# Chapter 5

# Selection on the Protein-Coding Genome

Carolin Kosiol and Maria Anisimova

## Abstract

Populations evolve as mutations arise in individual organisms and, through hereditary transmission, may become “fixed” (shared by all individuals) in the population. Most mutations are lethal or have negative fitness consequences for the organism. Others have essentially no effect on organismal fitness and can become fixed through the neutral stochastic process known as random drift. However, mutations may also produce a selective advantage that boosts their chances of reaching fixation. Regions of genes where new mutations are beneficial, rather than neutral or deleterious, tend to evolve more rapidly due to positive selection. Genes involved in immunity and defense are a well-known example; rapid evolution in these genes presumably occurs because new mutations help organisms to prevail in evolutionary “arms races” with pathogens. In recent years, genome-wide scans for selection have enlarged our understanding of the evolution of the protein-coding regions of the various species. In this chapter, we focus on the methods to detect selection in protein-coding genes. In particular, we discuss probabilistic models and how they have changed with the advent of new genome-wide data now available.

**Key words:** Conserved and accelerated regions, Positive selection scans, Codon models, Time and space heterogeneity of genome evolution, Phylo-HMMs, Selection-mutation models

## 1. Introduction

Protein-coding genes are the DNA sequences used as templates for the production of a functional protein. Such sequences consist of nucleotide triplets called codons. During the protein production phase, codons are transcribed and then translated into amino acids (AAs) according to the organism’s genetic code. In the past, selection studies on coding DNA mainly focused on the analysis of particular proteins of interest. With the availability of comparative genomic data, the emphasis has shifted from the study of individual proteins to genome-wide scans for selection. The overview of genomic data underlying the genome-wide analysis of protein-coding genes is included in Subheading 2.

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods, Volume 2*,
Methods in Molecular Biology, vol. 856, DOI 10.1007/978-1-61779-585-5_5,

The analysis of coding sequences can be performed on three different levels: using DNA, AA, or codon sequences. The mutational processes at these three levels can be described by probabilistic models, which set the basis for evaluating selective pressures and selection tests. The fundamental properties of these models are summarized in Subheading 3.1.

There is accumulating evidence that the evolutionary process varies between sites in biological sequences. Even in nonfunctional genomic regions, there appears to be variation in the mutational process. This variation is even more pronounced in active genomic segments. In protein-coding sequences, changes that impede function are unlikely to be accepted by selection (e.g., mutation in active site) while those altering less vital areas are under lower selective constraints (e.g., mutation in nonfunctional loop regions). Furthermore, systematic studies have shown that variability is not determined exclusively by selection on protein structure and function, but is also affected by the genomic position of the encoding genes, their expression patterns, their position in biological networks and their robustness to mistranslation (see ref. 1 for a review of these factors).

In Fig. 1, we summarize the different levels of modeling selection on protein-coding sequences. The wedges represent the three data types: DNA, AA, and codons. Temporal heterogeneity is represented by the tree branches from lineage-specific models to analyses considering genealogies and population properties, such as the effective population size and the distribution of selective coefficients. For example, temporal heterogeneity is included in models that detect regions with accelerated regions in DNA, rate shifts in AA sequences, or the branch-specific codon models.

Furthermore, the concentric layers in Fig. 1 describe different levels of modeling spatial heterogeneity in cDNA, such as phylogenetic hidden Markov models (phylo-HMMs) for DNA or branch-site models for codon sequences. Within the "Methods"

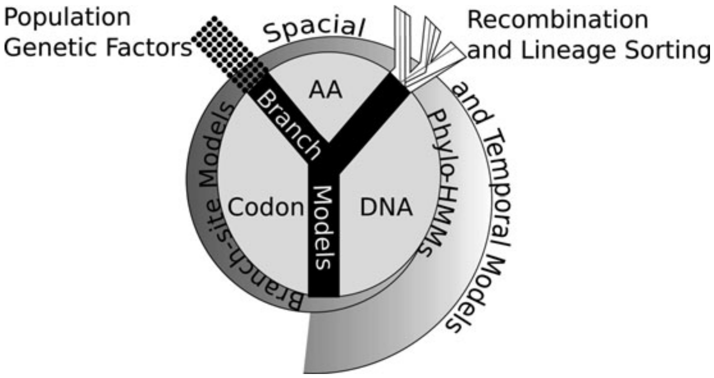

Fig. 1. A diagram illustrating the different data levels to analyze protein-coding sequences and the relationship of the various approaches modeling spatial and temporal heterogeneity.

section, Subheadings 3.2–3.4 are devoted to models allowing for temporal and spatial heterogeneity and give an overview of state-of-the-art methods to analyze selection of protein-coding regions. Subheadings 4.1–4.5 discuss possible sources of errors in genome-wide analyses. Finally, we conclude with the "Discussion" section providing insights to emerging directions in studying selection at the genomics level.

## 2. Comparative Genome Data

Several whole-genome sequence data sets are now available for selection scans. Mammalian genomes are well represented (in particular primates), and insect genomes are becoming more numerous (in particular *Drosophila*). These data can be downloaded as orthologous alignments from the Ensembl (2) and UCSC (3) browsers. Methods for constructing orthologous sets of genes are reviewed in Chapter 9 of Volume 1 (4).

In light of recent advances in DNA sequencing, with the so-called next-generation sequencing (NGS) technologies that have dramatically reduced the cost and time needed to sequence an organism's entire genome, large-scale (involving many organisms) sequencing projects have been and are currently being undertaken. In particular, genome projects resequencing 1000 *Human*, 1000 *Drosophila melanogaster*, and 1001 *Arabidopsis* individuals are ongoing. These polymorphism data from multiple individuals from several species enable us to detect very recent selection.

Together with the progress in sequencing technologies, algorithmic advances now allow the de novo assembly of genomes from NGS data (see Chapter 5 in Volume 1 (5)), including complex mammalian genomes (e.g., giant panda genome (6)). Announced shortly after the *Human* 1000 Genomes Project, the 1000 Plant Genomes Project is yet another, similar highly large-scale genomics endeavor to take advantage of the speed and efficiency of NGS. The Genome 10 K project aims to assemble a genomic zoo—a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately 1 for every vertebrate genus. All these genomes can be subject to scans for selection, for which we outline methods below.

## 3. Methods

### *3.1. Probabilistic Models for Genome Evolution*

The statistical modeling of the evolutionary process is of great importance when performing selection studies. When comparing reasonably divergent sequences, counting the raw sequence identity (percentage of sites with observed changes) underestimates the amount of evolution that has occurred because, by chance alone, some sites will have incurred multiple substitutions. In this chapter, we discuss maximum likelihood (ML) and Bayesian methods to detect selection based on probabilistic models of character evolution. Such substitution models provide more accurate evolutionary distance estimates by accounting for these unobserved changes and explicitly model the selection pressure on the protein-coding level.

One of the primary assumptions made in defining probabilistic substitution models is that future evolution is only dependent on its current state and not on previous (ancestral) states. Statistical processes with this lack of memory are called Markov processes. The assumption itself is reasonable because during the evolution mutation and natural selection can only act upon the molecules present in an organism and have no knowledge of what came previously. However, some large-scale mutational events, such as recombination (7), gene conversion (e.g., see refs. 8 and 9), or horizontal transfer (10), might not satisfy this "memoryless" condition.

To reduce the complexity of evolutionary models, it is often further assumed that each site in a sequence evolves independently from all other sites. There is evidence that the independence of sites assumption is violated. In real proteins, chemical interactions between neighboring sites or the protein structure affect how other sites in the sequence change. Steps have been made toward context-dependent models, where the specific characters at neighboring sites affect the sites' evolution (e.g., see refs. 11 and 12).

The Markov model asserts that one protein sequence is derived from another by a series of independent substitutions, each changing one character in the first sequence to another character in the second during the evolution. Thereby, we assume independence of evolution at different sites. A continuous-time Markov process is fully defined by its instantaneous rate matrix $Q = \{q_{ij}\}_{i,j=1\ldots N}$.

The diagonal elements of $Q$ are defined by a mathematical requirement that the rows sum up to zero. For multiple sequence alignments, the substitution process runs in continuous time over a tree representing phylogenetic relations between the sequences. The transition probability matrix $P(t) = \{p_{ij}(t)\} = e^{Qt}$ consists of transition probabilities from residue $i$ to residue $j$ over time $t$, and is found as a solution of the differential equation $\mathrm{d}P(t)/\mathrm{d}t = P(t)Q$ with $P(0)$ being the identity matrix. In order for tree branches to be measured by the expected number of substitutions per site, the matrix Q is scaled so that the average substitution rate at equilibrium equals 1.

