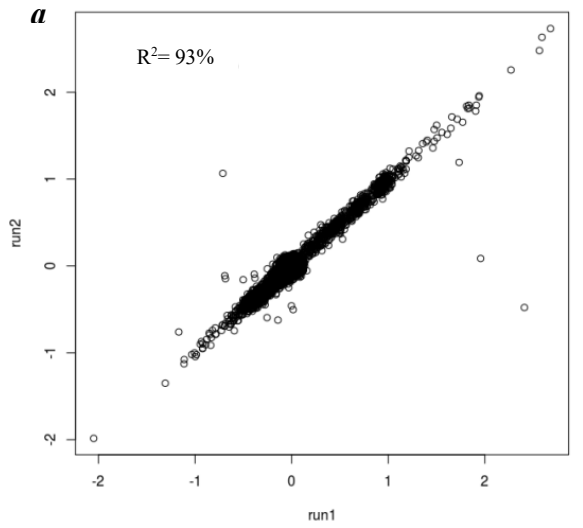
**T3**

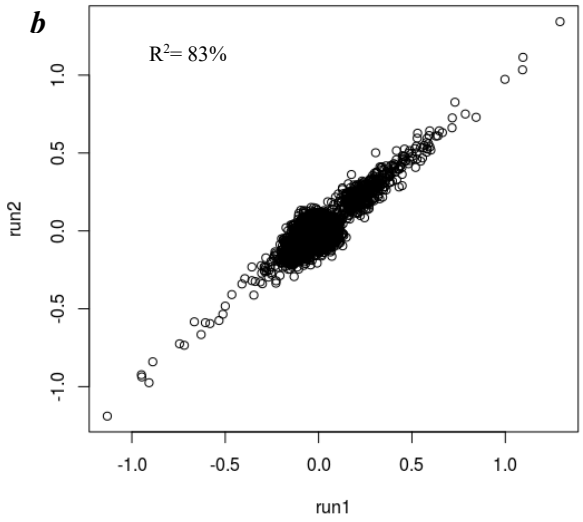
(AY331282:1.0,AY331283:1.0,(((AY331284:1.0,AY332236:1.0):1.0,((AY331287:1.0,((HM586
191:3.0,HM586193:3.0,HM586194:3.0):3.0,HM586196:3.0):3.0):0.0,(JF320185:3.0,JF320188:3
.0,JF320190:3.0,JF320192:3.0,JF320194:3.0):3.0):0.0):0.0,(AY331285:1.0,AY331286:1.0):1.0,(
(AY331289:1.0,AY331290:1.0):1.0,((AY786880:1.0,(AY786882:1.0,AY786885:1.0):1.0,AY78
6888:1.0):1.0,(((AY786881:1.0,(AY786886:1.0,AY786887:1.0):1.0):1.0,AY786884:1.0):1.0,(A
Y786890:1.0,AY786891:1.0,(AY786892:1.0,AY786893:1.0,AY786895:1.0,AY786896:1.0,AY7
86899:1.0):1.0,AY786894:1.0,AY786897:1.0,AY786898:1.0):1.0):1.0,(AY786883:1.0,AY7868
89:1.0):1.0,((AY786900:1.0,AY786901:1.0):1.0,(AY786903:1.0,AY786904:1.0):1.0):1.0,AY78
6902:1.0):1.0):0.0,AY331293:1.0,(AY331296:1.0,AY331297:1.0):1.0,(AY423381:1.0,AY42338
2:1.0,(AY423384:1.0,(AY423385:1.0,AY423386:1.0):1.0):1.0):1.0,((((AY779550:1.0,AY77955
1:1.0):1.0,AY779564:1.0):1.0,AY779552:1.0):1.0,((((((AY779557:1.0,AY779562:1.0):1.0,AY7
79558:1.0):1.0,AY779563:1.0):1.0,AY779560:1.0):1.0,AY779561:1.0):1.0,AY779559:1.0):1.0):
0.0,(((((AY786790:1.0,AY786791:1.0,AY786794:1.0,AY786797:1.0):1.0,(AY786792:1.0,(AY7
86793:1.0,AY786795:1.0,AY786796:1.0,AY786798:1.0,AY786799:1.0):1.0):1.0):1.0,((AY7868
00:1.0,AY786801:1.0,AY786802:1.0,AY786803:1.0,AY786804:1.0,AY786805:1.0,AY786806:
1.0,(AY786807:1.0,AY786809:1.0):1.0,AY786808:1.0):1.0,(AY786810:1.0,AY786811:1.0,AY7
86812:1.0,AY786813:1.0,AY786814:1.0,AY786815:1.0,AY786816:1.0,AY786817:1.0,AY7868
18:1.0,AY786819:1.0,((AY786820:1.0,AY786822:1.0):1.0,((AY786821:1.0,AY786827:1.0,AY
786829:1.0):1.0,AY786823:1.0,AY786824:1.0,AY786826:1.0,AY786828:1.0):1.0,AY786825:1.
0):1.0):1.0):1.0):0.0,(JF320615:3.0,JF320617:3.0,JF320620:3.0,JF320621:3.0,JF320623:3.0,JF3
20624:3.0,JF320625:3.0,JF320626:3.0,JF320627:3.0):3.0):0.0,((AY786865:1.0,AY786858:1.0,
AY786850:1.0):1.0,(((AY786840:1.0,AY786841:1.0,(AY786842:1.0,AY786847:1.0):1.0,AY78
6843:1.0,(AY786844:1.0,AY786848:1.0):1.0,AY786845:1.0,AY786849:1.0):1.0,AY786852:1.0
):1.0,AY786854:1.0):1.0,AY786846:1.0,AY786851:1.0,((AY786853:1.0,AY786856:1.0):1.0,A
Y786869:1.0):1.0,AY786860:1.0,(AY786861:1.0,AY786864:1.0,AY786868:1.0):1.0,(AY78686

