Chapter 12

Selection Acting on Genomes

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 genomes 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 genome evolution of various species. In this chapter, we will focus on methods to detect selection on the genome. In particular, we will 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, Selection-mutation models, Polymorphism-aware phylogenetic models

1 Introduction

In the past selection studies mainly focused on the analysis of particular loci such as genes, proteins, or regular elements 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 search for selection can be performed on different levels comparing homologous nucleotide sequences or protein-coding genes in one or multiple genomes. The evolutionary processes in all these levels can be described by probabilistic models, which set the basis for evaluating selective pressures and selection tests. This book chapter will give an introduction into fundamental properties of the probabilistic models used to detect selection in the Subheading 3 as well as examples of genome-wide scans.

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology, vol. 1910, https://doi.org/10.1007/978-1-4939-9074-0_12, © The Author(s) 2019

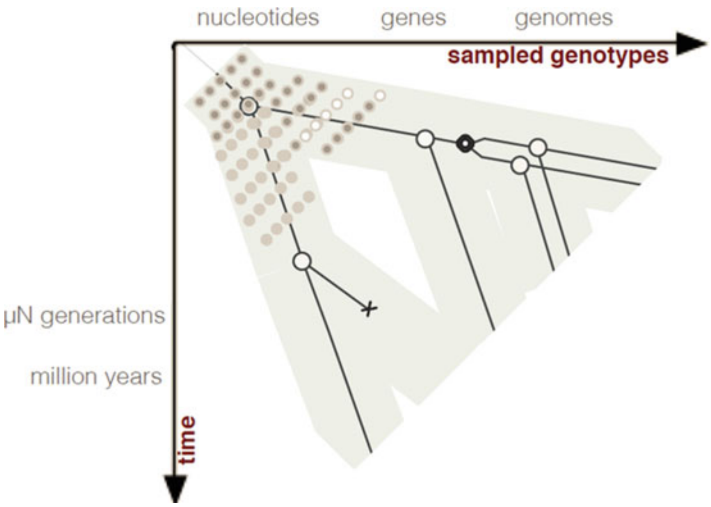

Fig. 1 A diagram illustrating the different data and levels to analyze genomic sequences and the relationship of the various approaches modeling selection

In Fig. 1, we summarize the different data levels and time scales of modeling selection on genomes.

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 [1] and UCSC [2] browsers.

In light of recent advances in DNA sequencing, with 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. Just to name a few, genome projects re-sequencing 1000 *D. melanogaster* [3] and 1001 *Arabidopsis* [4] were accomplished, and the 100,000 human genome project [5] is 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, including complex mammalian genomes (e.g., giant panda genome [6]). Therefore, not only international consortia but also small groups and individual labs can now envisage to sequence the organisms of their interest. As a consequence platforms for sharing this data have been established. For example, the Genome 10K project aims to assemble a genomic zoo—a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one 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 often explicitly model the selection pressures.

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* [8, 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 affects 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* [11, 12]).

The Markov model asserts that one 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 Q-matrix 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 (Q-matrix of a reversible process has only real eigenvectors and eigenvalues [13]). Fully unrestrained Q-matrix 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* [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 ≥ 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 *Drosophila 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* [16–19]). Other methods search for evidence of clade-specific rate shifts at individual sites (*see* [20–26]). For example, Gu [21] proposed a simple stochastic model for estimating the degree of divergence between two pre-specified clusters. The statistical significance was tested using site-specific profiles based on a hidden Markov model, 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 Codon Models: Site, Branch, and Branch-Site Specificity

3.3.1 Basic Codon Models

In protein-coding sequences, nucleotide sites at different codon positions usually evolve with highly heterogeneous patterns (e.g., [29]). 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 ω 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* [30]). However, ML estimates of pairwise d_N and d_S based on a codon model were shown to outperform all other approaches [31]. 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 [32, 33]. The model of Goldman and Yang [32] included the transition/transversion rate ratio κ, and modeled the selective effect indirectly using a multiplicative factor based on Grantham [34] distances, but was later simplified to estimate the selective pressure explicitly using the ω parameter [35]. 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 [32]) or (2) proportional to the frequency of a target nucleotide (as in Muse and Gaut [33]).