As a matter of mathematical and computational convenience rather than biological reality, several simplifying assumptions are usually made. Standard substitution models allow any state to change into any other. Such Markov process is called *irreducible* and has a unique *stationary* distribution corresponding to the equilibrium codon frequencies $\pi = \{\pi_i\}$. *Time reversibility* implies that the direction of the change between two states, $i$ and $j$, is indistinguishable so that $\pi_i\, p_{ij}(t) = \pi_j\, p_{ji}(t)$. This assumption helps to reduce the number of model parameters and is convenient when calculating the matrix exponential (the matrix Q of a reversible process has only real eigenvectors and eigenvalues (13)). The fully unrestrained matrix Q for $N$ characters defines an irreversible model with $[N(N-1)-1]$ free parameters while for a reversible process this number is $[(N(N+1)/2)-2]$.

By comparing how well-substitution models explain sequence evolution and by examining the parameters estimated from data, ML and Bayesian inference can be used to address many biologically important questions. In this section, we focus on probabilistic models that are used to detect selection.

### 3.2. Detecting Regions of Accelerated Genome Evolution

Understanding the forces shaping the evolution of specific lineages is one of the most exciting areas in evolutionary genomics. In particular, regions of accelerated evolution in mammalian and insect species have been studied (e.g., see ref. 14). To eliminate nonfunctional regions, one strategy is to begin with a search for regions that are conserved through the mammalian history or longer. A likelihood ratio test (LRT) may be used to detect acceleration of rates in a lineage of interest, for example the human lineage. Such LRT compares the likelihood of the alignment data under two probabilistic models. The null model has a single-scale parameter representing shortening (more conserved) and lengthening (less conserved) of all branches of the tree. The alternative model has an additional parameter for the human lineage, which is constraint to be $\geq 1$. This extra parameter allows the human branch to be relatively longer (accelerated) than the branches in the rest of the tree.

For example, this approach was used to identify genomic regions that are conserved in most vertebrates, but have evolved rapidly in humans. Interestingly, the majority of the human accelerated regions (HARs) were noncoding and many were located near protein-coding genes with protein functions related to the nervous system (14).

In contrast, the majority of *D. melanogaster*-accelerated regions (DMARs) are found in protein-coding regions and primarily result from rapid adaptive change at synonymous sites (15). This could be because flies have much more compact genomes compared to humans; however, even after considering the genomic content, in *Drosophila*, a significant excess of DMARs occur in protein-coding regions. Furthermore, Holloway and colleagues observed a mutational bias from G|C to A|T, and therefore the accelerated divergence in DMARs might be attributed to a shift in codon usage and a fixation of many suboptimal codons.

In a similar manner, amino acid-based models search for site- or lineage-specific rate accelerations and residues subject to altered functional constraints. Such sites are likely to be contributing to the change in protein function over time. The advantage of amino acid-based models is that they might be suitable for the analysis of deep divergences of fast-evolving genes, where sequences rapidly saturate over time. Also amino acid methods are not influenced by the effects of codon bias, a topic that is discussed at the end of this chapter. The idea is that adaptive change on the level of amino acid sequences may not necessarily correspond to an adaptive change in protein function but rather to peaks in the protein-adaptive landscape reflecting the optimization of the protein function in a particular species to long-term environmental changes. One class of methods for detecting functional divergence searches for a lineage-specific change in the shape parameter of the gamma distribution that is used to model rate heterogeneity (see refs. 16–18 and 19). Other methods search for evidence of clade-specific rate shifts at individual sites (see refs. 20–25 and 26). For example, Gu (21) proposed a simple stochastic model for estimating the degree of divergence between two prespecified clusters. The statistical significance was tested using site-specific profiles based on an HMM, which was used to identify amino acids responsible for these functional differences between two gene clusters. More flexible evolutionary models were incorporated in the maximum likelihood approach applicable to the simultaneous analysis of several gene clusters (27). This was extended (28) to evaluate site-specific shifts in amino acid properties, in comparison with site-specific rate shifts. Pupko and Galtier (24) used the LRT to compare ML estimates of the replacement rate at an amino acid site in distinct subtrees.

### 3.3. Phylogenetic Hidden Markov Models

Phylo-HMMs are probabilistic models that consider not only the way substitutions occur along an evolutionary history represented by a tree, but also the way this process changes from site to site in a genome. Phylo-HMMs describe evolution as a combination of two Markov processes—one that operates in the dimension of space (along the genome) and one that operates in the dimension of time (along the branches of a phylogenetic tree). In the assumed process, a character is drawn at random from the background distribution and assigned to the root of the tree. Character substitutions occur randomly along the tree branches from root to leaves. The characters that are found at the leaves when the process has been completed define an alignment column having a correlation structure that reflects the phylogeny and the substitution process. The different phylogenetic models associated with the states of the phylo-HMM may reflect different overall rates of substitution (for example, conserved and nonconserved as in Fig. 2) and different patterns of substitution or background distributions (as in different codon positions). The idea is to identify

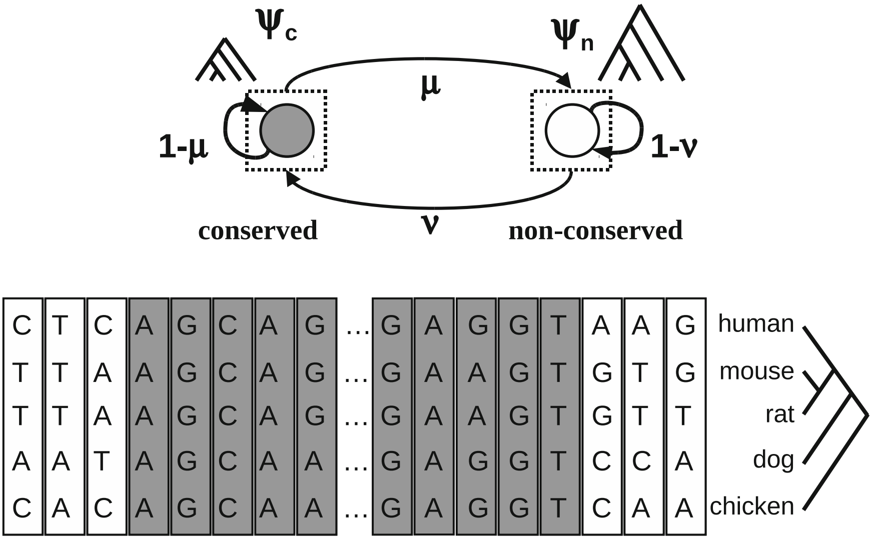
$\Psi_c$
$\Psi_n$
$\mu$
$1-\mu$
$1-\nu$
$\nu$
conserved
non-conserved
C T C A G C A G … G A G G T A A G human
T T A A G C A G … G A A G T G T G mouse
T T A A G C A G … G A A G T G T T rat
A A T A G C A A … G A G G T C C A dog
C A C A G C A A … G A G G T C A A chicken

Fig. 2. Visualization of an example phylo-HMM showing the probabilistic graph and the input alignment. The *grey* columns represent the conserved state; the *white* columns the fast state. At each time step, a new state is visited according to the transition probabilities ($\mu$ and $\nu$ parameters on arcs) and a multiple alignment column is emitted according to the conserved and nonconserved phylogenetic models $\Psi_c$ and $\Psi_n$. Thereby, the phylogenetic models include the parameters describing the tree and the pattern of substitution.

highly conserved genomic regions indicating purifying selection or accelerated regions indicating positive selection in a set of multiple aligned sequences. Such regions are good candidates for further selection analysis and they are likely to be functionally important. Hence, the identification of regions through phylo-HMMs has become a subject of considerable interest in comparative genomics (see refs. 29 and 30).

### *3.4. Codon Models: Site, Branch, and Branch-Site Specificity*

#### *3.4.1. Basic Codon Models*

In protein-coding sequences, nucleotide sites at different codon positions usually evolve with highly heterogeneous patterns (e.g., see ref. 31). Thus, DNA substitution models fail to account for this heterogeneity unless the sequences are partitioned by codon positions for the analysis. But even then, DNA models do not model the structure of genetic code or selection at the protein level. Indeed, one advantage of studying protein-coding sequences at the codon level is the ability to distinguish between nonsynonymous (AA replacing) and synonymous (silent) codon changes. Based on this distinction, the selective pressure on the protein-coding level can be measured by the ratio $\omega = d_N/d_S$ of the nonsynonymous-to-synonymous substitution rates. The nonsynonymous substitution rate may be higher than the synonymous rate and thus $\omega > 1$ due to fitness advantages associated with recurrent AA changes in the protein, i.e., positive selection on the protein. In contrast, purifying selection acts to preserve the protein sequence so that the nonsynonymous substitution rate is lower than the synonymous rate, with $\omega < 1$. Neutrally evolving sequences exhibit similar nonsynonymous and synonymous rates, with $\omega \approx 1$.