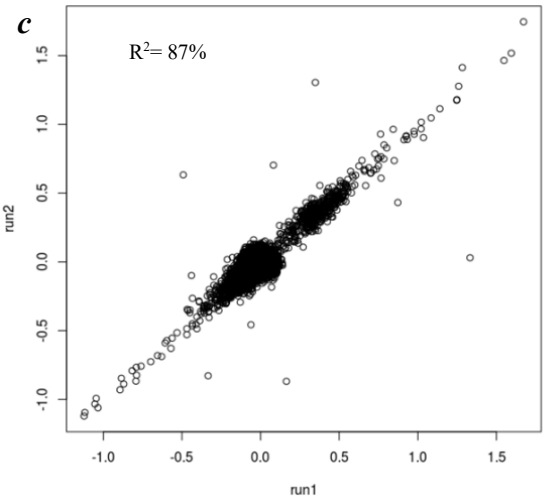
3:1.0,AY786866:1.0):1.0,AY786867:1.0,((AY786859:1.0,AY786857:1.0):1.0,AY786855:1.0,A
Y786862:1.0):1.0):1.0):0.0,((EF363125:2.0,EF363126:2.0):2.0,(FJ495937:2.0,FJ495939:2.0,FJ4
95940:2.0,FJ495941:2.0,FJ495942:2.0,FJ495943:2.0,((FJ495957:2.0,FJ495999:2.0):2.0,FJ49597
7:2.0,FJ495992:2.0,FJ495995:2.0,FJ495996:2.0,FJ495997:2.0,FJ495998:2.0):2.0,(FJ495958:2.0,
((FJ495961:2.0,FJ495963:2.0):2.0,FJ495962:2.0):2.0,FJ495976:2.0,FJ495991:2.0,FJ495993:2.0)
:2.0,FJ495973:2.0,FJ495974:2.0,FJ495975:2.0,FJ495978:2.0,FJ495979:2.0,FJ495980:2.0,FJ495
981:2.0,FJ495994:2.0):2.0):0.0,((EU807832:2.0,((EU807833:2.0,EU807834:2.0,EU807835:2.0,
EU807836:2.0,EU807838:2.0):2.0,EU807837:2.0):2.0):0.0,(DQ487190:3.0,DQ487191:3.0):3.0):
0.0,(FJ496000:2.0,FJ496001:2.0,FJ496002:2.0,FJ496003:2.0,FJ496004:2.0,FJ496005:2.0,(FJ496
006:2.0,FJ496058:2.0):2.0,FJ496007:2.0,FJ496024:2.0,((FJ496025:2.0,((FJ496033:2.0,(FJ49603
4:2.0,FJ496036:2.0):2.0,FJ496038:2.0,FJ496041:2.0,(FJ919955:2.0,FJ919956:2.0,(((FJ919957:2
.0,FJ919962:2.0):2.0,FJ919961:2.0):2.0,FJ919960:2.0):2.0,FJ919958:2.0,FJ919959:2.0):2.0):2.0,
FJ496035:2.0,FJ496037:2.0,FJ496040:2.0):2.0):2.0,FJ496027:2.0,FJ496039:2.0):2.0,FJ496026:2
.0,FJ496059:2.0,FJ496068:2.0,FJ496069:2.0,FJ496070:2.0,FJ496071:2.0,FJ496123:2.0,FJ49612
8:2.0,FJ496130:2.0,FJ496131:2.0,FJ496132:2.0,FJ496133:2.0,FJ496134:2.0,FJ496135:2.0,FJ49
6136:2.0):2.0,(((JF320363:2.0,JF320364:2.0,JF320366:2.0,JF320519:2.0):2.0,JF320365:2.0,JF3
20369:2.0,JF320373:2.0,JF320374:2.0,JF320514:2.0,JF320515:2.0,JF320516:2.0,JF320518:2.0):
2.0,(HM208363:3.0,(JF320381:3.0,JF320384:3.0):3.0):0.0):0.0,(((JF320559:2.0,JF320561:2.0,JF
320562:2.0,JF320563:2.0):2.0,(JF320186:3.0,JF320193:3.0,JF320201:3.0,JF320215:3.0):3.0):0.
0,(JF320028:3.0,JF320029:3.0,JF320031:3.0,JF320032:3.0):3.0):0.0,(((((JF320044:3.0,JF32004
7:3.0,JF320049:3.0,JF320051:3.0,JF320062:3.0):3.0,JF320064:3.0):3.0,((JF320053:3.0,JF32005
8:3.0,JF320060:3.0):3.0,JF320055:3.0):3.0):3.0,(JF320145:3.0,JF320147:3.0,JF320152:3.0,JF32
0153:3.0,JF320154:3.0):3.0):0.0,((JF320048:3.0,JF320056:3.0,JF320065:3.0,JF320068:3.0,JF32
0071:3.0):3.0,(JF320179:3.0,JF320181:3.0,JF320182:3.0):3.0):0.0,((JF320059:3.0,(JF320307:3.
0,JF320309:3.0,JF320311:3.0):3.0):0.0,((JF320197:3.0,JF320200:3.0,JF320202:3.0,JF320207:3.
0):3.0,JF320205:3.0):3.0):0.0):0.0,(JF320183:2.0,(JF320187:2.0,JF320209:2.0,JF320212:2.0):2.
0):2.0,((((JF320409:3.0,JF320569:3.0,JF320582:3.0):3.0,JF320591:3.0):3.0,(JF320571:3.0,JF32
0577:3.0):3.0,JF320572:3.0,JF320575:3.0):3.0,(JF320573:3.0,JF320574:3.0,JF320576:3.0):3.0):
3.0,((((JF320460:3.0,JF320464:3.0):3.0,((JF320461:3.0,JF320468:3.0):3.0,JF320462:3.0,JF3204
63:3.0,JF320465:3.0,JF320466:3.0):3.0,JF320467:3.0):3.0,JF320469:3.0):3.0,JF320470:3.0):3.0)
:0.0);

# Appendix C

Posterior mean differential selection factors for all amino acids at all sites for two independent runs, for within-patients (*a*), B57+ patients (*b*) and B35+ patients (*c*). The correlation coefficient $R^2$ is provided for each plot.