In 2006, empirical codon models have been estimated (*see* [36, 37]) 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 κ, selective pressure ω, 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. [37] also found that the affiliations between 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 [37] 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 ω, κ, and codon frequencies π_j [38]. In 2013, De Maio et al. [39] extended the ECM approach to accommodate site-specific variation of selective pressure and lineage-specific variation. Simulations showed that ECMs allowing for double and triple mutations is more conservative: they reduce the number of false positives and have less power to detect positive selection [39].

3.3.2 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 a 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 χ^2 due to boundary problems or if some parameters become not estimable (e.g., *see* [40, 41]).

3.3.3 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 [42] 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 78 genes associated with this pathway in KEGG (*see* http://www.genome.jp/kegg-bin/show_pathway?map04610 for the complement cascades), nine were under positive selection (FDR < 0.05), and five others had nominal $P < 0.05$. Most of 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 ω value using the basic M0 model [32]. 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.

3.3.4 Selective Variability Among Codons: Site Models

The simplest site models use the general discrete distribution with a pre-specified number of site classes. Each site class i has an independent parameter ω_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 χ^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(log L2 − log L1) = 2 (−6377.35 − (−6369.67)) = 2 × 7.68 = 15.36. This is much greater than the critical value of the chi-square distribution χ^2 (df = 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.

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

M1a (neutral)			
Site class	0	1	
Proportion	$p_0 = 0.69$	$(p_1 = 1 - p_0 = 0.31)$	
ω ratio	$\omega_0 = 0.07$	$(\omega_1 = 1)$	
Log-likelihood L1 $= -6377.35$			
M2a (selection)			
Site class	0	1	2
Proportion	$p_0 = 0.70$	$p_1 = 0.29$	$(p_2 = 1 - p_0 - p_1 = 0.01)$
ω 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

Another LRT can be performed on the basis of the modified model M8 with two site classes: one with sites where the ω 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 flexible null hypothesis, whereby all evolution can be explained by sites with ω from the beta distribution or from a discrete site class with $\omega = 1$. Significance of the LRT is tested 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 (PP) to belong to the site class under positive selection (usually PP $>$ 0.95, but *see* [43, 44]). Such posterior probabilities are estimated using the naïve empirical Bayesian approach (NEB, [45]), full hierarchical Bayesian approach ([46]; BEB [44]), or a mid-way approach – the Bayes empirical Bayes (BEB [44]). For a discussion on these approaches, *see* Scheffler and Seoighe [47] and Aris-Brosou [48]. Alternatively, Massingham and Goldman [49] 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 acids 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* [50, 51]); 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* [52]). Given a biologically meaningful a priori partitioning, FE models are useful to study heterogeneity among partitions. However, a priori information is not always available.

3.3.5 Selective Variability over Time: Branch Models

A simple way to include the variation of the selective pressure over time is by using separate parameters ω for each branch of a phylogeny (known as *free-ratio* model; [35]). Compared with the *one-ratio* model (which assumes constant selection over time), the free-ratio model requires additional $2T - 4$ ω parameters for T species. Figure 2 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 ω. LRTs are used to decide (1) whether selective pressure is significantly different on a pre-specified set of branches and (2) whether these branches are under positive selection.

However, branch models have relatively poor power to detect selection [53] 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* Subheading 4 later).

