Modeling Codon Rate Variation Improves Protein Positive Selection Inference and Detects Nucleotide Selection

Iakov I. Davydov[1,2,3], Nicolas Salamin[1,3], Marc Robinson-Rechavi[2,3,*]

1. Department of Computational Biology, Biophore, University of Lausanne, 1015 Lausanne, Switzerland
2. Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland
3. Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Abstract

There are numerous sources of variation in the rate of synonymous substitutions inside genes, such as direct selection on the nucleotide sequence, or mutation rate variation. However the majority of the codon models which are developed and widely used today still incorporate an assumption of effectively neutral synonymous substitution rate, constant between sites of each gene. Here we propose a simple yet effective extension to codon models, which incorporates codon substation rate variation along the gene sequence. We assess the performance of our approach in simulations and on real data. We find strong effects of nucleotide rate variation on positive selection inference, both under models with variation of protein selection and with branch-site variation of protein selection. We also demonstrate that the computational load of our approach remains tractable, and therefore we are able to apply it to genome scale positive selection scans. We apply our new method to two datasets: 767 vertebrate orthologs and 8,606 orthologs from twelve Drosophila species. We demonstrate that our new model is strongly favored by the data, and the support of the model increases with the amount of information. Moreover, it is able to capture signatures of nucleotide level selection acting on translation initiation and on splicing sites within the coding region. Finally, we show that rate variation is highest in the highly recombining regions, and we hypothesize that recombination and mutation rate variation, such as high CpG mutation rate, are the two main sources of nucleotide rate variation. Overall, nucleotide rate variation in substitutions is an important feature to capture, both to detect positive selection and to understand gene evolution, and the approach that we propose allows to do this in genome-wide scans.

[*] Corresponding author, marc.robinson-rechavi@unil.ch

I. Introduction

Detecting the selective pressure affecting protein coding genes is an important component of molecular evolution and evolutionary genomics. Codon models are one of the main tools used to infer selection on protein coding genes (Koonin and Wolf, 2010). This is done by comparing the rate of nonsynonymous substitutions (d_N) that are changing the amino-acid sequence with the rate of synonymous substitutions (d_S) that are supposed not to affect the gene function.

While there is overwhelming evidence of negative and positive selection acting on the amino acid sequence of the proteins (Boyko et al., 2008), the synonymous substitutions affecting the protein coding genes are assumed to be effectively neutral. This is a reasonable first approximation, especially for species with low effective population size, such as many mammals (Keightley et al., 2005; Romiguier et al., 2014). Therefore the synonymous substitution rate can be used as a proxy for the neutral substitution rate, and comparison between d_N and d_S can be used to identify the selection acting on the level of amino acids (Yang and Bielawski, 2000).

Two similar approaches to model the evolution of codons have been proposed by Goldman and Yang (1994) and Muse and Gaut (1994). In the Goldman and Yang (1994) model, selection pressure on the protein sequence is represented by a single parameter (ω), which defines the ratio of nonsynonymous to synonymous substitutions (d_N/d_S). In the Muse and Gaut (1994) model, both d_N and d_S are estimated as two independent parameters called α (d_S) and β (d_N). Neither of these approaches makes a molecular clock assumption, i.e. overall substitution rates are free to vary between the branches of a phylogenetic tree. On the other hand, both approaches originally incorporated an assumption of constant rates or ratio of nonsynonymous and synonymous substitutions between sites over the gene sequence. Moreover, they originally incorporated an assumption of constant ratio between branches. Various extensions have been proposed over the years to those models, allowing the incorporation of variation of selection acting on the protein sequence between codons (Yang et al., 2000) and/or between phylogenetic branches (Zhang et al., 2005). Yet for almost all of those new models an assumption of effectively neutral synonymous substitution rate, which is constant between sites, was kept. This leads to an assumption that all variability along the sequence is due to variability of the selection at the protein level. For the rest of this paper, we will use the term "uniform rate" to denote the assumption of constant nucleotide rates between sites within a gene, whether there is variation in rates or in selection between branches or not.

There is no biological reason to assume that the rate of synonymous substitutions is uniform in this way. In practice, it has been suggested (Yang, 2014) that the effect of the variation in the rate of synonymous substitutions should not substantially bias the inference of selection strength. The idea is that since the comparison between d_N and d_S is a contrast between the rates before and after the action of selection on the protein coding gene, while all the other factors will affect both rates in the same way. This reasoning might work when assigning a single ω value to the whole alignment, and the estimated value is an average. But for more sophisticated models, where ω varies between branches and sites, violation of the assumption of uniformity of the synonymous rate can affect the model performance (Rubinstein et al., 2011).