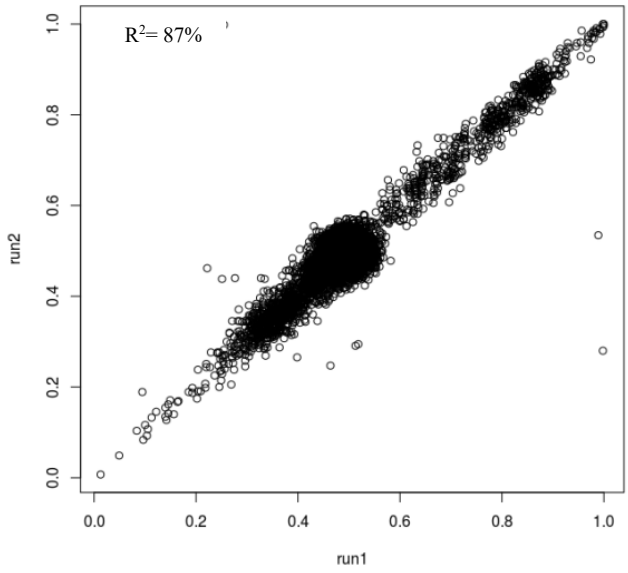




c
R2= 87%
run2
run1
-1.0
-0.5
0.0
0.5
1.0
1.5

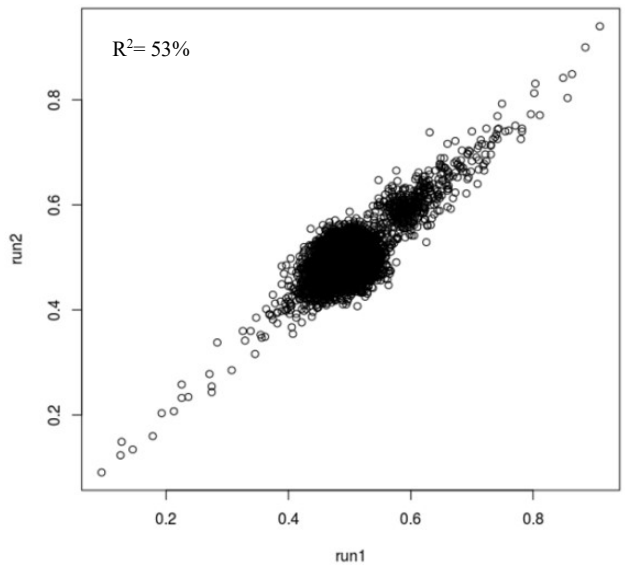
# Appendix D

Posterior probability correlation for all amino acids at all sites for two independent runs, for within-patients (a), B57+ patients (b) and B35+ patients (c). The correlation coefficient $R^2$ is provided for each plot.

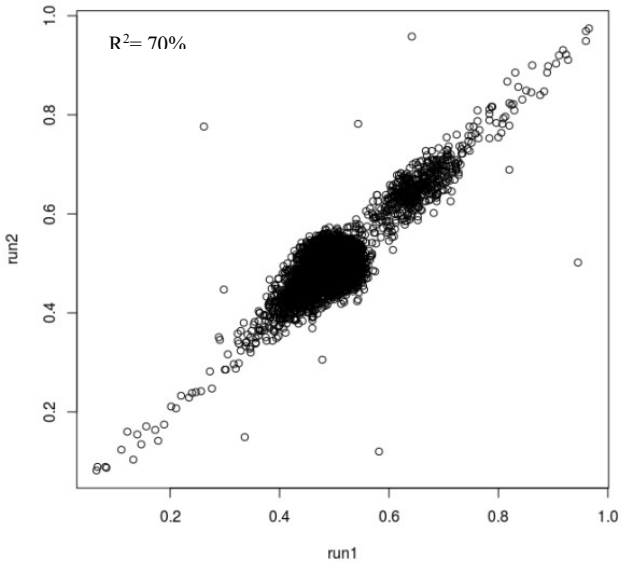
***a***




***b***



***c***



# Appendix E

## List of 179 Amaranthaceae species

The photosynthetic type (C3 or C4) of each species is given.

| Species | Type of Photosynthesis |
|---|---|
| Achyranthes_aspera | C3 |
| Acroglochin_chenopodioides | C3 |
| Aerva_javanica | C4 |
| Agriophyllum_squarrosum | C3 |
| Allenrolfea_occidentalis | C3 |
| Alternanthera_caracasana | C4 |
| Alternanthera_pungens | C4 |
| Alternanthera_repens | C4 |
| Amaranthus_blitum | C4 |
| Amaranthus_greggii | C4 |
| Amaranthus_hypochondriacus | C4 |
| Amaranthus_tricolor | C4 |
| Anabasis_aphylla | C4 |
| Anabasis_brevifolia | C4 |
| Anabasis_elatior | C4 |
| Anabasis_eriopoda | C4 |
| Anabasis_salsa | C4 |
| Anabasis_truncata | C4 |
| Anthochlamys_multinervis | C3 |
| Aphanisma_blitoides | C3 |

| | |
|---|---|
| Archiatriplex_nanpinensis | C3 |
| Arthrocnemum_macrostachyum | C3 |
| Atriplex_aucherii | C3 |
| Atriplex_australasica | C3 |
| Atriplex_centralasiatica | C4 |
| Atriplex_coriacea | C4 |
| Atriplex_glauca | C4 |
| Atriplex_halimus | C4 |
| Atriplex_lampa | C4 |
| Atriplex_lentiformis | C4 |
| Atriplex_parryi | C4 |
| Atriplex_patula | C3 |
| Atriplex_phyllostegia | C4 |
| Atriplex_powellii | C4 |
| Atriplex_rosea | C4 |
| Atriplex_serenana | C4 |
| Atriplex_spongiosa | C4 |
| Atriplex_undulata | C4 |
| Axyris_prostrata | C3 |
| Bassia_dasyphylla | C3 |
| Bassia_diffusa | C3 |
| Bassia_prostrata | C4 |
| Bassia_sedoides | C4 |
| Beta_nana | C3 |
| Beta_vulgaris | C3 |
| Bienertia_cycloptera | C4 |
| Blutaparon_vermiculare | C4 |