3.3.6 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 $\omega_0 < 1$; another with $\omega_1 = 1$. The other two site

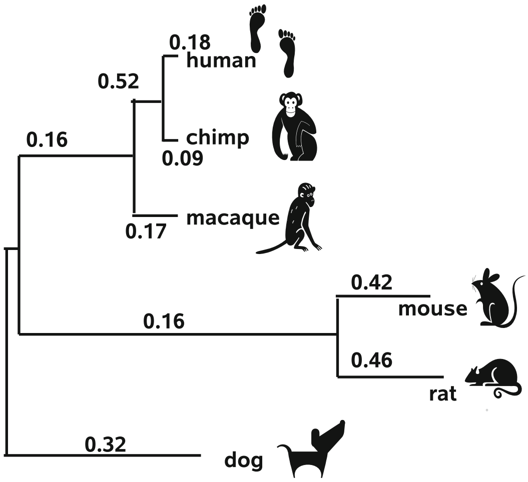

Fig. 2 An estimate of ω for each branch of a six-species phylogeny. Shown is the maximum likelihood estimate for the gene 8B. Each branch is labeled with the corresponding estimate of ω

classes allow selective pressure at a site to change over time on a pre-specified 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 a 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 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(log L2 − log L1) = 2 × 2.23 = 4.46 for the mouse lineage and 11.2 for the rate lineage, respectively.

A major drawback of described branch-site models is their reliance on a biologically viable a priori hypothesis. In 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 [54]. In the example of six species (Fig. 2), 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. 2), 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 [42].

Other models (e.g., with d_S variation among sites [55]) 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 [55] 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. 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. [56] 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).

3.3.7 Polymorphism-Aware Phylogenetic Models

Polymorphism-aware phylogenetic models (POMOs, [57, 58]) use polymorphism and divergence data simultaneously to estimate relative mutation rates and scaled selection coefficients. Similar to DNA substitution models, the PoMo approach is based on a continuous-time Markov process to model evolution of hereditary sequences along a species tree. However, not only evolution of a single reference site but rather evolution of a population is considered.

PoMo includes polymorphisms as states of the Markov chain, in addition to the four nucleotide states of classical nucleotide models. Sequence evolution is modeled as a gradual process made by small allele frequency changes. PoMo accounts for ancestral polymorphisms and in particular for ancestral shared polymorphisms and incomplete lineage sorting (when two speciation events are separated by a lapse of time not sufficient for polymorphisms to reach fixation, *see* Maddison and Knowles [59]). The parameters in PoMo do not merely describe substitution rate but are also informative of mutation rates, fixation biases, root nucleotide frequencies, and branch lengths. All these parameters are estimated within a ML framework. De Maio et al. [57] performed a comprehensive study of evolutionary patterns of fourfold-degenerate sites in great apes populations. They show evidence in favor of variation in mutation and fixation rates between genomic regions with different base composition, contributing to the long-standing debate regarding the origin and maintenance of GC content variation (e.g., *see* Eyre-Walker and Hurst [60]). They found that both mutation rates and biased gene conversion vary with GC content. They also found lineage-specific differences, with weaker fixation biases in orangutan species, suggesting a reduced historical effective population size. As PoMo can distinguish between the contributions of mutation and fixation biases, it might also contribute to addressing the problem of disentangling signatures of selection and biased gene conversion (*see* Subheading 4.2).

3.4 Software

The software PHAST (PHylogenetic Analysis with Space/Time 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 (Siepel et al. [61]). PhyloP is designed to compute *p*-values for conservation or acceleration, either lineage-specific or across all branches (Pollard et al. [62]). Recently, the software can also be run through a webportal at http://compgen.cshl.edu/phastweb/.

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]. PoMo has been implemented as part of the IQ-TREE software package (http://www.iqtree.org/) by Schrempf et al. [65].

These programs are primarily developed for maximum likelihood inference on a fixed tree. ML inference of phylogeny under codon models is possible with CodonPhyML, which allows to explicitly account for selection on the protein level [66].

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* [67] 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* [43, 68]) and is now commonly used in genome-wide selection scans (e.g., *see* [69]). Recently, however, another simulation study by Nozawa et al. [70] suggested that this modification also showed an excess of false positives. Yang and Dos Reis [52] defended the branch-site test by examining the null distribution and showing that Nozawa and colleagues [70] misinterpreted their simulation results. However, it is clear that even tests with good statistical properties will be 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 [71] 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. [72] 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. [73] and Gibbs et al.