First methods that used the $\omega$-ratio as a criterion to detect positive selection were based on pairwise estimation of $d_N$ and $d_S$ rates with "counting" methods (e.g., see ref. 32). However, ML estimates of pairwise $d_N$ and $d_S$ based on a codon model were shown to outperform all other approaches (33). Moreover, a Markov codon model is naturally extended to multiple sequence alignments, unlike the counting methods. This, together with the benefits of the probabilistic framework within which codon models are defined, made codon models very popular in studies of positive selection in protein-coding genes.

The first two codon models were proposed simultaneously in the same issue of Molecular Biology and Evolution ((34) and (35)). The model of Goldman and Yang (34) included the transition/transversion rate ratio $\kappa$, and modeled the selective effect indirectly using a multiplicative factor based on Grantham (36) distances, but was later simplified to estimate the selective pressure explicitly using the $\omega$ parameter (37). The main distinction between the first codon models concerns the way to describe the instantaneous rates with respect to equilibrium frequencies: (1) proportional to the equilibrium frequency of a target codon (as in Goldman and Yang (34)) or (2) proportional to the frequency of a target nucleotide (as in Muse and Gaut (35)).

Recently, empirical codon models have been estimated (see refs. 38 and 39) that summarize substitution patterns from large quantities of protein-coding gene families. In contrast to the parametric codon models that estimate gene-specific parameters (e.g., transition–transversion $\kappa$, selective pressure $\omega$, etc.), the empirical codon models do not explicitly consider distinct factors that shape protein evolution. Standard parametric models assume that protein evolution proceeds only by successive single-nucleotide substitutions. However, empirical codon models indicate that model accuracy is significantly improved by incorporating instantaneous doublet and triplet changes. Kosiol et al. (39) also found that the affiliations among codon, the amino acid it encodes, and the physicochemical properties of the amino acid are main driving factors of the process of codon evolution. Neither multiple nucleotide changes nor the strong influence of the genetic code nor amino acid properties form a part of the standard parametric models.

On the other hand, parametric models have been very successful in applications studying biological forces shaping protein evolution of individual genes. Thus, combining the advantages of parametric and empirical approaches offers a promising direction. Kosiol, Holmes, and Goldman (39) explored a number of combined codon models that incorporated empirical AA exchangeabilities from ECM while using parameters to study selective pressure, transition/transversion biases, and codon frequencies. Similarly, AA exchangeabilities from (suitable) empirical AA matrices may be used to alter probabilities of nonsynonymous changes, together with traditional parameters $\omega$, $\kappa$, and codon frequencies $\pi_j$ (40). Such an approach accommodates site-specific variation of selective pressure and can be further extended to include lineage-specific variation. Combined empirical and parametric models will, therefore, become more frequent in selection studies. However, selecting an appropriate model is of utmost importance and needs further study. In particular, parameter interpretations may change with different model definitions, since empirical exchangeabilities already include average selective factors and other biases (39). Thus, selection among alternative parameterizations requires detailed attention.

#### 3.4.2. Accounting for Variability of Selective Pressures

First codon models assumed constant nonsynonymous and synonymous rates among sites and over time. Although most proteins evolve under purifying selection most of the time, positive selection may drive the evolution in some lineages. During episodes of adaptive evolution, only a small fraction of sites in the protein have the capacity to increase the fitness of the protein via AA replacements. Thus, approaches assuming constant selective pressure over time and over sites lack power in detecting genes affected by positive selection. Consequently, various scenarios of variation in selective pressure were incorporated in codon models, making them more powerful at detecting positive selection, and short episodes of adaptive evolution in particular. Evidence of positive selection on a gene can be obtained by an LRT comparing two nested models: a model that does not allow positive selection (constraining $\omega \leq 1$ to represent the null hypothesis) and a model that allows positive selection ($\omega > 1$ is allowed in the alternative hypothesis). Positive selection is detected if a model $\omega > 1$ fits data significantly better compared to the model restricting $\omega \leq 1$ at all sites and lineages. However, the asymptotic null distribution may vary from the standard $\chi^2$ due to boundary problems or if some parameters become not estimable (e.g., see refs. 41 and 42).

#### 3.4.3. Case Study: Application of a Genome-Wide Scan of Positive Selection on Six Mammalian Genomes

In 2006, six high-coverage genome assemblies became available for eutherian mammals. The increased phylogenetic depth of this data set permitted Kosiol and colleagues (43) to perform several new lineage- and clade-specific tests using branch-site codon models. Of ~16,500 human genes with high-confidence orthologs in at least two other species, 544 genes showed significant evidence of positive selection using branch-site codon models and standard LRTs.

Interestingly, several pathways were found to be strongly enriched in genes with positive selection, suggesting possible coevolution of interacting genes. A striking example is the complement immunity system, a biochemical cascade responsible for the elimination of pathogens. This system consists of several small proteins found in the blood that cooperate to kill target cells by disrupting their plasma membranes. Of 28 genes associated with this pathway in KEGG (see http://www.genome.jp/kegg-bin/show_pathway?map04610 for the complement cascades), 9 were under positive selection (FDR $< 0.05$) and 5 others had nominal $P < 0.05$. Most of the genes under positive selection are inhibitors (DAF, CFH, CFI) and receptors (C5AR1, CR2), but some are part of the membrane attack complex (C7, C9, C8B), which punctures cell membranes to initiate cell lysis. Here, we focus on the analysis of these proteins of the membrane attack complex.

First, we calculate gene-averaged $\omega$ value using the basic M0 model (34). The ML estimates of $\omega < 1$ ($\omega = 0.31$ for C7, $\omega = 0.25$ for C8B, and $\omega = 0.44$ for C9) indicate that most sites in these genes are under purifying selection. However, selection pressure could be variable at different locations of the membrane proteins and we, therefore, continue our analysis by applying models that allow for variation in selective pressure across sites.

#### 3.4.4. Selective Variability Among Codons: Site Models

The simplest site models use the general discrete distribution with a prespecified number of site classes. Each site class $i$ has an independent parameter $\omega_i$ estimated by ML together with proportions of sites $p_i$ in each class. Since a large number of site categories require many parameters, three categories are usually used (requiring five independent parameters). To test for positive selection, several pairs of nested site models were defined to represent the null and alternative hypotheses in LRTs. For example, model M1a includes two site classes, one with $\omega_0 < 1$ and another with $\omega_1 = 1$, representing the neutral model of evolution (the null hypothesis). The alternative model M2a extends M1a by adding an extra site class with $\omega_2 \geq 1$ to accommodate sites evolving under positive selection. Significance of the LRT is tested using the $\chi^2_2$ distribution for the M1 vs. M2 comparison. We test the C7 gene for positive selection by the LRT comparing nested models M1a and M2a (Table 1).

Model M2a has two additional parameters compared to model M1a. The resulting LRT statistic is $2\times$ (log L2 $-$ log L1) $= 2\times$ $(-6377.35 - (-6369.67)) = 2 \times 7.68 = 15.36$. This is much greater than the critical value of the chi-square distribution $\chi^2$ (d$f = 2$, at 5%) $= 5.99$, and we calculate a $p$-value of $P = 5.0\text{e–}04$. However, the M1a vs. M2a comparison for genes C8B and C9 is not significant.

Another LRT can be performed on the basis of the modified model M8 with two site classes: one with sites, where the $\omega$-ratio is drawn from the beta distribution (with $0 \leq \omega \leq 1$ describing the neutral scenario), and the second, discrete class, with $\omega \geq 1$. Constraining $\omega = 1$ for this second class provides a sufficiently

### Table 1
### Parameter estimates and log likelihoods for an LRT of positive selection for the complement immunity component C7

| | | | |
|---|---|---|---|
| *M1a (nearly neutral)* | | | |
| Site class | 0 | 1 | |
| Proportion | $p_0 = 0.69$ | $(p_1 = 1 - p_0 = 0.31)$ | |
| $\omega$ ratio | $\omega_0 = 0.07$ | $(\omega_1 = 1)$ | |
| Log likelihood L1 $= -6377.35$ | | | |
| *M2a (selection)* | | | |
| Site class | 0 | 1 | 2 |
| Proportion | $p_0 = 0.70$ | $p_1 = 0.29$ | $(p_2 = 1 - p_0 - p_1 = 0.01)$ |
| $\omega$ ratio | $\omega_0 = 0.08$ | $(\omega_1 = 1)$ | $\omega_2 = 10.89$ |
| Log likelihood L2 $= -6369.67$ | | | |

The model M2a is the alternative model with a class of sites with $\omega_2 \geq 1$. The null hypothesis M1a is the same model but with $\omega_2 = 1$ fixed

flexible null hypothesis, whereby all evolution can be explained by sites with $\omega$ from the beta distribution or from a discrete site class with $\omega = 1$. Significance of the LRT is tested using the mixture $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ for the M8 ($\omega = 1$) vs. M8 comparison. If the LRT for positive selection is found to be significant, specific sites under positive selection may be predicted based on the values of posterior probabilities (PPs) to belong to the site class under positive selection (usually, PP $> 0.95$, but see refs. 44 and 45). Such posterior probabilities are estimated using the naïve empirical Bayesian (NEB) approach (46), full hierarchical Bayesian approach (47), or a mid-way approach — the Bayes empirical Bayes (BEB (45)). For a discussion on this approaches, see Scheffler and Seoighe (48) and Aris-Brosou (49). Alternatively, Massingham and Goldman (50) proposed a site-wise likelihood ratio estimation to detect sites under purifying or positive selection.