| | |
|---|---|
| Bosea_yervamora | C3 |
| Calicorema_capitata | C3 |
| Camphorosma_monspeliaca | C4 |
| Celosia_argentea | C3 |
| Celosia_trigyna | C3 |
| Ceratocarpus_arenarius | C3 |
| Chamissoa_altissima | C3 |
| Charpentiera_obovata | C3 |
| Charpentiera_ovata | C3 |
| Chenoleoides_tomentosa | C4 |
| Chenopodium_acuminatum | C3 |
| Chenopodium_album | C3 |
| Chenopodium_ambrosioides | C3 |
| Chenopodium_auricomum | C3 |
| Chenopodium_bonushenricus | C3 |
| Chenopodium_botrys | C3 |
| Chenopodium_coronopus | C3 |
| Chenopodium_cristatum | C3 |
| Chenopodium_desertorum | C3 |
| Chenopodium_foliosum | C3 |
| Chenopodium_frutescens | C3 |
| Chenopodium_murale | C3 |
| Chenopodium_sanctaeclarae | C3 |
| Climacoptera_brachiata | C4 |
| Climacoptera_lanata | C4 |
| Corispermum_filifolium | C3 |
| Cremnophyton_lanfrancoi | C3 |

| | |
|---|---|
| Cycloloma_atriplicifolium | C3 |
| Deeringia_amaranthoides | C3 |
| Dissocarpus_paradoxus | C3 |
| Dysphania_glomulifera | C3 |
| Einadia_nutans | C3 |
| Girgensohnia_oppositiflora | C4 |
| Gomphrena_elegans | C3 |
| Gomphrena_haageana | C4 |
| Gomphrena_serrata | C4 |
| Guilleminea_densa | C4 |
| Hablitzia_tamnoides | C3 |
| Halimione_pedunculata | C3 |
| Halimione_verrucifera | C3 |
| Halimocnemis_karelinii | C4 |
| Halimocnemis_villosa | C4 |
| Halocharis_hispida | C4 |
| Halogeton_arachnoideus | C4 |
| Halogeton_glomeratus | C4 |
| Halopeplis_amplexicaulis | C3 |
| Halosarcia_indica | C4 |
| Halostachys_belangeriana | C3 |
| Halothamnus_bottae | C4 |
| Haloxylon_ammodendron | C4 |
| Haloxylon_persicum | C4 |
| Haloxylon_tamariscifolium | C4 |
| Hebanthe_occidentalis | C3 |
| Hemichroa_diandra | C3 |

| | |
|---|---|
| Hermbstaedtia_glauca | C3 |
| Horaninovia_ulicina | C4 |
| Iljinia_regelii | C4 |
| Iresine_palmeri | C3 |
| Kalidium_caspicum | C3 |
| Kalidium_cuspidatum | C3 |
| Kalidium_foliatum | C3 |
| Kochia_americana | C3 |
| Kochia_densiflora | C4 |
| Krascheninnikovia_ceratoides | C3 |
| Maireana_brevifolia | C3 |
| Manochlamys_albicans | C3 |
| Microgynoecium_tibeticum | C3 |
| Micromonolepis_pusilla | C3 |
| Monolepis_nuttalliana | C3 |
| Nanophyton_erinaceum | C4 |
| Nitrophila_occidentalis | C3 |
| Noaea_mucronata | C4 |
| Nototrichium_humile | C3 |
| Ofaiston_monandrum | C4 |
| Oreobliton_thesioides | C3 |
| Pachycornia_triandra | C3 |
| Panderia_pilosa | C4 |
| Pandiaka_angustifolia | C3 |
| Patellifolia_patellaris | C3 |
| Petrosimonia_glaucescens | C4 |
| Petrosimonia_nigdeensis | C4 |

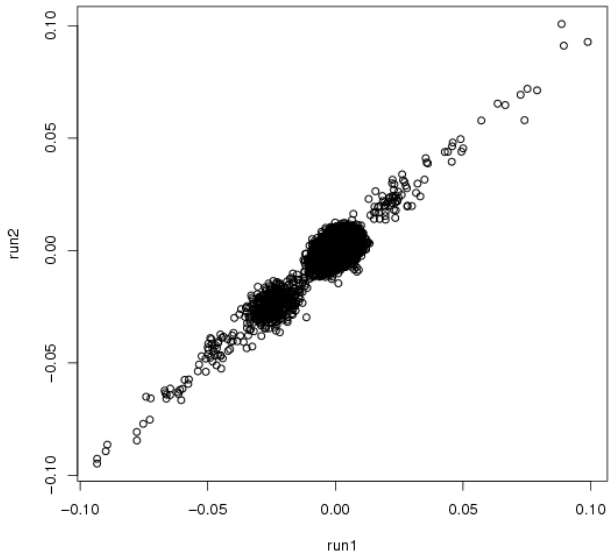
| | |
|---|---|
| Petrosimonia_sibirica | C4 |
| Petrosimonia_squarrosa | C4 |
| Polycnemum_perenne | C3 |
| Pseudoplantago_friesii | C3 |
| Ptilotus_manglesii | C3 |
| Pupalia_lappacea | C3 |
| Rhagodia_drummondi | C3 |
| Rhaphidophyton_regelii | C3 |
| Roycea_divaricata | C3 |
| Salicornia_dolichostachya | C3 |
| Salicornia_europaea | C3 |
| Salsola_affinis | C4 |
| Salsola_arbuscula | C4 |
| Salsola_arbusculiformis | C3 |
| Salsola_chinghaiensis | C4 |
| Salsola_collina | C4 |
| Salsola_dshungarica | C4 |
| Salsola_ferganica | C4 |
| Salsola_foliosa | C4 |
| Salsola_genistoides | C3 |
| Salsola_heptapotamica | C4 |
| Salsola_implicata | C4 |
| Salsola_kali | C4 |
| Salsola_komarovii | C4 |
| Salsola_laricifolia | C3 |
| Salsola_micranthera | C4 |
| Salsola_orientalis | C4 |

| | |
|---|---|
| Salsola_paulsenii | C4 |
| Salsola_pellucida | C4 |
| Salsola_praecox | C4 |
| Salsola_rosacea | C4 |
| Salsola_ruthenica | C4 |
| Salsola_sukaczevii | C4 |
| Salsola_vermiculata | C4 |
| Salsola_zaidamica | C4 |
| Sarcocornia_utahensis | C3 |
| Sclerolaena_obliquicuspis | C3 |
| Sclerostegia_moniliformis | C3 |
| Sericostachys_scandens | C3 |
| Spinacia_oleracea | C3 |
| Suaeda_altissima | C4 |
| Suaeda_crassifolia | C3 |
| Suaeda_maritima | C3 |
| Suaeda_microphylla | C4 |
| Suaeda_physophora | C3 |
| Suckleya_suckleyana | C3 |
| Sympegma_regelii | C3 |
| Tecticornia_australasica | C3 |
| Tecticornia_disarticulata | C3 |
| Teloxys_aristata | C3 |
| Tidestromia_lanuginosa | C4 |