[74]. Mallick et al. [75] re-examined 59 genes detected to be under positive selection on the chimpanzee lineage by Bakewell et al. [72], 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. [75], 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* [76, 77]). Indeed, most commonly used alignment programs tend to place nonhomologous codons or amino acids into the same column (*see* [78, 79]), generating the wrong impression that multiple nonsynonymous substitutions occurred at the same site and misleading the codon models into detecting positive selection [77]. In 2012, Jordan and Goldman [80] investigated the effect of various multiple alignment and filtering programs on the identification of positive selection. They found that alignment software PRANK [79] and the filter Guidance [81] performed best in simulations. However, it remains very challenging to develop a pipeline to detect 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 Biased Gene Conversion and Recombination

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* [82, 83]), 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* [84, 85]). Therefore, the putatively selected genes should be checked for GC content and closeness to recombination hotspots and telomeres.

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* [86]) suggesting 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 [87]). Also Drummond and Suchard [88] 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 [89] 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 principle the correlation structure of a collection of orthologous sequences can be fully described by a network known as an ancestral recombination graph (ARG). However, methods for ARG inferences have not been fast enough for practical use, and for applications on large-scale genomic data, approximations are necessary (Rassmussen et al. [90]).

4.3 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 [91] demonstrated purifying selection on genes of *Drosophila 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 [92], as well as genes of fast-evolving pathogens like viruses [93]. The standard approach to study selection on codon usage computes various codon adaptation indexes on full-length protein-coding genes (*see* [94] for review). More recently, methods to study selection on synonymous changes adopted more sophisticated approaches, mainly the following strategies: (1) account for synonymous rate variation within sequences; (2) include codon fitness parameters within a modeling framework that connects population and intraspecific parameters; and (3) allow for selection on synonymous substitutions by introducing the dependency on the rate of protein production and nonsense error rates. Below we elaborate on these approaches.

In the past decade, evidence has accumulated to suggest that codon bias may vary not only between genomes and genes of the same genome but also within genes. Rather than just measuring codon biases in single sequences, a more powerful approach is to model evolution and selection across a set of homologous sequences. Taking the evolutionary perspective into account, Resch et al. [95] 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 exemplify different scenarios explaining how synonymous mutation may be affected by positive or negative selection. Codon bias to match skews of tRNA abundances may influence translation [96]. 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* [92, 97]). Silent changes may affect gene regulation via constraints for efficient binding of miRNA to sense mRNA (e.g., [92, 98]). Selection may act on the choice of synonymous codons near miRNA targets, improving the binding site accessibility, binding efficiency and consequently the function of miRNA itself [99]. Programmed ribosomal frameshifting may be another reason for selection to act on specific codon positions [100]. Speed-dependent protein folding also has been proposed to be a result of selective pressure [101]. According to the co-translational protein folding hypothesis, slower production could cause the protein to take an altered final form (as has been shown in multidrug resistance-1, [102]). Finally, synonymous changes may act to modulate expression by altering mRNA secondary structure, affecting protein abundance [103].

Models of codon evolution currently provide the most powerful approach for studying selection on silent sites. Models with variable synonymous rates (*see* [64, 104]) have been used to evaluate the extent of variability of synonymous rates in a gene and to predict specific sites with most extreme—low or high—synonymous rates (for example *see* [93]). A large-scale study of synonymous rate variation [105] described some intriguing general patterns and showed that the phenomenon is widespread in protein-coding genes. Genes displaying significantly varying synonymous rates increased association with several genetic diseases (especially cancers and diabetes) and were enriched for metabolic pathways. Other studies specifically focusing on human oncogenes revealed that a significant proportion of all cancer driver mutations were synonymous [106]. This suggests that synonymous rates cannot be automatically assumed fitness-neutral. Note that $\omega = d_N/d_S$, an accepted measure of selection on the protein, is not designed to detect selection on synonymous codons, particularly when d_S is assumed constant. Yet, some cautioned that low synonymous rates preserved by purifying selection might erroneously lead to the detection of positive selection on the protein (e.g., Rubinstein et al. [107]). However, the usage of the ω ratio does not rely on the assumption that synonymous sites are neutral (pages 58–59 of Yang [108]; and Section 6.3 of Anisimova and Liberles [109]); rather, it is defined as a ratio of two ratios, comparing the proportions of nonsynonymous and synonymous sites after and before selection has operated on the protein ($\omega = 1$). In general we can assume that the evolutionary forces apply equally to synonymous and nonsynonymous sites. Forces that act differentially on synonymous and nonsynonymous sites should be rare in real data, but they can affect the validity of the ω measure. The only known example of such a natural force is probably synonymous phasing, considered by Xing and Lee [110]. But even in this case, and with a worst case scenario, the estimated effect is very weak. More crucially, an adequate description of mutational processes at the DNA level allows to circumvent biases in the estimation of the ω ratio [106].