For the C7 gene, using BEB, we identified several amino acid sites to be putatively under selection: residue R at position 223 (PP $= 0.94$), H at position 239 (PP $= 0.93$), and N at position 331 (PP $= 0.93$). Unfortunately, the crystal structures of C7 (as well as C8B and C9) are not known, and we cannot relate the location of amino acids in the protein sequence to relevant 3D data, such as sites of protein–protein interaction or binding sites of the protein. If such structural information were known, it would also be possible to use this biological knowledge in a model that is aware of the position of the different structural elements.

Site models that do not use a priori partitioning of codons (as those described above) are known as random-effect (RE) models. In contrast, fixed-effect (FE) models categorize sites based on a prior knowledge, e.g., according to tertiary structure for single genes, or by gene category for multigene data. Site partitions for FE models can be defined also based on inferred recombination breakpoints, useful for inferences of positive selection from recombining sequences (see refs. 51 and 52), although the uncertainty of breakpoint inference is ignored in this way. FE models with each site being a partition should be avoided, as they lead to the "infinitely many parameter trap" (e.g., see ref. 53). Given a biologically meaningful a priori partitioning, FE models are useful to study heterogeneity among partitions. However, a priori information is not always available.

### *3.4.5. Selective Variability Over Time: Branch Models*

A simple way to include the variation of the selective pressure over time is by using separate parameters $\omega$ for each branch of a phylogeny (known as *free-ratio* model (37)). Compared with the *one-ratio* model (which assumes constant selection over time), the free-ratio model requires additional $2T - 4$ $\omega$-parameters for $T$ species. Figure 3 shows the estimates of the free-ratio model for the C8B gene. Although the ML estimates of ω values on the rodent lineages are visibly higher than on the primate lineages, none of the branches has $\omega > 1$.

Other branch models can be defined by constraining different sets of branches of a tree to have an individual $\omega$. LRTs are used to

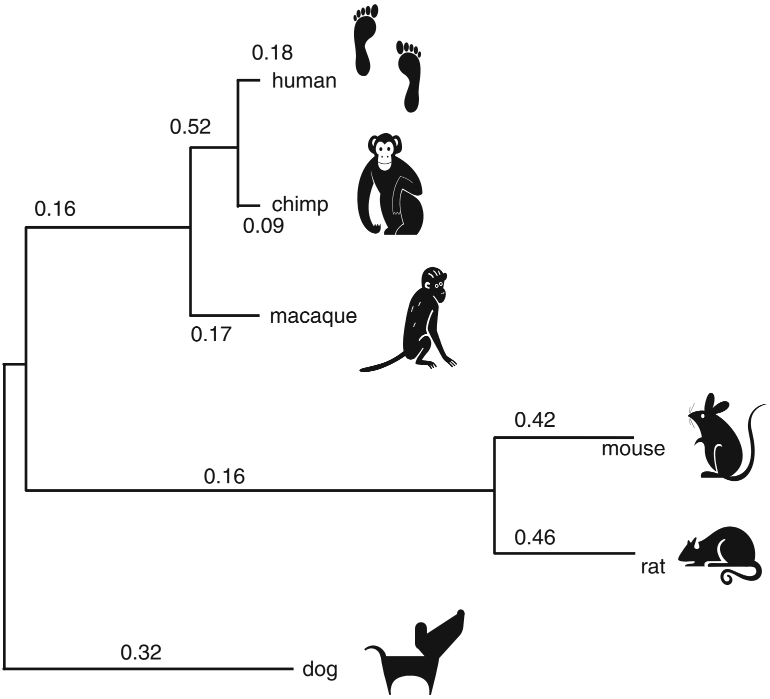

Fig. 3. An estimate of $\omega$ for each branch of a six-species phylogeny. Shown is the maximum likelihood estimate for the gene C8B. Each branch is labeled with the corresponding estimate of $\omega$.

decide (1) whether selective pressure is significantly different on a prespecified set of branches and (2) whether these branches are under positive selection.

However, branch models have relatively poor power to detect selection (54) in comparison to branch-site models that are discussed in the next section. Also note that testing of multiple hypotheses on the same data requires a correction, so the overall false-positive rate is kept at the required level (most often 5%). Correction for multiple testing further reduces the power of the method, especially when many hypotheses are tested simultaneously (see discussion later).

### 3.4.6. Temporal and Spatial Variation of Selective Pressure

Several solutions were proposed to simultaneously account for differences in selective constraints among codons and the episodic nature of molecular evolution at individual sites. One of the first models—model MA (45)—assumes four site classes. Two classes contain sites evolving constantly over time: one under purifying selection with $\omega_0 < 1$ and another with $\omega_1 = 1$. The other two site classes allow selective pressure at a site to change over time on a prespecified set of branches, known as *the foreground*. The two variable classes are derived from the constant classes so that sites typically evolving with $\omega_0 < 1$ or $\omega_1 = 1$ are allowed to be under positive selection with $\omega_2 \geq 1$ on the foreground. Testing for positive selection on the rodent clade involves an LRT comparing a constrained version of MA (with $\omega_2 = 1$) vs. an unconstrained MA model. Compared to branch models, the branch-site formulation improves the chance of detecting short spills of adaptive pressure in the past even if these occurred at a small fraction of sites.

Returning to our example of gene C8B of the complement pathway, we perform a branch-site LRT for positive selection using the M1a vs M2a comparison. Thereby, we take mouse and the rat lineage, respectively, as foreground branches, and all other branches as background branches. Significance of the LRT is tested using the mixture $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ with critical values to be 2.71 at 5%. For the C8B gene, we calculate $2\times$ (log L2 $-$ log L1) $= 2 \times 2.23 = 4.46$ for the mouse lineage and 11.2 for the rat lineage.

A major drawback of described branch-site models is their reliance on a biologically viable a priori hypothesis. In the context of detecting sites and lineages affected by positive selection, one possible solution is to perform multiple branch-site LRTs, each setting a different branch at the foreground (55). In the example of six species (Fig. 3), a total of nine tests (for an unrooted tree) are necessary in the absence of an a priori hypothesis. Multiple test correction has to be applied to control excessive false inferences. This strategy tends to be conservative but can be sufficiently powerful in detecting episodic instances of adaptation. As with all model-based techniques, precautions are necessary for data with unusual heterogeneity patterns, which may cause deviations from the asymptotic null distribution and thus result in an elevated false-positive rate.

In the case of episodic selection where any combination of branches of a phylogeny can be affected, a Bayesian approach in lieu of the standard LRTs and multiple testing have been suggested. The multiple LRT approach is most concerned with controlling the false-positive rate of selection inference, and is less suited to infer the best-fitting selection history. In the hypothetical example (Fig. 3), a total of $2^9 - 1 = 511$ selection histories (excluding the history without selection on any branch) need to be considered. The Bayesian analysis allows a probability distribution over possible selection histories to be computed, and therefore permits estimates of prevalence of positive selection on individual branches and clades. Such approach evaluates uncertainty in selection histories using their posterior probabilities and allows robust inference of interesting parameters, such as the switching probabilities for gains and losses of positive selection (43).

Other models (e.g., with $d_S$-variation among sites (56)) also may be extended to allow changes of selective regimes on different branches. This is achieved by adding further parameters, one per branch, describing the deviation of selective pressure on a branch from the average level on the whole tree under the site model. Such model is parameter rich and can be used for exploratory purposes on data with long sequences, but does not provide a robust way of testing whether $\omega > 1$ on a branch is due to positive selection on a lineage or due to inaccuracy of the ML estimation.

Kosakovsky Pond and Frost (56) suggested detecting lineage-specific variation in selective pressure using the genetic algorithm (GA)—a computational analogue of evolution by natural selection. The GA approach was successfully applied to phylogenetic reconstruction (see refs. 57, 58, and 59). In the context of detecting lineage-specific positive selection, GA does not require an a priori hypothesis. Instead, the algorithm samples regions of the whole hypotheses space according to their "fitness" measured by $\mathrm{AIC_C}$. The branch-model selection with GA may also be adapted to incorporate $d_N$ and $d_S$ among-site variation, although this imposes a much heavier computational burden.