There are numerous sources of variation in the rate of synonymous substitutions inside genes. First, the neutral mutation rate across each genome significantly varies. One of the strongest effects on the mutation rate in mammals are CpG sites. Transitions at CpGs are more that 10-fold more likely than transitions at non-CpG sites (Leffler et al., 2013) due to spontaneous deamination, which causes a mutation from C to T, or from G to A. Both mutation frequency and repair efficiency are highly dependent on the context. E.g., the mammalian

CpG mutation rate is lower in high GC regions (Fryxell and Zuckerkandl, 2000). This is probably related to strand separation and hydrogen bonding in the neighboring region (Segurel et al., 2014). High GC regions themselves are characterised by a higher mutation rate, which is probably caused by less efficient repair by the exonuclease domain. There are other context-dependent effects which are known, many of which lack a mechanistic explanation, such as a higher mutation rate away from T with an increasing number of flanking purines (Hwang and Green, 2004) (for reviews see Hodgkinson and Eyre-Walker (2011) and Segurel et al. (2014))

Mutation rate is also affected by replication time: it is usually higher in the late-replicating regions (Stamatoyannopoulos et al., 2009). This effect is caused by the variation in the efficiency of mismatch repair (Supek and Lehner, 2015). It is not clear that this affects variation within genes, as opposed to between genes, but it could affect very long genes.

Mutation rates are also correlated with recombination rates. Some suggest (Lercher and Hurst, 2002; Hellmann et al., 2003, 2008) that recombination itself can have a mutagenic effect, possibly through an interaction with indels. Alternatively this correlation can be a result of GC-biased gene conversion (GC-BGC), whereby mutations increasing GC content have a higher chance of fixation in the population (Duret and Galtier, 2009). While GC-BGC is a fixation bias, in some cases it can create a pattern which is hard to distinguish from positive selection (Ratnakumar et al., 2010).

Finally, the synonymous substitution rate can be affected by selection on the nucleotide level. First, while synonymous substitutions do not affect the protein sequence, they might affect translation efficiency. This effect is not limited to species with large effective population size, such as Drosophila (Carlini and Stephan, 2003), since selection for codon usage was identified even in Homo sapiens (Comeron, 2004) and other mammals, especially for highly expressed genes. It has been suggested that bias in codon usage reflects the tRNAs abundance, and thereby provides a fitness advantage through increased translation efficiency/accuracy of protein synthesis (Bulmer, 1991), although in many cases there is no dependency between tRNA abundance and codon frequency, and the source of the bias remains unknown (Plotkin and Kudla, 2011).

Selection on the nucleotide sequence can be also caused by secondary structure avoidance, as secondary structure can reduce translation efficiency (Kudla et al., 2009; Kertesz et al., 2010). Other potentially important sources of selection on the nucleotide sequence, independent of the coding frame, include splicing motifs located within exons, exon-splicing enhancers (Majewski and Ott, 2002), or functional non-coding RNAs, such miRNAs or siRNAs, which often reside within coding sequences (Mattick and Makunin, 2006).

Because of all these mutational and selective biases, it is important to model rate variation not only for protein selection, but also at the nucleotide level. There are in principle two different approaches to incorporate rate variation into codon models. First, it is possible to model synonymous and non-synonymous rates separately extending a two-rate model, as in Pond and Muse (2005). Second, it is possible to incorporate site-specific rates as an independent parameter into one-rate models (Scheffler et al., 2006; Rubinstein et al., 2011). In the second case, the rate parameter captures biological factors, such as mutation rates, fixation rates, or nucleotide selection, which act on all substitutions, both synonymous and nonsynonymous.

Here we focus on the second approach, due to its more straightforward selection parameter (ω) and interpretation, and to its superior performance for d_N/d_S estimation (Spielman et al., 2016).

While codon models accounting for nucleotide rate variation are available for more than a decade, they are still rarely used for large-scale selection analyses, such as Kosiol et al. (2008); Moretti et al. (2014); Zhang et al. (2014). This is probably because these models have even higher computational demands, and the performance of different approaches to nucleotide rate variation was never compared.