Further testing, however, is necessary to decide whether any specific site has been affected by selection on synonymous codon usage. For example, Zhou, Gu, and Wilke [111] 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 non-conserving 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 a likelihood ratio test. 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 [112].

Mutation-selection models include selective and mutational effects separately and allow estimating the fitness of various codon changes (*see* [113–115]). 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 ω 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 [115] 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. Nevertheless, 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 to how changing demographic factors influence observed intraspecific patterns. Mutation-selection models also allow a new perspective on understanding codon models in the context of fitness landscapes with statistical implications as discussed in Subheading 4.2 of Chapter 13 by Jones, Susko, and Bielawski.

Finally, it is also possible to study selection on synonymous changes by introducing a parametric relationship between fitness and protein production cost. The idea was first described by Gilchrist [116], who assumed that, in addition to mutation and drift, the codon bias evolved under selection to reduce the cost of nonsense errors. Protein production cost can be computed as a ratio of the expected cost to the expected benefit [117]. Kubatko and colleagues [118] have extended a standard codon model to include the difference in protein production due to the usage of different codons (and therefore different elongation probabilities). However, such a model requires position-specific instantaneous rate matrices, and consequently also the probability transition matrices, making the approach computationally very intensive. To circumvent this, a GPU-based implementation was developed and used for phylogeny inference from 104 gene data set from *Saccharomyces cerevisiae*. Based on the standard model selection measure AIC, the new model outperformed the simplest model M0 as well as the mutation-selection model FMutSel of Yang and Nielsen.

5 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

1. 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 GRCh38/hg38) belonging to the mammalian clade. The USCS genome browser tracks provide the summary of previous analysis of coding regions. Switch the "Cons_30_Primates" under "Comparative Genomics" to full and "refresh." Why are only a few bases in the HAVCR1 gene conserved according to the PhastCons track? Click on the "Cons_30_Primates" track to learn more about the conservation scores used.
2. To retrieve the multiple sequence alignments for the HAVCR1 gene, go to "Tools" and "Table Browser" at the top bar of the webpage. This will open a new page. Choose the table "ccdsGene" under the "Genes and Gene Predictions" group and "CCDS" track. Select "CDS FASTA alignment from multiple alignment" option in the output format and "Show nucleotides" to download the aligned coding sequences of the HAVCR1 gene. Alternatively you can retrieve the multiple alignments from Ensembl using BioMart. Here, you have options for more file formats including PHYLIP that is needed for the PAML software.
3. 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 trees by comparing models M1a and M2a with a likelihood ratio test.
4. Use PAML to identify sites under positive selection by using the Bayes Empirical Bayes approach. Do you find the same sites to be under selection as in Fig. 2 of Kosiol et al. [43]?

Q3. Selection-Mutation Models

Selection-mutation rely on a theoretical relationship between the nonsynonymous-synonymous rate ratio ω and the scaled selection coefficient $\gamma = 2Ns$. The probability that a new mutation eventually becomes fixed is

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

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 (Kimura and Ohta [119]). Furthermore, assume that synonymous substitutions are neutral and nonsynonymous have equal (and small) selection coefficients. Derive the relationship:

$$\omega = 4s/\left(1 - e^{-4Ns}\right) = 2\gamma/\left(1 - e^{-2\gamma}\right)$$

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

Acknowledgments

C. K. is supported by a grant of the Vienna Science and Technology Fund (WWTF—MA016-061). M. A. receives funding from the Swiss National Science Foundation (grant 31003A_176316).