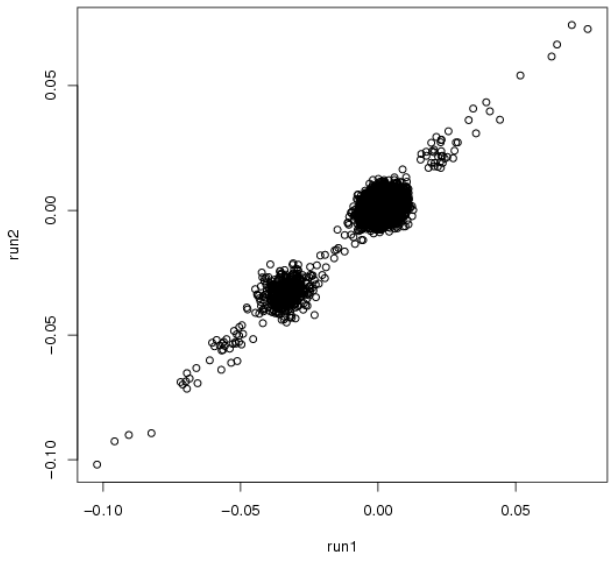
# Appendix F

Estimates of posterior mean differential selection effects across all amino acids and all sites for two independent chains, for C3 plants (*a*) and C4 plants (*b*). The correlation coefficient is 0.78 and 0.81, respectively.