In branch and branch-site models, change in selection regime is always associated with nodes of a tree, but the selective pressure remains constant over the length of each branch. Guindon et al. (60) proposed a Markov-modulated model, where switches of selection regimes may occur at any site and any time on the phylogeny. In a covarion-like manner, this codon model combines two Markov processes: one governs the codon substitution while the other specifies rates of switches between selective regimes. These models can be used to study the patterns of the changes in selective pressures over time and across sites by estimating the relative rates of changes between different selective regimes (purifying, neutral, and positive).

### 3.5. Software

The software PHylogenetic Analysis with Space/Time (PHAST) models includes several phylo-HMM-based programs. Two programs in PHAST are particularly interesting in the context of selection studies: PhastCons is a program for conservation scoring and identification of conserved elements (61). PhyloP is designed to compute *p*-values for conservation or acceleration, either lineage specific or across all branches (62). PHAST is designed for use on DNA sequences only.

A variety of codon models to detect selection, including branch-site models and the recent selection-mutation model, are implemented in the CODEML program of PAML (63). HYPHY is another implementation that includes a large variety of codon models (64). FitModel is the ML implementation of the switching codon model (60). Selecton Web server (65) offers several site models as well as the combined model described in Doron-Faigenboim and Pupko (40).

Xrate (66) is a generic tool to implement complex probabilistic models in the form of context-free stochastic grammars. Grammars for codon models can be defined such that they lead to estimates consistent with those at PAML, but for features of particular proteins (e.g., see analysis of transmembrane proteins (67)). However, Xrate is slower than PAML.

## 4. Notes/Discussion

With the wider use of codon models to detect selection, some questioned the statistical basis of testing based on branch-site models. In 2004, Zhang found that the original branch-site *test* (68) produced excessive false positives when its assumptions were not met. The modified branch-site test was shown to be more robust to model violations (see refs. 45 and 69), and is now commonly used in genome-wide selection scans (e.g., see ref. 70). Recently, however, another simulation study by Nozawa et al. (71) suggested that this modification also showed an excess of false positives. Yang and Dos Reis (54) defended the branch-site test by examining the null distribution and showing that Nozawa and colleagues (71) misinterpreted their simulation results. However, it is clear that even tests with good statistical properties are affected by data quality and the extent of models' violations. Below, we list factors that can affect the test, and so should be taken into account when analyzing genome-wide data.

### 4.1. Quality of Multiple Alignments

The impact of the quality of sequence and the alignment is a major concern when performing positive selection scans. For example, in their analysis of 12 genomes, Markova-Raina and Petrov (72) found that the results were highly sensitive to the choice of an alignment method. Furthermore, visual analysis indicated that most sites inferred as positively selected are in fact misaligned at the codon level. The rate of false positives ranged ~50% and more depending on the aligner used. Some of these results can be ascribed to the high divergence level of the 12 Drosophila species, and could be addressed by better filtering of the data. Nevertheless, even in mammals where alignment is easier, problems have been observed.

Bakewell et al. (73) used the branch-site test to analyze ~14,000 genes from the human, chimpanzee, and macaque, and detected more genes to be under positive selection on the chimpanzee lineage than on the human lineage (233 vs. 154). The same pattern was also observed by Arbiza et al. (74) and Gibbs et al. (75). Mallick et al. (76) reexamined 59 genes detected to be under positive selection on the chimpanzee lineage by Bakewell et al. (73), using more stringent filters to remove less reliable nucleotides and using synteny information to remove misassembled and misaligned regions. They found that with improved data quality, the signal of positive selection disappeared in most of the cases when the branch-site test was applied. It now appears that, as suggested by Mallick et al. (76), the earlier discovery of more frequent positive selection on the chimpanzee lineage than on the human lineage is an artifact of the poorer quality of the chimpanzee genomic sequence. This interpretation is also consistent with a few recent studies analyzing both real and simulated data, which suggest that sequence and alignment errors may cause excessive false positives (see refs. 77 and 78). Indeed, most commonly used alignment programs tend to place nonhomologous codons or amino acids into the same column (see refs. 79 and 80), generating the wrong impression that multiple nonsynonymous substitutions occurred at the same site and misleading the codon models into detecting positive selection (78).

It appears very challenging to develop a test of positive selection that is robust to errors in the sequences or alignments. Instead, we advise to carefully check the alignments of genes that are putatively under selection by any method described here.

### 4.2. Overlapping Reading Frames

Another line of development in modeling the evolution of protein-coding genes concerns evaluating selective pressures on overlapping reading frames (ORFs). In particular, viruses are known to frequently encode genes with ORFs to maximize information content of their short genomes. This may increase codon bias and affect evolutionary constraints on overlapping regions. Indeed, regions of genes that encode several protein products evolve under constraints imposed on each frame, which is disregarded in standard codon models. Although less common, ORFs are also found in eukaryotic genomes.

Some solutions for modeling overlapping regions have been proposed. A nonstationary model can fully accommodate complex site dependencies caused by ORFs and other effects, such as methylation, but requires a conditional Markov process of a higher order with $61N$x$61N$ instantaneous rate matrix so that instantaneous rates at a base are dependent on the neighboring nucleotide states (see refs. 81 and 82). The ML parameter estimation is analytically intractable for such model. When applied only to pairs of sequences, the model requires MCMC for parameter estimation. To speed up the computation under such site-dependent model, an approximate estimation method can be used, based on the pseudo-likelihood via expectation–maximization (EM) algorithm (83). The process of context-dependent substitution may be extended to a general phylogeny at the expense of limiting the full process-based Jensen–Pedersen model (84). A second-order Markov process running at the tips of a tree is an approximation since interdependencies in the ancestral sequences are ignored. The likelihood is calculated with a modified pruning algorithm and optimized with EM.

Instead, computationally simple approximations may be used. For example, Sabath, Landan, and Graur (85) extended the simple GY codon model to accommodate different average selective pressures in two overlapping genes using an additional $\omega$-parameter for the second gene. This model, however, assumes a multiplicative selective effect in ORF and uniform selective pressures within each gene. Another alternative is to define a phylo-HMM with hidden classes being the degeneracy classes, which include the possible outcomes of ORFs (see refs. 86, 87, and 88). Such phylo-HMM also assumes the constancy of selective pressure over time and in the sequence and that degeneracy of a site is constant over time. It is not known whether for the estimates of selective pressure in overlapping genes these assumptions are more detrimental compared to those made in the model of Sabath et al. (85). Further improvements in codon models are needed to describe the evolution of ORFs more realistically to provide more accurate estimates of selection in gene regions with ORFs.

### 4.3. Recombination

Most codon models assume a single phylogeny and a constant synonymous rate among sites, implying that rate variation among codons is solely due to the variation of the nonsynonymous rate. Recent studies question whether such assumptions are generally realistic (e.g., see ref. 89) suggested that failure to account for synonymous rate variation may be one of the reasons why LRTs for positive selection are vulnerable on data with high recombination rates. Some selection scans try to control this problem by checking putatively selected genes for recombination either manually or automated with traditional detection software (e.g., RDP (90)). Also Drummond and Suchard (91) have recently developed a Bayesian approach to detect recombination within a gene.

Another approach is to explicitly consider recombination. For example, Scheffler, Martin, and Seoighe (92) extended codon models with both $d_N$ and $d_S$ site variation and allowed changes of topology at the detected recombination breakpoints. Certainly, fast-evolving pathogens (such as viruses) undergo frequent recombination which often changes either the whole shape of the underlying tree or only the apparent branch lengths. While the efficiency of the approach depends on the success of inferring recombination breakpoints, the study demonstrated that taking into account alternative topologies achieves a substantial decrease of false-positive inferences of selection while maintaining reasonable power. In a related development, Wilson and McVean (93) used an approximation to a population genetics coalescent with selection and recombination. Inference was performed on both parameters simultaneously using the Bayesian approach with reversible-jump MCMC.

### 4.4. Biased Gene Conversion

Mutation rate variation can also cause genomic regions to have different substitution rates without any change in fixation rate. Recent studies of guanine and cytosine (GC)-isochores in the mammalian genome have suggested the importance of another selectively neutral evolutionary process that affects nucleotide evolution. As described in the work of Laurent Duret and others (see refs. 94 and 95), biased gene conversion (BGC) is a mechanism caused by the mutagenic effects of recombination combined with the preference in recombination-associated DNA repair toward strong (GC) versus weak (adenine and thymine [AT]) nucleotide pairs at non-Watson–Crick heterozygous sites in heteroduplex DNA during crossover in meiosis. Thus, beginning with random mutations, BGC results in an increased probability of fixation of G and C alleles. In particular, methods looking for accelerated regions in coding DNA but also codon models cannot distinguish positive selection from BGC (see refs. 96 and 97). Therefore, the putatively selected genes should be checked for GC content, and closeness to recombination hot spots and telomeres. A recent study by Yap et al. (98) suggests that modeling nucleotide target frequencies to be conditional on the other nucleotides in the codon should help to alleviate codon-dependent biases, like BGC and CpG biases.