References

1. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P (2018) Ensembl 2018. Nucleic Acids Res 46:D754–D761
2. Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, Hinrichs AS, Haeussler M, Guruvadoo L, Navarro Gonzalez J, Gibson D, Fiddes IT, Eisenhart C, Diekhans M, Clawson H, Barber GP, Armstrong J, Haussler D, Kuhn RM, Kent WJ (2018) The UCSC Genome Browser database: update 2018. Nucleic Acids Res 46: D762–D769
3. Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE (2016) A thousand fly genomes: an expanded drosophila genome nexus. Mol Biol Evol 33:3308–3313
4. Weigel D, Mott R (2009) The 1001 Genomes Project for Arabidopsis thaliana. Genome Biol 10:107
5. Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, Halai D, Baple E, Craig C, Hamblin A, Henderson S, Patch C, O'Neill A, Devereaux A, Smith K, Martin AR, Sosinsky A, McDonagh EM, Sultana R, Mueller M, Smedley D, Toms A, Dinh L, Fowler T, Bale M, Hubbard T, Rendon A, Hill S, Caulfield MJ, 100,000 Genomes Project (2018) The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ 361:k1687
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) Phylogenetic 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, NY
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) Forces shaping the fastest evolving regions in the human genome. PLoS Genet 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. Bofkin L, Goldman N (2007) Variation in evolutionary processes at different codon positions. Mol Biol Evol 24:513–521
30. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167–170
31. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32–43
32. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736
33. 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
34. Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864
35. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573
36. Schneider A, Cannarozzi GM, Gonnet GH (2005) Empirical codon substitution matrix. BMC Bioinformatics 6:134
37. Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. Mol Biol Evol 24:1464–1479
38. Doron-Faigenboim A, Pupko T (2007) A combined empirical and mechanistic codon model. Mol Biol Evol 24:388–397
39. De Maio N, Holmes I, Schlötterer C, Kosiol C (2013) Estimating empirical hidden Markov models. Mol Biol Evol 30:725–736
40. 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
41. 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
42. Kosiol C, Vinar T, Da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six mammalian genomes. PLoS Genet 4: e10000144
43. 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
44. Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118
45. Yang Z, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449
46. Huelsenbeck JP, Dyer KA (2004) Bayesian estimation of positively selected sites. J Mol Evol 58:661–672

47. Scheffler K, Seoighe C (2005) A Bayesian model comparison approach to inferring positive selection. Mol Biol Evol 22:2531–2540

48. 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

49. Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. Genetics 169:1753–1762

50. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) GARD: a genetic algorithm for recombination detection. Bioinformatics 22:3096–3098

51. Kosakovsky PSL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23:1891–1901

52. Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland, MA

53. Yang Z, Dos Reis M (2011) Statistical properties of the branch-site test of positive selection. Mol Biol Evol 28:1217–1228

54. 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

55. Kosakovsky Pond SL, Frost SD (2005) A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol 22:478–485

56. Guindon SA, 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

57. De Maio N, Schlötterer C, Kosiol C (2013) Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. Mol Biol Evol 30:2249–2262

58. De Maio N, D Schrempf D, Kosiol C (2016) PoMo: an allele frequency-based approach for species tree estimation. Syst Biol 64:1018–1031

59. Maddison W, Knowles L (2006) Inferring phylogeny despite incomplete lineage sorting. Syst Biol 55:21–30

60. Eyre-Walker A, Hurst L (2001) The evolution of isochores. Nat Rev Genet 2:549–555

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. Schrempf D, Minh BQ, De Maio N, von Haeseler A, Kosiol C (2016) Reversible polymorphism-aware phylogenetic models and their application to tree inference. J Theor Biol 407:362–370

66. Gil M, Zanetti MS, Zoller S, Anisimova M (2013) CodonPhyML: fast maximum likelihood phylogeny estimation under codon substitution models. Mol Biol Evol 30:1270–1280