***a***



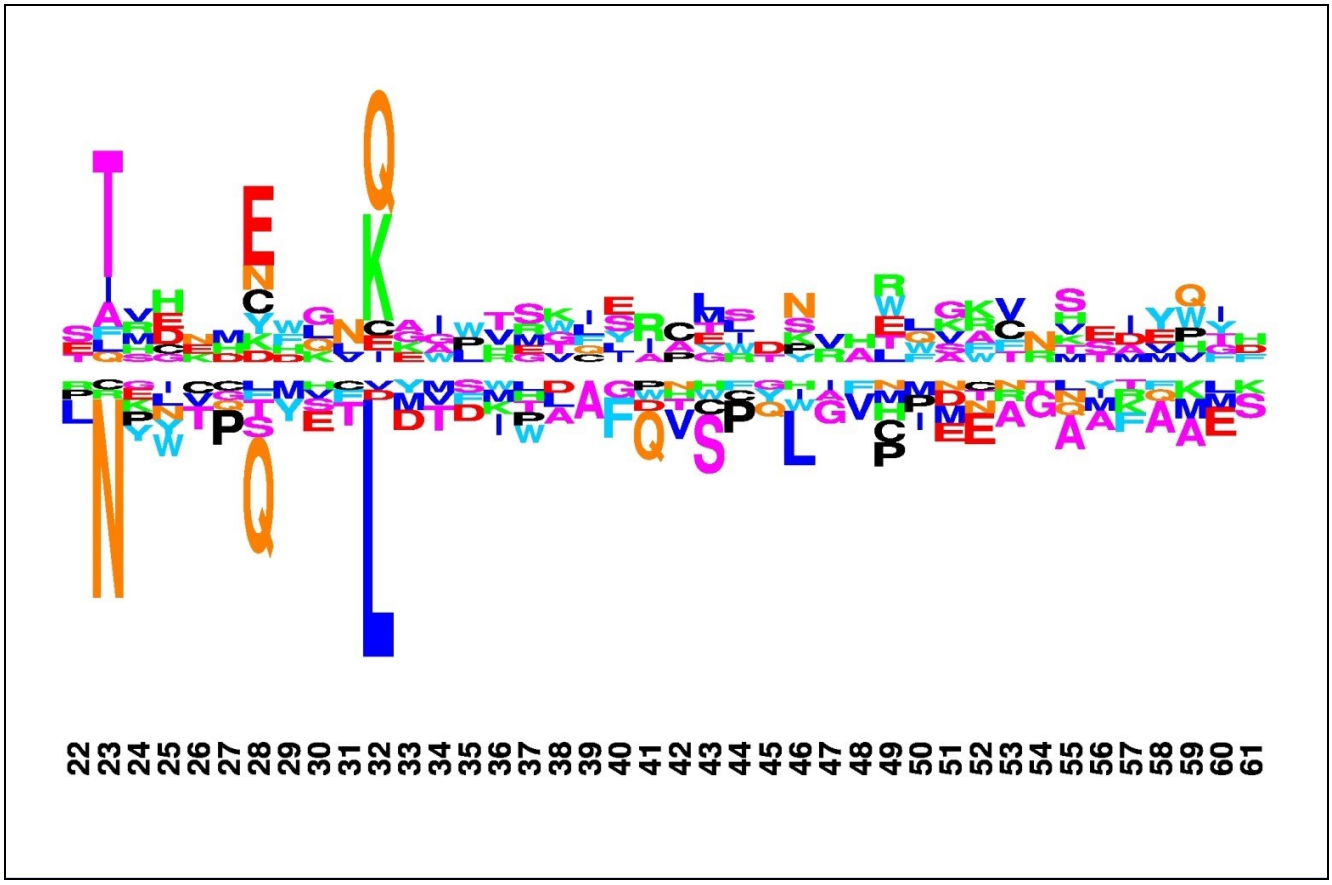

***b***

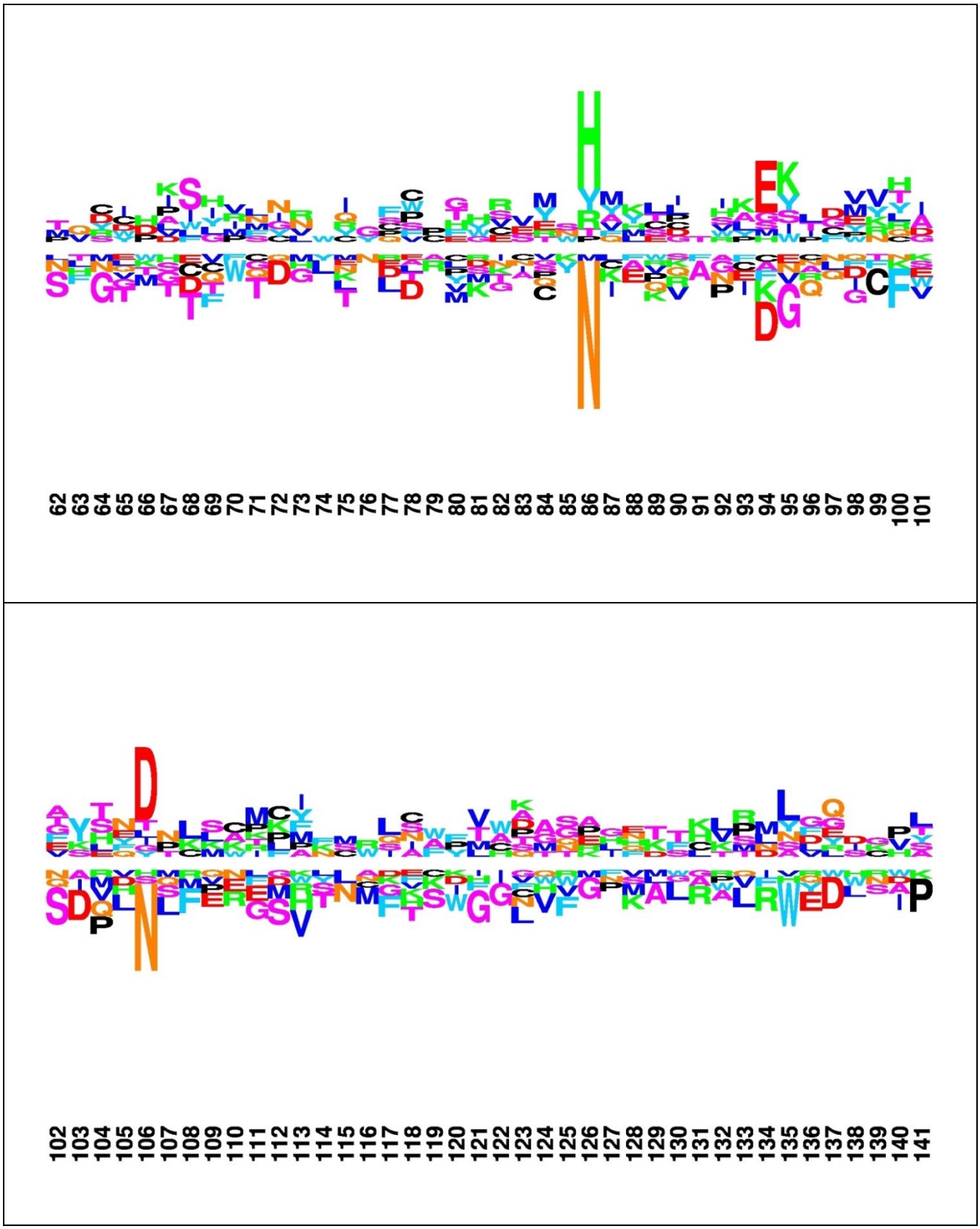


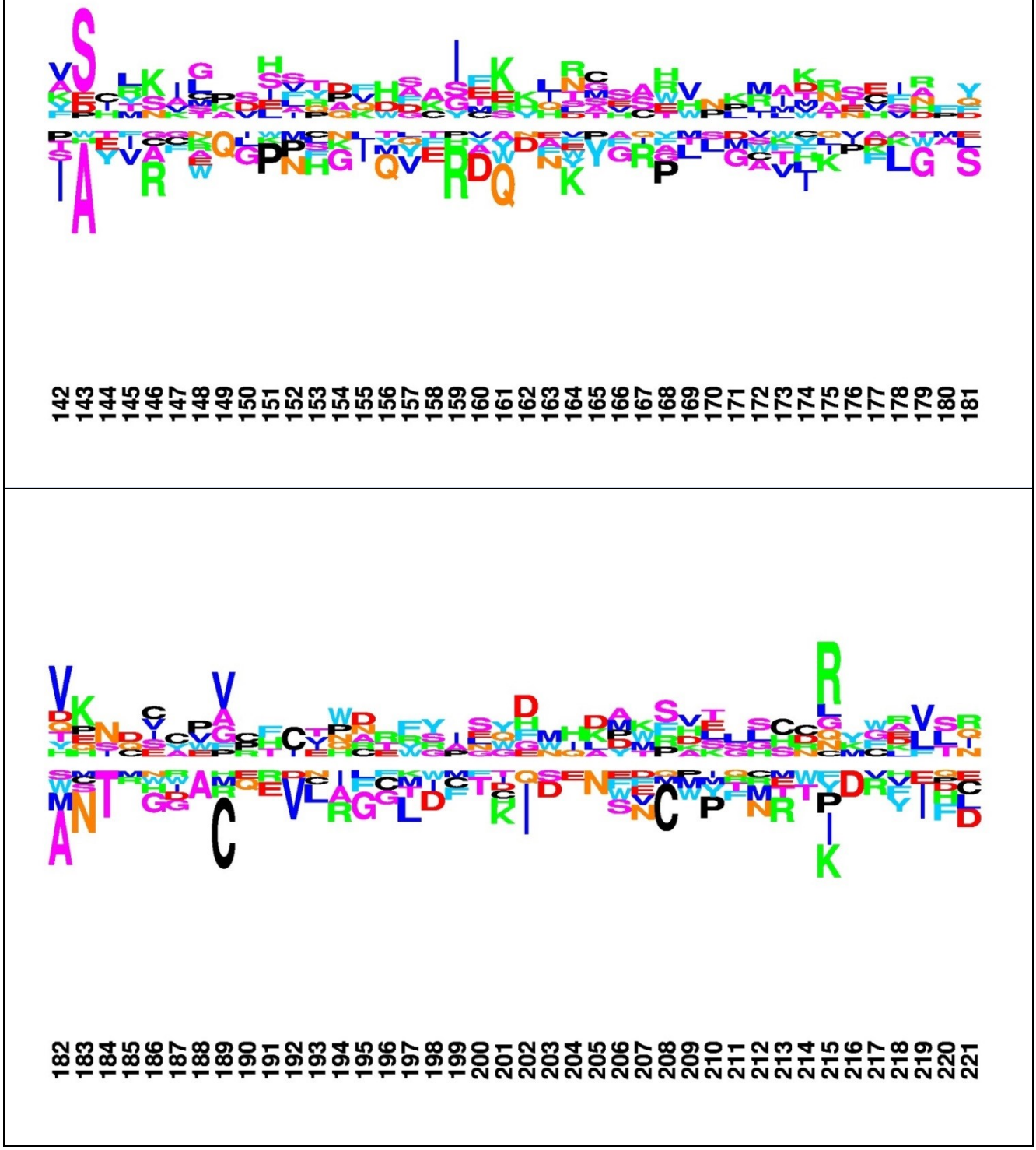