Here we extend the Scheffler et al. (2006) model, which captures variation between codons, i.e. uses a single rate per codon, and perform a direct comparison with Rubinstein et al. (2011), which captures variation between nucleotides, i.e. with three rate parameters per codon. We also assess the impact of nucleotide rate variation on the BS-REL-family model (Murrell et al., 2015). We chose Murrell et al. (2015) as a comparison, since it is the only BS-REL model for gene-wide identification of positive selection, while other positive selection models in that family are intended for inference of selection on individual sites.

We first use simulations to compare different approaches of modeling synonymous rate variation. Then we use our model to detect positive selection in twelve Drosophila species and in a vertebrate dataset.

We detect positive selection on genes from those two datasets under our new model, and we demonstrate that it is important to take rate variation into account for such positive selection inference. We investigate factors affecting the nucleotide substitution rate, and we show that the new model successfully detects synonymous selection acting on regulatory sequences within the coding sequence. We also identify which gene features most affect rate heterogeneity.

II. New approaches

We model the process of codon substitution as a Markov process defined by the instantaneous rate matrix Q. In a general case the Q can be written as (Rubinstein et al., 2011):

$$q_{ij} = \begin{cases} \rho^{(h)}\lambda_{ij}\pi_j & \text{i and j differ by one synonymous substitution at site h} \\ \rho^{(h)}\lambda_{ij}\omega\pi_j & \text{i and j differ by one nonsynonymous substitutio at site h} \\ 0 & \text{i and j differ by more than one nucleotide} \end{cases}$$

where $\rho^{(h)}$ is the substitution rate for a site h, and λ_{ij} is the substitution factor to change from nucleotide i to nucleotide j, which is typically used to account for the difference between transitions and transversion rates (Hasegawa et al., 1985). Here the rate $\rho^{(h)}$ is used to account for various effects that are not captured by the variation in ω; in particular it accounts for variation in mutation rate and selection acting on the nucleotide sequence.

In Rubinstein et al. (2011), $\rho^{(h)}$ is modeled using a one parameter gamma distribution across sites of the alignment, such that the mean substitution rate is equal to 1, i.e. $\rho^{(h)} \sim Gamma(\alpha, 1/\alpha)$. Keeping a mean rate of 1 is important to avoid biases in the estimation of branch lengths. There is no implicit assignment of rates to sites, as in the CAT model (Lartillot and Philippe, 2004). Instead, a random-effects model is used: the gamma distribution is split into equally probable discrete categories using quantiles, and the site likelihood is computed as the average of the likelihoods for each possible rate assignment. This approach adds only one extra parameter to the model, but it is computationally intensive, since for k discrete categories, k^3 likelihoods have to be computed per site.

In Scheffler et al. (2006), unlike Rubinstein et al. (2011), the three positions of each codon have the same rates. Here a codon belongs to one of three categories, each one represented by a single rate value. The rates and their respective proportions are estimated from the data, which leads to the estimation of four different parameters (two rates, as the third one is fixed by the constraint that the mean rate $\bar{\rho} = 1$, and two proportions). This approach is virtually equivalent to adding a branch length multiplier for certain site classes, and therefore likelihood can be computed efficiently.

Here we propose having one rate per codon, while allowing this rate to vary following the gamma distribution, $\rho_k \sim Gamma(\alpha, 1/\alpha)$.

With this approach we are combining the strengths of both Rubinstein et al. (2011) and Scheffler et al. (2006): we only increase parameter space by one parameter (α), allow a flexible distribution of rates, and keep the computational tractability of the model. Moreover, having similar parametrization and rate distributions allows us to compare the statistical performance of site rate and codon rate variation models.

Using the proposed approach we extended two widely used codon models: M8 (Yang et al., 2000) and the branch-site model (Zhang et al., 2005). In principle our approach could be applied to any GY94-based model. These models were implemented in `Godon`, a codon model optimizer in Go, in three variants: no rate variation, site rate variation (Rubinstein et al., 2011) and codon rate variation as described above.

Three models out of six were implemented and used for the first time to our knowledge: branch-site model with the site rate variation similar to Rubinstein et al. (2011) and M8 and branch-site models with gamma distributed codon rate variation, as proposed above.