### 4.5. Selection on Synonymous Sites

Most selection studies to date focused on detecting selection on the protein, since synonymous changes are often presumed neutral and so unaffected by selective pressures. However, selection on synonymous sites has been documented more than a decade ago. Codon usage bias is known to affect the majority of genes and species. In his seminal work, Akashi (99) demonstrated purifying selection on genes of *D. melanogaster*, where strong codon bias favoring certain (optimal) codons serves to increase the translational accuracy. Pressure to optimize for translational efficiency, robustness, and kinetics leads to synonymous codon bias, which was shown to widely affect mammalian genes (100), as well as genes of fast-evolving pathogens like viruses (101). Positive selection on synonymous sites has been unheard of until recently when Resch et al. (102) conducted a large-scale study of selection on synonymous sites in mammalian genes. They measured selection by comparing the average rate of synonymous substitutions ($d_S$) to the average substitution rate in the corresponding introns ($d_I$). While purifying selection was found to affect 28% of genes ($d_S/d_I < 1$), 12% of genes were found to have been affected by positive selection on synonymous sites ($d_S/d_I > 1$). The signal of positive selection correlated with lower predicted mRNA stability compared to genes with negative selection on synonymous sites, suggesting that mRNA destabilization (affecting mRNA levels and translation) could be driving positive selection on synonymous sites.

An increasing number of experimental studies may now explain how synonymous mutation may be affected by positive or negative selection. Codon bias to match skews of tRNA abundances may influence translation (103). Changes at silent sites can disrupt splicing control elements and create new “cryptic” splice sites, as well as mRNA and transcript stability can be affected through preference or avoidance of certain sequence motifs (see refs. 104 and 100). Silent changes may affect gene regulation via constraints for efficient binding of miRNA to sense mRNA (see refs. 105 and 100). Cotranslational protein folding hypothesis suggests that speed-dependent protein folding may be another source of selective pressure (106) because slower production could cause the protein to take an altered final form (as has been shown in multidrug resistance-1 (107)). Finally, synonymous changes may act to modulate expression by altering mRNA secondary structure, affecting protein abundance (108).

Models of codon evolution currently provide the best approach for studying selection on silent sites. In particular, models with variable synonymous rates (see refs. 64 and 109) may be applied to evaluate the extent of variability of synonymous rates in a gene and to predict the positions of most conserved and most variable synonymous sites (for example, see ref. 101). Whether or not the site has been affected by selection requires further testing. For example, Zhou, Gu, and Wilke (110) suggested distinguishing two types of synonymous substitution rates: the rate of conserving synonymous changes $d_{SC}$ (between “preferred” codons or between “rare” codons) and the rate of nonconserving synonymous changes $d_{SN}$ (between codons from the two different groups “rare” and “preferred”). Silent sites with $d_{SN}/d_{SC} > 1$ may be considered to be under positive selection, and significance can be tested based on an LRT. Alternatively, synonymous rates at sites may be compared to the mean substitution rate in the corresponding intron, which can be implemented in a joint codon and DNA model, similar to the approach proposed by Wong and Nielsen (111).

While selection on codon usage bias is typically studied with various codon adaptation indexes (see ref. 112 for review), several codon models were developed for this task (see refs. 113, 114, and 115). The mutation-selection models include selective and mutational effects separately and allow estimating the fitness of various codon changes. The relative rate of substitution for selected mutations to neutral mutations is given by $\omega = 2\gamma/(1 - e^{-2\gamma})$ , where $\gamma = 2Ns$ is the scaled selection coefficient (see Exercise 3 for a derivation). Nielsen et al. (114) assumed that all changes between preferred and rare codons have the same fitness (and so the same selection coefficient). They used one selection coefficient for optimal codon usage for each branch of a phylogeny, and estimated these jointly with the $\omega$-ratio by ML. Using this approach to study ancestral codon usage bias, Nielsen et al. (114) confirmed the reduction in selection for optimal codon usage in *D. melanogaster.* In contrast, Yang and Nielsen (2008) estimated individual codon fitness parameters and used them to estimate optimal codon frequencies for a gene across multiple species. LRT is used to test whether the codon bias is due to the mutational bias alone. Finally, one remarkable contribution of the mutation-selection models is the connection they make between the interspecific and population parameters. Exploiting this further should provide insights into how changing demographic factors influence observed intraspecific patterns.

## 5. Outlook: Selection Scans Using Population Data

By modeling genome evolution as a process by which a single genome sequence mutates along the branches of a species phylogeny, standard phylogenetic methods reduce the entire populations to single points in genotypic space. In reality, each population consists of many individuals—or more precisely, chromosomes from these individuals—that are related by trees of genetic ancestry known as genealogies. With the publication of large amounts of genome-wide polymorphism data, it is now possible to study the role of advantageous mutations. Many population genomic techniques can be applied to noncoding and coding regions. Here, we focus on scans for selection acting on protein-coding genes. Methods for the analysis of noncoding regions are discussed in Chapter 6 of this Volume (116).

### *5.1. Neutrality Tests with a Focus on Protein-Coding Genes*

Many methods have been proposed for population data. Tajima's $D$-test (for DNA data) compares the estimate of the population-scaled mutation rate based on the number of pairwise differences with that based on the number of segregating sites in a sample (117). Under neutrality, Tajima's $D \approx 0$ and significant deviations may indicate a selective sweep $(D < 0)$ or balancing selection $(D > 0)$. Other neutrality tests are based on a similar idea but use different summary statistics (e.g., see refs. 118 and 119). The Hudson–Kreitman–Aguade (HKA) test for DNA data evaluates the neutral hypothesis by comparing variability within and between species for two or more loci (120). Under neutrality, levels of polymorphism (variability within species) and divergence (variability between species) should be proportional to the mutation rate, resulting in a constant polymorphism-to-divergence ratio. Tests of selective neutrality based solely on simple summary statistics are successful at rejecting the strictly neutral model but are sensitive to demographic assumptions, such as constant population size, no population structure, and migration (see refs. 121 and 122). While simple neutrality tests are not specific to coding data, performing such tests separately for synonymous and nonsynonymous changes can potentially help separating selective and demographic effects. Indeed, the popular McDonald–Kreitman (MK) test for protein-coding data exploits the underlying idea of the HKA test, but classifies the observed changes into synonymous and nonsynonymous (123). The MK test compares the ratio of nonsynonymous (amino acid altering) to synonymous (silent) substitutions within and between species, which should be the same in the absence of selection. This test is more robust to demographic assumptions, as the effect of the demographic model should be the same for both nonsynonymous and synonymous sites (122). Whereas the population demographic process is expected to affect all genomic loci, selection should be nonuniform. Several studies (see refs. 124, 125, and 126) took a genomic approach and confirmed that polymorphism-to-divergence ratios differed significantly only for a few genes, although the high amounts of inferred adaptation exceeded expectations.

Apart from biasing the mutation frequency distribution, selection may also affect the distribution of genealogical shapes in population data. Drummond and Suchard (91) proposed a Bayesian test for neutrality that takes into account the distribution of genealogical shapes and can test for both selection and recombination. Such test should be relevant particularly for protein-coding sequences, where most selection is expected to operate. More generally, methods that use information from both the mutation frequency spectrum and the shape of the genealogies are expected to be more powerful than when either used individually.

Unlike neutrality tests that do not explicitly model selection, the Poisson random-field framework (see refs. 127–130 and 131) enables estimation of mutation and selection parameters in various population genetics scenarios. The rationale behind the approach is that natural selection alters the site-frequency spectrum, making it possible to estimate the strength of selection that has contributed to the observed deviation from neutrality. Boyko et al. (132) estimated ~10% of adaptive amino acid changes in humans, but the proportion of adaptively driven substitutions is higher than 50% in some microorganisms and Drosophila (see refs. 125, 133, and 134). Also current estimates might be biased downwardly in the presence of slightly deleterious mutations and decreasing population *size* (135).

Recently, Gutenkunst et al. (136) have developed methods for multidimensional site frequency spectra. These allow the joint inference of the demographic history of multiple populations. Nielsen et al. (137) used a 2D site frequency spectrum to study the Darwinian and demographic forces in protein-coding genes from two human populations. In the future, we can expect to study selection on protein-coding genes in more populations from more species as new sequencing technologies and new methods for detecting selection in population data will be developed.

## 6. Exercises

**Q1. Amino acid and codon substitution models**: How many parameters need to be estimated in the instantaneous rate matrix Q defining a reversible empirical AA model? How many such parameters are necessary to estimate for a reversible empirical codon model? How many parameters are to be estimated in both cases if a model is nonreversible?