# Appendix G

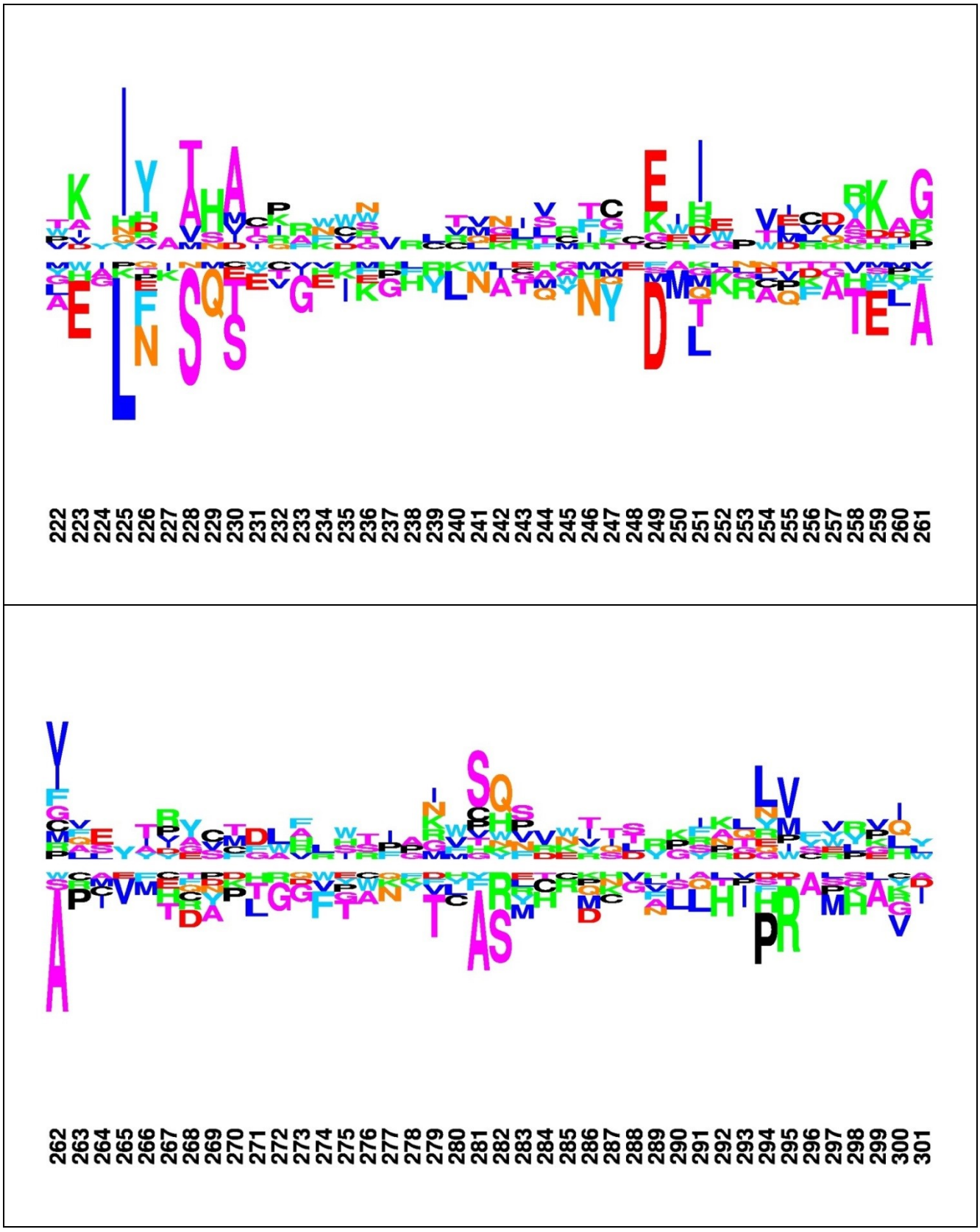
## C4/C3 Differential Selection profiles for rbcL sequence estimated by model DS3