67. 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

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. 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

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

71. 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 21:863. https://doi.org/10.1101/gr.115949.110

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

73. 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

74. 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

75. 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
76. 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
77. Fletcher W, Yang Z (2010) The effect of insertions, deletions and alignment errors on the branch-site test of positive selection. Mol Biol Evol 27:2257–2267
78. 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
79. Löytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents error in sequence alignment and evolutionary analysis. Science 320:1632–1635
80. Jordan G, Goldman N (2012) The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol Biol Evol 29:1125–1139
81. Penn O, Privman E, Landan G, Graur D, Pupko T (2010) An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol 27:1759–1767
82. Duret L, Semon M, Piganeau G, Mouchiroud D, Galtier N (2002) Vanishing GC-rich isochores in mammalian genomes. Genetics 162:1837–1847
83. Meunier J, Duret L (2004) Recombination drives the evolution of GC content in the human genome. Mol Biol Evol 21:984–990
84. Berglund J, Pollard KS, Webster MT (2009) Hotspots of biased nucleotide substitutions in human genes. PLoS Biol 7:e26
85. 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 R Soc B 365:2571–2580
86. 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
87. Martin DP, Williamson C, Posada D (2005) RDP2: recombination detection and analysis of sequence alignments. Bioinformatics 21:260–262
88. Drummond AJ, Suchard MA (2008) Fully Bayesian tests of neutrality using genealogical summary statistics. BMC Genet 9:68
89. Scheffler K, Martin DP, Seoighe C (2006) Robust inference of positive selection from recombining coding sequences. Bioinformatics 22:2493–2499
90. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A (2014) Genome-wide inference of ancestral recombination graphs. PLoS Genet 10(5): e1004342
91. Akashi H (1994) Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics 136:927–935
92. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98–108
93. 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
94. Roth A, Anisimova M, Cannarozzi GM (2012) Measuring codon usage bias. Codon evolution: mechanisms and models. Oxford University Press, New York, NY
95. 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
96. 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
97. Hurst LD, Pál C (2001) Evidence of purifying selection acting on silent sites in BRCA1. Trends Genet 17:62–65
98. 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
99. Gu W, Wang X, Zhai C, Xie X, Zhou T (2012) Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. Mol Biol Evol 29:3037–3044
100. Garcia V, Anisimova M (2018) Accounting for programmed ribosomal frameshifting in the computation of codon usage bias indices. G3 (Bethesda) 8:3173
101. 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, NY

102. 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
103. Nackley AG, Shabalina SA, 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
104. 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–i327
105. Dimitrieva S, Anisimova M (2014) Unraveling patterns of site-to-site synonymous rates variation and associated gene properties of protein domains and families. PLoS One 9 (7):e102721
106. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ (2017) Cell 171:1029–1041.e21
107. Rubinstein ND, Doron-Faigenboim A, Mayrose I, Pupko T (2011) Evolutionary models accounting for layers of selection in protein-coding genes and their impact on the inference of positive selection. Mol Biol Evol 28:3297–3308
108. Yang Z (2006) Computational molecular evolution. Oxford University Press, New York, NY
109. Anisimova M, Liberles DA (2012) Detecting and understanding natural selection. Codon evolution: mechanisms and models. Oxford University Press, New York, NY
110. Xing Y, Lee C (2006) Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes. Nat Rev Genet 7:499–509
111. 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
112. Wong WSW, Nielsen R (2004) Detecting selection in non-coding regions of nucleotide sequences. Genetics 167:949–958
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. Gilchrist MA (2007) Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. Mol Biol Evol 24:2362–2372
117. Gilchrist MA, Shah P, Zaretzki R (2009) Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. Genetics 183:1493–1505
118. Kubatko L, Shah P, Herbei R, Gilchrist MA (2016) A codon model of nucleotide substitution with selection on synonymous codon usage. Mol Phylogenet Evol 94:290–297
119. Kimura M, Ohta T (1969) The average number of generations until fixation of a mutant gene in a finite population. Genetics 61:763–771