**Q2. Positive selection scans**: Go to the UCSC genome browser (http://genome.ucsc.edu). Search for the HAVCR1 (hepatitis A virus cellular receptor 1) in the human genome (assembly NCBI36/hg18) belonging to the mammalian clade.

Genome browser tracks provide the summary of previous analysis of coding regions. Switch “Pos Sel Genes” under “Genes and Gene Prediction Tracks” to “full” and collect information on the LRTs that were performed for the six species scan. Next, switch the “17-Way Cons” under “Comparative Genomics” to full. Why are only a few bases in the HAVCR1 gene conserved? Is this consistent with the results obtained by LRTs?

Click on the “Conservation” track to retrieve the multiple sequence alignment for the HAVCR1 gene. Use the PAML software (http://abacus.gene.ucl.ac.uk/software/paml.html) to test the models for positive selection on any lineage of the mammalian tress by comparing models M1a and M2a with an LRT.

Use PAML to identify sites under positive selection by using the BEB approach. Do you find the same sites to be under selection as in Fig. 3 of Kosiol et al. (43)?

**Q3. Selection-mutation models**: Models incorporating selection and mutation rely on a theoretical relationship between the nonsynonymous–synonymous rate ratio $\omega$ and the scaled selection coefficient $\gamma = 2Ns$. The probability that a new mutation eventually becomes fixed is

$$\Pr(\text{fixation}) = \frac{1 - e^{-2s}}{1 - e^{-4Ns}} = \frac{2s}{1 - e^{-4Ns}}$$

if we assume that the selection coefficient $s$ is small and $N$ is large and represents the effective population size, which is constant in time (138). Furthermore, assume that synonymous substitutions are neutral and nonsynonymous have equal (and small) selection coefficients. Derive the relationship

$$\omega = \frac{4s}{1 - e^{-4Ns}} = \frac{2\gamma}{1 - e^{-2\gamma}}$$

that combines phylogenetic with population genetic quantities and is crucial for mutation-selection models.

## Acknowledgments

C.K. is supported by the University of Veterinary Medicine Vienna. M.A. is supported by the ETH Zurich and also receives funding from the Swiss National Science Foundation (grant 31003A_127325).

### References

1. Pal C, Papp B, Lercher MJ (2006) An integrated view on protein evolution. Nature Rev Genet 7:337–348
2. Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Gräf S, Haider S, Hammond M, Howe K, Jenkinson A, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Koscielny G, Kulesha E, Lawson D, Longden I, Massingham T, McLaren W, Megy K, Overduin B, Pritchard B, Rios D, Ruffier M, Schuster M, Slater G, Smedley D, Spudich G, Tang YA, Trevanion S, Vilella A, Vogel J, White S, Wilder SP, Zadissa A, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Herrero J, Hubbard TJ, Parker A, Proctor G, Smith J, Searle SM (2010) Ensembl's 10th year. Nucleic Acids Research 38: D557–D562
3. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Res 39:D876-D882
4. Altenhoff AM, Dessimoz C (2012) Inferring orthology and paralogy. In: Anisimova M (ed) Evolutionary genomics: statistical and

computational methods (volume 1). Methods in Molecular Biology, Springer Science+ Business Media New York

5. Lee H, Tang H (2012) Next generation sequencing technology and fragment assembly algorithms. In: Anisimova M (ed) Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business Media New York
6. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Xuan Z, Wang W, Li J *et al.* (2010) The sequence and de novo assembly of the giant panda genome. Nature 463:311–317
7. Posada D, Crandall KA (2002) The effect of recombination on the accuracy of phylogenetic estimation. J Mol Evol 54:396–402
8. Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538
9. Semple C Wolfe KH (1999) Gene duplication and gene conversion in the caenorhabditis elegans genome. J Mol Evol 48:555–564
10. Doolittle WF (1999) Phylogentic classification and the universal tree. Science 284:2124–2129
11. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003) Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20:1692–1704
12. Choi SC, Holboth A, Robinson DM, Kishino H, Thorne JL (2007) Quantifying the impact of protein tertiary structure on molecular-evolution. Mol Biol Evol 24:1769–1782
13. Keilson J (1979). Markov Chain Models-Rarity and Exponentiality. Springer, New-York
14. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Perdersen JS, Berjerano G, Baertsch R, Rosenblum KR, Kent J, Haussler D (2006) Frorces shaping the fastest evolving regions in the human genome, PLoS Genetics 2(10): e168.
15. Holloway AK, Begun DJ, Siepel A, Pollard K (2008) Accelerated sequence divergence of conserved genomic elements in Drosophila melanogaster. Genome Res 18:1592–1601
16. Miyamoto MM, Fitch WM (1995) Testing the covarion hypothesis of molecular evolution. Mol Biol Evol 12:503–513
17. Lockhart PJ, Steel MA, Barbrook AC, Huson DH, Charleston MA, Howe CJ (1998) A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol 15:1183–1188
18. Penny D, McComish BJ, Charleston MA, Hendy MD (2001) Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol 53:711–753
19. Siltberg J, Liberles DA (2002) A simple covarion-based approach to analyse nucleotide substitution rates. J Evol Biol 15:588–594
20. Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Evol 257:342–358
21. Gu X (1999) Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664–1674
22. Armon A, Graur D, Ben-Tal N (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 307:447–463
23. Gaucher EA, Gu X, Miyamoto MM, Benner SA (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci 27: 315–321
24. Pupko T, Galtier N (2002) A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc Biol Sci 269:1313–1316
25. Blouin C, Boucher Y, Roger AJ (2003) Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. Nucleic Acids Res 31:790–797
26. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33: W299–W302
27. Gu X (2001) Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol 18:453–464
28. Gu X (2006) A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. Mol Biol Evol 23:1937–1945
29. Siepel A, Haussler D (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. J Comput Biol 11:413–428
30. Siepel A, Haussler D (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 21:468–488
31. Bofkin L, Goldman N (2007) Variation in evolutionary processes at different codon positions. Mol Biol Evol 24:513–521

32. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167–170
33. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32–43
34. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736
35. Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–724
36. Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864
37. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573
38. Schneider A, Cannarozzi GM, Gonnet GH (2005) Empirical codon substitution matrix. BMC Bioinformatics 6:134
39. Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. Mol Biol Evol 24:1464–1479
40. Doron-Faigenboim A, Pupko T (2007) A combined empirical and mechanistic codon model. Mol Biol Evol 24:388–397
41. Whelan S, Goldman N (1999) Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol Biol Evol 16:1292–1299
42. Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18:1585–1592
43. Kosiol C, Vinar T, Da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, and Siepel A (2008) Patterns of positive selection in six mammalian genomes. PLoS Genet 4: e10000144
44. Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950–958
45. Yang Z, Wong WS, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118
46. Yang Z, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431–449
47. Huelsenbeck JP, Dyer KA (2004) Bayesian estimation of positively selected sites. J Mol Evol 58:661–672
48. Scheffler K, Seoighe. C (2005) A Bayesian model comparison approach to inferring positive selection. Mol Biol Evol 22:2531–2540
49. Aris-Brosou S, Bielawski JP (2006) Large-scale analyses of synonymous substitution rates can be sensitive to assumptions about the process of mutation. Gene 378:58–64
50. Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. Genetics 169:1753–1762
51. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) GARD: a genetic algorithm for recombination detection. Bioinformatics 22:3096–3098
52. Kosakovsky Pond SL, Posada, D Gravenor MB, Woelk,CH and Frost SD (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23:1891–1901
53. Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland Massachusetts
54. Yang Z, Dos Reis M (2011) Statistical properties of the branch-site test of positive selection. Mol Biol Evol 28:1217–1228
55. Anisimova M, Yang Z (2007) Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol 24:1219–1228
56. Kosakovsky Pond SL., and Frost SD (2005) A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol 22:478–485
57. Lemmon AR, and Milinkovitch MC (2002) The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. Proc Natl Acad Sci U S A 99:10516–10521
58. Jobb G, von Haeseler A, and Strimmer K (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol 4:18
59. Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD dissertation, The University of Texas, Austin.
60. Guindon S.A, Rodrigo G, Dyer KA, Huelsenbeck JP (2004) Modeling the site-specific variation of selection patterns along lineages. Proc Natl Acad Sci U S A 101:12957–12962
61. Siepel A, Bejerano G, Pedersen JS, Hinrichs A, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ,

Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 20: 1034–1050

62. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of non-neutral substitution rates on mammalian phylogenies. Genome Res 20: 110–121

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

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

65. Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, and Pupko T (2007) Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res 35:W506-511

66. Klosterman PS, Uzilov AV, Bendana YR, Bradley RK, Chao S, Kosiol C, Goldman N, Holmes I (2006) XRate: a fast prototyping, training and annotation tool for phylo-grammars. BMC Bioinformatics 7: 428