III. Results

I. Simulations

Site models

We have simulated three datasets using various flavours of the M8 model: a dataset without rate variation, a dataset with site rate variation, and a dataset with codon rate variation (Table 1). We then used three corresponding models to infer positive selection in those datasets. In all three cases, as expected, the model corresponding to the simulations shows the best result (Fig. 1, Table 2 and Supplementary Table S1).

In the absence of rate variation, the statistical performance of the three methods is very similar even though the M8 model without rate variation has a slightly better performance (Fig. 1A, Supplementary Fig. S1A). With the dataset with site rate variation (Fig. 1B, Supplementary Fig. S1B), there is a large underperformance when not accounting for variation. On the other hand, codon variation performs almost as well as site variation, and clearly better than the model with no variation. With the dataset with codon rate variation (Fig. 1C, Supplementary Fig. S1C), there is a relatively large decrease in the performance of both other models, while, as expected, accounting for site rate variation performs better compared to the model without rate variation. False positive rates exceed the significance levels when rate variation is not taken into account (Supplementary Fig. S1C). Codon rate variation even increases the false positive rate above 50% for the model without rate variation at the significance level of 0.05. Stronger rate variation (i.e. smaller α value) causes a higher false positive rate (Supplementary Fig. S2).

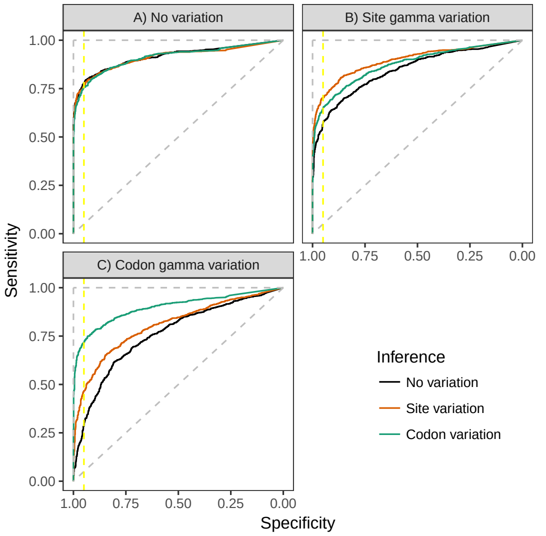

Figure 1: *Performance of three M8-based models (M8 with no rate variation, M8 with site rate variation and M8 with codon rate variation) on datasets A) without rate variation, B) with site rate variation, and C) with codon rate variation. The yellow dashed line indicates the 0.95 specificity threshold (i.e. false positive rate of 0.05). The dashed diagonal line shows theoretical performance of the random predictor, the dashed vertical and horizontal lines indicate theoretical performance of the perfect predictor.*

From this we can conclude that a) the performance of models accounting for codon variation is acceptable in all three scenarios, i.e.

		Estimation						
		M8			Branch-site			BUSTED
Simulations		No var.	Site var.	Codon var.	No var.	Site var.	Codon var.	
M8	No variation	•	•	•				•
	Site variation	•	•	•				•
	Codon variation	•	•	•				•
BS	No variation				•	•	•	•
	Site variation				•	•	•	•
	Codon variation				•	•	•	•

Table 1: *Summary of estimations performed on the simulated datasets. M8: M8 model of Yang et al. (2000); BS: branch-site model of Zhang et al. (2005); BUSTED: BUSTED model from the BS-REL-family (Murrell et al., 2015).*

	Simulation		
Estimation	No var.	Site var.	Codon var.
No var.	0.916/ 100%	0.846/94.4%	0.758/84.5%
Site var.	0.912/99.6%	0.897/ 100%	0.806/89.7%
Codon var.	0.912/99.6%	0.875/97.6%	0.898/ 100%

Table 2: *Area under curve (AUC) for all M8-based simulations (see Fig. 1). Second number computed as proportion of maximum AUC for a particular simulation.*

no rate variation, site rate variation and codon rate variation; b) in the presence of codon rate variation in the data, models not accounting for this kind of variation suffer from a notable loss of statistical performance.

The total inference computation time was 76 CPU hours for the model with no rate variation, 232 CPU hours for codon rate variation, and 4,023 CPU hours for site rate variation.

Branch-site models

The simulations based on the branch-site model show a qualitatively similar behaviour to the simulations based on the M8-type models (Fig. 2 and Supplementary Fig. S3, Supplementary Tables S2, S3), although the performances are more similar between models. As with M8-type models, codon rate variation models performs well in all three cases, while simulating with codon rate variation causes a clear underperformance in both other models. Unlike in the case of M8-type models, false positive rates are only marginally inflated compared to theoretical expectations (Supplementary Fig. S4). Nevertheless, the model with codon rate variation shows the best performance.