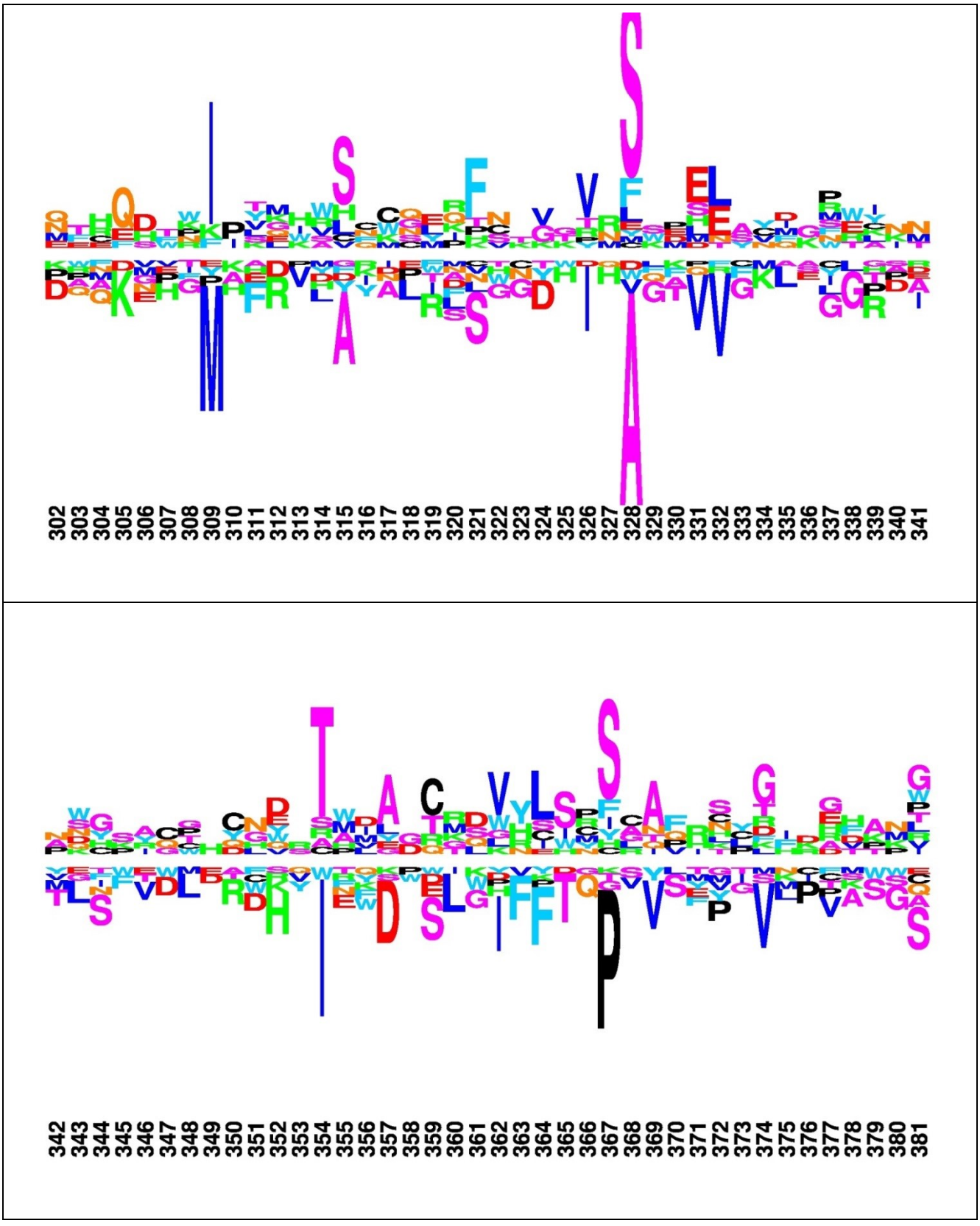
The first 21 amino acids are missing.

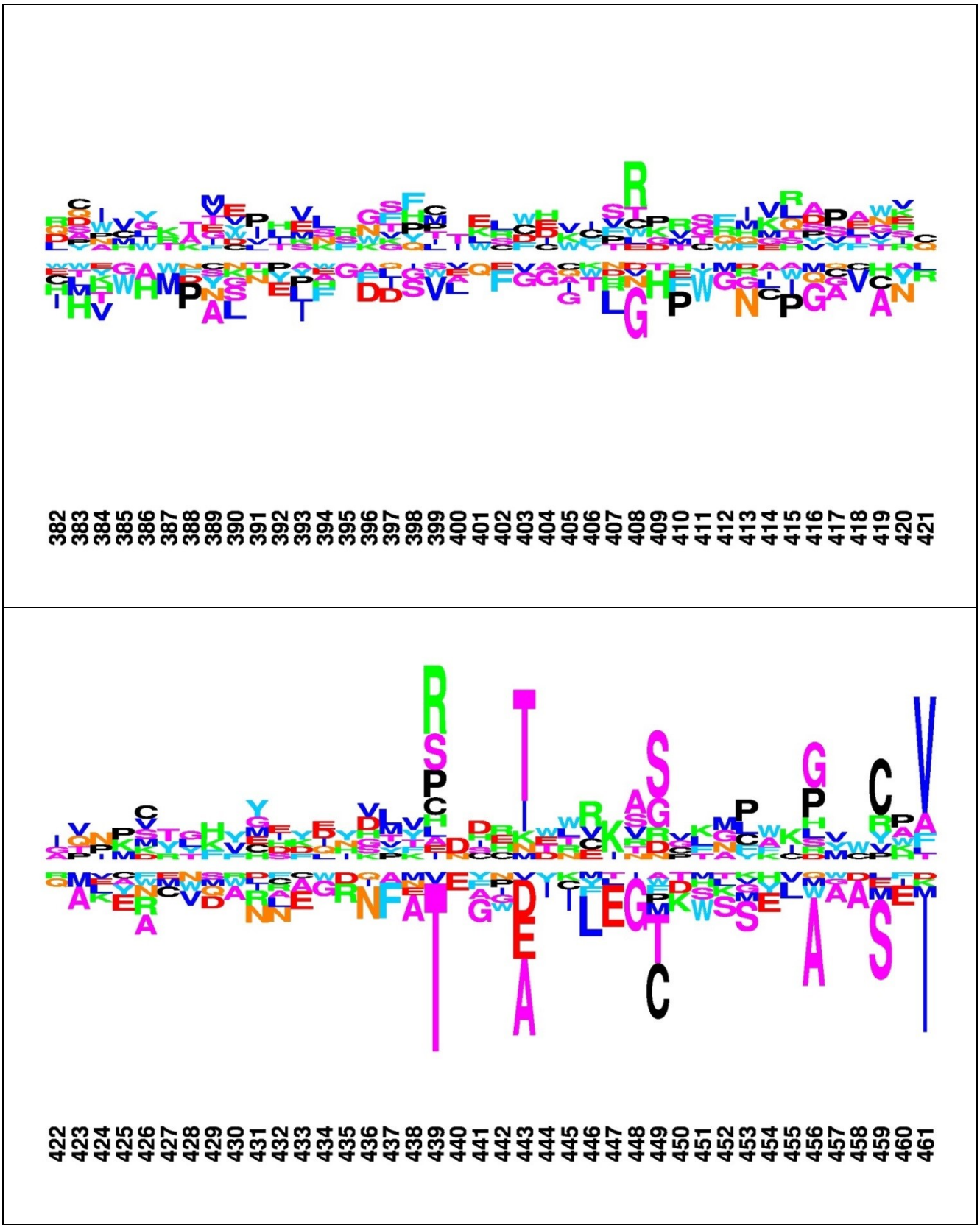



62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101
102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141


142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181
182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221


222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261
262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301


302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341
342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381


382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421
422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461

462
463
464
465
466
467
468