67. Heger A, Ponting CP, Holmes I (2009) Accurate estimation of gene evolutionary rates using XRATE, with an application to transmembrane proteins. Mol Biol Evol 26:1715–1721

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

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

70. Vamathevan JJ, Hasan S, Emes RD, Amrine-Madsen H, Rajagopalan D, Topp SD, Kumar V, Word M, Simmons MD, Foord SM, Sanseau P, Yang Z, Holbrook JD (2008) The role of positive selection in determining the molecular cause of species differences in disease. BMC Evol Biol 8:273

71. Nozawa M, Suzuki Y, Nei M (2009) Reliabilities of identifying positive selection by the branch-site and site-prediction methods. Proc Natl Acad Sci USA 106:6700–6705

72. Markova-Raina P, Petrov D (2011) High sensitivity to aligner and high rate of false positives in the estimates of positive selection in 12 Drosophila genomes. Genome Res. doi:10.1101/gr.115949.110

73. Bakewell MA, Shi P, Zhang J (2007) More genes underwent positive selection in chimpanzee than in human evolution. Proc Natl Acad Sci USA 104:E97

74. Arbiza L, Dopazo J, Dopazo H (2006) Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput Biol 2:e38

75. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK *et al.* (2007) Evolutionary and biomedical insights from the macaque genome. Science 316:222–234

76. Mallik S, Gnerre S, Muller P, Reich D (2010) The difficulty of avoiding false positives in genome scans for natural selection. Genome Res 19:922–933

77. Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH (2009) Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol 1:114–118

78. Fletcher W, Yang Z (2010) The effect of insertions, delections and alignment errors on the branch-site test of positive selection. Mol Biol Evol 27:2257–2267

79. Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A 102:10557–10562

80. Löytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents error in sequence alignment and evolutionary analysis. Science 320:1632–1635

81. Jensen JL, Pedersen AK (2000) Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv Appl Probab 32:499–517

82. Pedersen AK, Jensen JL (2001) A Dependent-Rates Model and an MCMC-Based Methodology for the Maximum-Likelihood Analysis of Sequences with Overlapping Reading Frames. Mol Biol Evol (2001) 18:763–776

83. Christensen OF, Hoboth A, Jensen JL (2005) Pseudo-likelihood analysis of context dependent codon substitution models. J Comp Biol 12:1166–1182

84. Siepel A, Haussler D (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 21:468–488

85. Sabath N, Landan G, Gaur D (2008) A method for the simultaneous estimation of selection intensities in overlapping genes. PLoS One 3:e3996

86. De Groot S, Mailund T, Hein J (2007). Comparative annotation of viral genomes

with non-conserved genestructure. Bioinformatics 23:1080–1089

87. McCauley S, Hein J (2006) Using hidden Markov models (HMMs) and observed evolution to annotate ssRNA Viral Genomes. Bioinformatics 22: 1308–1316
88. McCauley S, de Groot S, Mailund T, Hein J (2007) Annotation of selection strength in viral genomes. Bioinformatics 23:2978–2986
89. Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229–1236
90. Martin DP, Williamson C, Posada D (2005) RDP2: recombination detection and analysis of sequence alignments. Bioinformatics 21:260–262
91. Drummond AJ, Suchard MA (2008) Fully Bayesian tests of neutrality using genealogical summary statistics. BMC Genet 9:68
92. Scheffler K, Martin DP, Seoighe C (2006) Robust inference of positive selection from recombining coding sequences. Bioinformatics 22:2493–2499
93. Wilson DJ, McVean G (2006) Estimating diversifying selection and functional constraint in the presence of recombination. Genetics 172:1411–1425
94. Duret L, Semon M, Piganeau G, Mouchiroud D, Galtier N (2002) Vanishing GC-rich isochores in mammalian genomes. Genetics 162:1837–1847
95. Meunier J, Duret L (2004). Recombination drives the evolution of GC content in the human genome. Mol Biol Evol 21:984–990
96. Berglund J, Pollard KS, Webster MT (2009) Hotspots of biased nucleotide substitutions in human genes. PLoS Biology 7:e26
97. Ratnakumar A, Mousset S, Glemin S, Berglund J, Galtier N, Duret L, Webster MT (2010) Detecting positive selection within genomes: the problem of biased gene conversion. Phil Trans Roy Soc B 365:2571–2580
98. Yap B, Lindsay H, Easteal S, Huttley G (2010) Estimates of the effect of natural selection on protein-coding content. Mol Biol Evol 27:726–734
99. Akashi H (1994) Synonymous codon usage in Drosophila melanogaster: Natural selection and translational accuracy. Genetics 136:927–935
100. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98–108
101. Ngandu N, Scheffler K, Moore P, Woodman Z, Martin D, Seoighe C (2009) Extensive purifying selection acting on synonymous sites in HIV-1 Groug M sequences. Virol J 5:160
102. Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV (2007) Widespread Positive Selection in Synonymous Sites of Mammalian Genes. Mol Biol Evol 24:1821–1831
103. Cannarozzi GM, Faty M, Schraudolph NN, Roth A, von Rohr P, Gonnet P, Gonnet GH, Barral Y (2010) A role for codons in translational dynamics, Cell 141:355–367
104. Hurst LD, Pál C (2001) Evidence of purifying selection acting on silent sites in BRCA1. Trends Genet 17: 62–65
105. Chamary JV, Hurst LD (2005) Biased usage near intron-exon junctions: selection on splicing enhancers, splice site recognition or something else? Trends Genet 21:256–259
106. Komar AA (2008) Protein translational rates and protein misfolding: Is there any link? In: O'Doherty CB, Byrne AC (eds) Protein Misfolding: New Research. Nova Science Publisher Inc, New York.
107. Kimichi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM (2007) A silent polymorphism in the MDR1 gene changes substrate specificity. Science 315:525–528
108. Nackley AG, SA Shabalina, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L (2006) Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science 314:1930–1933
109. Mayrose I, Doron-Faigenboim A, Bacharach E, Pupko T (2007) Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. Bioinformatics 23:i319-327
110. Zhou T, Gu W, Wilke CO (2010) Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol 27: 1912–1922
111. Wong WSW, Nielsen R (2004). Detecting selection in non-coding regions of nucleotide sequences. Genetics 167:949–958
112. Roth A, Anisimova M, Cannarozzi GM (2011) Measuring codon usage bias. In: Cannarozzi G, Schneider A (eds) Codon Evolution: mechanisms and models. Oxford University Press
113. Nielsen R, Yang Z (2003) Estimating the distribution of selection coefficients from phylogenetic data with applications to

mitochondrial and viral DNA. Mol Biol Evol 20:1231–1239

114. Nielsen R, Bauer DuMont VL, Hubisz MJ, Aquadro CF (2007) Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. Mol Biol Evol 24:228–235
115. Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25:568–579
116. Zhen Y, Andolfatto P (2012) Detecting selection on non-coding genomics regions. In: Anisimova M (ed) Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business Media New York
117. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595
118. Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709
119. Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413
120. Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159
121. Wayne ML, Simonsen K (1998) Statistical tests of neutrality in the age of weak selection. Trends Ecol Evol 13:1292–1299
122. Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. Heredity 86:641–647
123. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654
124. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158:1227–1234
125. Eyre-Walker A (2002) Changing effective population size and the McDonald–Kreitman test. Genetics 162:2017–2024
126. Smith NG, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415:1022–1024
127. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132:1161–1176
128. Hartl DL, Moriyama EN, Sawyer SA (1994) Selection intensity for codon bias. Genetics 138:227–234
129. Akashi H (1999) Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151:221–238
130. Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan, Hartl DL (2002) The cost of inbreeding: fixation of deleterious genes in Arabidopsis. Nature 416:531–534
131. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd-Hubisz M, Glanowski S, Hernandez R, Civello D, Tanebaum DM, White TJ, Sninsky JJ, Adams MD, Cargill M, Clark AG (2005) Natural selection on protein coding genes in the human genome. Nature 437:1153–1157
132. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genetics 4(5): e1000083
133. Bierne N, Eyre-Walker A (2004) Genomic rate of adaptive amino acid substitution in Drosophila. Mol Biol Evol 21:1350–1360
134. Welch JJ (2006) Estimating the genome-wide rate of adaptive protein evolution in Drosophila. Genetics 173: 821–837
135. Eyre-Walker A, and Keightley PD (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Bio Evol 26:2097–2018
136. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from SNP data. PLoS Genetics 5: e1000695
137. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, Indap A, Bustamante CD, Clark AG (2009) Darwinian and demographic forces affecting human protein coding genes. Genome Res 19:838–849
138. Kimura M, Ohta T (1969) The average number of generations until fixation of a mutant gene in a finite population. Genetics 61:763–771