The total inference computation time was 37 CPU hours for the model with no rate variation, 101 CPU hours for codon rate variation, and 1,630 CPU hours for site rate variation.

More complex models have a computational cost. Analyses with codon rate variation were 3.0 and 2.8 times slower compared to no rate variation for M8 and branch-site models respectively, while those with site rate variation were 52.7 and 44.3 times slower respectively. Thus codon rate variation captures biological signal at a much lower computational cost than site rate variation.

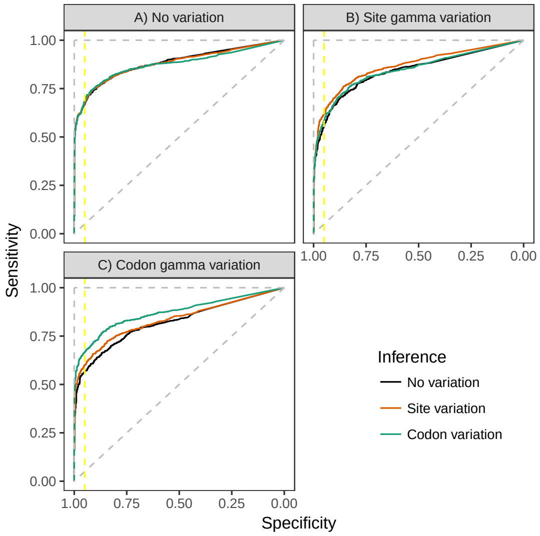

Figure 2: *Performance of three branch-site-based models (branch-site with no rate variation, branch-site with site rate variation and branch-site with codon rate variation) on datasets A) without rate variation, B) with site rate variation and C) with codon rate variation. The yellow dashed line indicates the 0.95 specificity threshold (i.e. false positive rate of 0.05). The dashed diagonal line shows theoretical performance of the random predictor, the dashed vertical and horizontal lines indicate theoretical performance of the perfect predictor.*

Comparisons with BS-REL

It was demonstrated (Murrell et al., 2015) that in certain cases the statistical power of BS-REL is superior to other methods, therefore it is important to study how rate variation affects the performance of those models. The only BS-REL model suitable for the gene-wide identification of positive selection is BUSTED (Murrell et al., 2015), and the current implementation supports neither rate variation nor d_S variation as implemented in Pond and Muse (2005).

Because of differences in the simulations and model assumptions we did not compare model accuracy and ROC, but we focused instead on the effect of the rate variation on the false positive rate. In other words we are asking whether the unaccounted rate variations in the evolutionary process can cause false positives in positive selection inference.

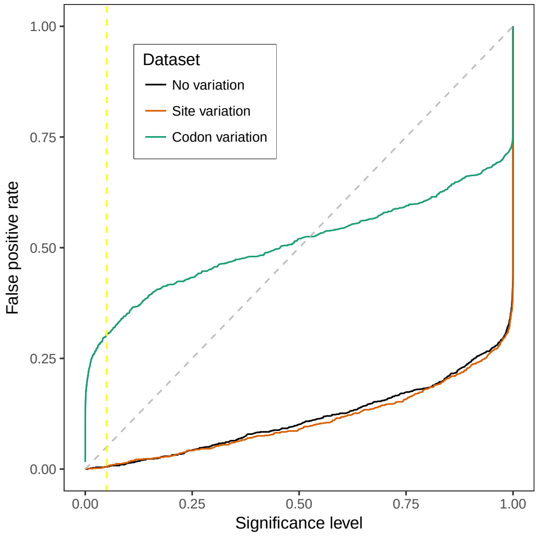

Figure 3: *False positive rate as a function of significance level for data simulated with the M8 model. A vertical line indicated typical significance level equal to 0.05. A diagonal dashed line corresponds to the identity line, significance level equal to false positive rate. Three line colors indicate three simulated datasets (see legend).*

BUSTED shows highly inflated rates of false positives in the presence of codon rate variation (Fig. 3, Supplementary Fig. S5); at a typical significance level of 0.05 the false positive rate of BUSTED is close to 0.3 and 0.2 for the M8 and branch-site simulations, respectively.

II. Vertebrate dataset

Given the good performance of our model in the simulations, we applied it to the real data. First we used 767 one-to-one orthologs from vertebrate species. This represents a set of genes with high divergence (more than 450 My), conservative evolution (Studer et al., 2008), and relatively low effective population sizes (although some vertebrates have high N_e, see Gossmann et al. (2012)), thus relatively weak impact of natural selection. We analyzed

A	Codon variation	
No variation	−	+
−	6,935	144
+	790	1,038

B	Codon variation	
No variation	−	+
−	7,022	57
+	486	1,342

Table 3: *Positive selection predictions with and without rate variation for the vertebrate dataset; A) codon rate variation, B) site rate variation.*

them with three variants of the branch-site model: no rate variation, site rate variation and codon rate variation.

We observed that in most of the cases (a branch of a gene tree) the data supports (Akaike information criterion, AIC) the codon rate variation model: out of 26,721 individual branches tested, data supports codon rate variation in 85% of the tests, site rate variation in 15%, and no rate variation model was favored only in a single test (0.01%).

A large proportion of branches detected to be under positive selection with the no rate variation model are not detected to be under positive selection with the codon rate variation model (Table 3). More than 40% of the positive predictions from the standard branch-site model are not supported when codon rate variation is accounted for. This suggests that evolution on these branches can be explained by nucleotide substitution rate variation without positive selection.

Supplementary Table S4 shows prediction agreement between each model and the best supported model out of three, confirming the good performance of the codon variation model.

The branch-site model analysis took 9, 76 and 1,732 CPU hours for no variation, codon rate variation and site rate variation models. Thus codon rate variation model was 8.4 times slower, while site rate variation model was 190 times slower.

With real data, differences between genes are not only stochastic, but are expected to be driven by underlying biological differences. It is thus interesting to find which factors affect rate variation as estimated by the model, as well as to know which genes favoured the model with codon rate variation the most. In order to perform this analysis we averaged parameters obtained by testing different branches of the tree.

The relative support of the model with codon rate variation is mostly affected by total branch length, alignment length, and mean GC content of the gene (Supplementary Table S5). The positive correlation with tree length and alignment lengths is probably related to the increase in total amount of information available for the model. The relation to average GC content might be due to the relationships between recombination rates, substitution rates, and GC content (Duret and Galtier, 2009; Rudolph et al., 2016).

For the shape parameter of the gamma distribution α, the strongest explanatory variable is also the length of the alignment (Supplementary Table S6). Counterintuitively, shorter alignments are characterized by larger rate variation (low α value corresponds to the large gamma distribution variance). We also observe a weak relation with maximal expression level. Highly expressed genes tend to have a higher rate variation, which could be explained by higher nucleotide level selection on certain parts of the gene.

III. Drosophila dataset

Second, we used 8,606 one-to-one orthologs from Drosophila genomes. The Drosophila data set is ten-fold larger than the vertebrate dataset. Since analyses on the simulated and vertebrate datasets show consistent superiority of codon rate over site variation rate, with a much lower computational cost, we ran only codon variation and not site variation model on this dataset. Therefore for the Drosophila data we are comparing models with and without codon rate variation. Drosophila have large effective population sizes on average (Gossmann et al., 2012), thus stronger impact of natural

	Codon variation	
No variation	−	+
−	55,717	301
+	5,203	5,166

Table 4: *Positive selection predictions for the Drosophila dataset with and without rate variation.*

selection; the genes studied are less biased towards core functions than in the vertebrate dataset, and have lower divergence: about 50 mya for Drosophila (Russo et al., 2013) compared to more than 450 mya for the vertebrate dataset (Betancur-R et al., 2015).

In total 66,387 branches were tested for positive selection. The model with codon rate variation was supported by the data using AIC in 97% of the tests. As with the vertebrate dataset, predictions were not consistent between the two models (Table 4). In this case the majority of predictions of positive selection given by the model without rate variation are not supported by the model accounting for rate variation.

The slowdown of branch-site model with codon rate variation was 9.3 times and it took about 10,200 CPU hours.

The relative support of the model with codon rate variation is mainly explained by average GC content, alignment length and tree length (Table 5), which is consistent with the vertebrate results (Supplementary Table S5). We also see a correlation with the number of sequences and alignment and sequence lengths, which also represent the amount of information available (higher coding sequence length means less gaps for the same alignment length).

We also observe a dependence on the number of exons and on recombination rate. A larger number of exons implies more exon-intron junctions, which might affect variation in levels of nucleotide sequence selection (see below). And recombination might affect GC-BGC, mutation rate, and selection strength acting on synonymous sites (Campos et al., 2014).

The rate variation parameter α can be explained by several features of genes (Supplementary Table S7). Most of the effects are not reproduced between the two datasets. Counterintuitively the direction of dependence is reversed for the number of sequences, but this dependency is not very strong and not very significant. The strongest and the most consistent effect between the two datasets is dependence of the rate variation on GC content (smaller α implies higher variance of gamma distribution, hence higher rate variation).

IV. Signatures of selection at the nucleotide level

Codon rate variation can be influenced by various factors such as mutation bias, fixation bias (e.g., gene conversion), or selection acting against synonymous substitutions. Notably, it is well known that exon regions adjacent to the splicing sites are evolving under purifying selection at the nucleotide level (e.g. see Majewski and Ott (2002)). We determined posterior rates for positions of protein coding gene regions located in the proximity of exon-intron and intron-exon junctions; first exons were excluded from the analysis.

We observe in Drosophila (Fig. 4) that our codon rate variation model captures these selection constraints: the codon rate is lower at the exon-intron junction than at the intron-exon junction, and both have lower rates than the rest of the exon. This is in agreement with splicing motif conservation scores (e.g. see Cartegni et al. (2002)), and consistent with negative selection acting on spicing sites.

We used the M8 with codon rate variation to simultaneously estimate the effect of factors which affect substitution rates of nucleotide and protein sequences, again in Drosophila. We observed that the model is able to recover opposing trends acting on the 5' region of the protein coding gene (Fig. 5). These trends are probably a results of the high functional importance of the 5' nucleotide sequence, but low functional importance of the corresponding amino acid sequence (see discussion). We observe that the top 25% most highly expressed genes show both stronger conservation of the

Variable	Estimate	Std. Error	t value	p-value
Number of sequences	0.341405	0.013961	24.455	$< 2 \cdot 10^{-16}$
Total branch length	0.574623	0.056136	10.236	$< 2 \cdot 10^{-16}$
Alignment length	1.552349	0.076787	20.216	$< 2 \cdot 10^{-16}$
Length of coding sequence	0.554160	0.079118	7.004	$2.69 \cdot 10^{-12}$
GC content (mean)	5.671729	0.492706	11.511	$< 2 \cdot 10^{-16}$
GC content (stdev)	1.373463	1.813904	0.757	0.44896
Total intron length	-0.028203	0.009466	-2.979	0.00290
Number of exons	0.426161	0.039296	10.845	$< 2 \cdot 10^{-16}$
Maximum expression	0.108736	0.033415	3.254	0.00114
Mean expression	-0.088538	0.033832	-2.617	0.00889
Recombination rate	0.243515	0.024765	9.833	$< 2 \cdot 10^{-16}$

Table 5: *Linear model of relative support of model with the codon rate variation; Drosophila dataset. Significant variables (p-value < 0.05) in bold. Model p-value is $< 2.2 \cdot 10^{-16}$, multiple R^2 is 0.6134.*

amino acid sequences (Supplementary Fig. S6) and more pronounced decrease in the substitution rate of the 5'-region (Supplementary Fig. S7). The relation with expression levels is consistent with the assumption that we are measuring natural selection on gene sequences, rather than mutation rates.

IV. Discussion

I. Nucleotide level selection in coding regions

There is strong evidence of selection acting on synonymous substitutions within protein coding sequences, and the strength of this selection is expected to vary across the coding region (Chamary et al., 2006).

In particular, negative selection strongly affects regulatory sequences, such as exonic splicing enhancers or exon junction regulatory sequences (Cartegni et al., 2002). This variation in the selection strength affecting both synonymous and nonsynonymous substitutions can affect the performance of codon models (Rubinstein et al., 2011) and it is essential to take it into account. While there are multiple ways to do this, for instance by modeling the synonymous and non-synonymous rates separately (Pond and Muse, 2005), here we focused on modeling the ratio of non-synonymous vs synonymous rates as a single parameter (ω), while allowing the substitution rate (ρ) to vary along the sequence.

Our approach succeeds in recovering a signal of splicing motif conservation jointly with positive selection acting on the gene sequence.

We also demonstrate that our model is able to disentangle opposite trends acting on the same sequence, i.e. stronger negative selection acting on the nucleotide sequence combined with weaker amino-acid selection towards the beginning of the reading frame.

Selection on the 5' nucleotide sequence is probably due to selection for translation initiation efficiency (Bentele et al., 2013), and is probably related to suppression of mRNA structures at the ribosome binding site. At the same time N-terminal amino acids are more likely to be unstructured, and they are relatively less important to protein function and stability compared to the core (Guharoy and Chakrabarti, 2005).

II. Determinants of rate variation

The vast majority of gene alignments in the study indicated better support for the model with codon rate variation. Moreover, the relative probability of the models incorporating codon rate variation increases with the amount

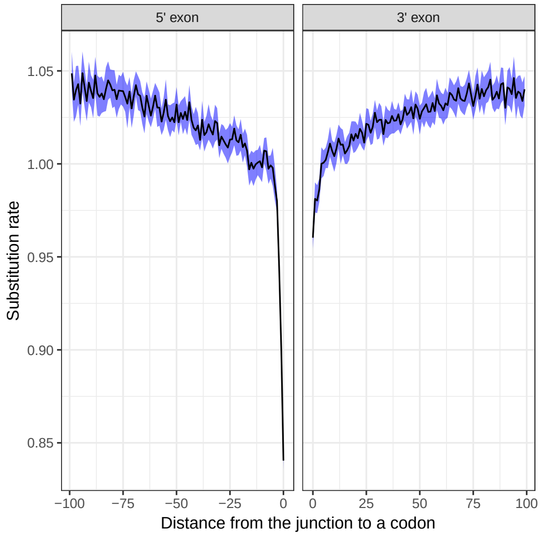

Figure 4: *Codon rate as a function of proximity to the exon-intron and intron-exon junction in the Drosophila dataset. The left panel depicts rates in 5' exon (prior to the exon-intron junction, negative distances), while the right panel depicts 5' exon (rates after the intron-exon junction, positive distances). A rate of 1 corresponds to the average rate of substitution over the gene; thus values above 1 do not indicate positive selection, but simply a rate higher than average for this gene. The blue ribbon indicates 98% confidence interval of mean estimate. Only alignment positions with more than 70% of the sequences defined were used in the plot.*

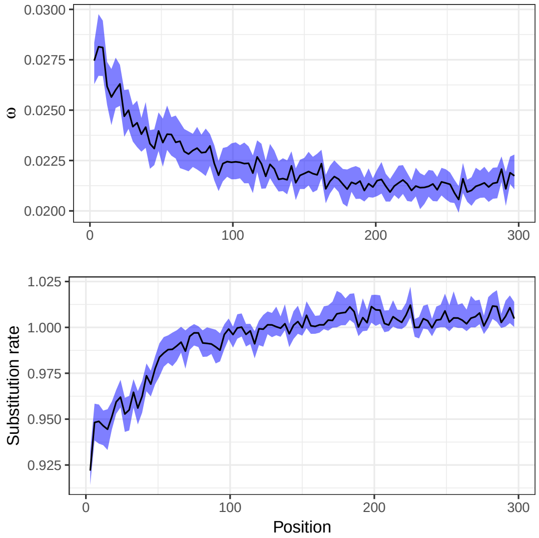

Figure 5: *Posterior estimates of median ω (d_N/d_S, top panel) and codon substitution rate ρ (bottom panel) as a function of distance from the start codon expressed in the number of nucleotides in Drosophila. The model M8 with codon rate variation was used to estimate both parameters simultaneously. Smaller values of ω (top panel) indicate stronger negative selection acting on the protein sequence. A substitution rate of 1 (bottom panel) corresponds to the average rate of substitution over the gene; thus values above 1 do not indicate positive selection, but simply a rate higher than average for this gene. The blue ribbons indicate 98% confidence intervals of median estimates. Start codons and alignment positions with less than three sequences were excluded from the plot.*

of information available, be it number of sequences, alignment length or total number of substitutions. This indicates that these models are better in describing the underlying evolutionary process, and if we have enough data these models are favored. We detect a strong signal of nucleotide variation in two quite different datasets: flies have high effective population size, thus natural selection is relatively strong, including on codon usage or splicing; and vertebrates have higher sequence divergence, which does not appear to mask the signal of nucleotide evolution, despite lower effective population sizes in many species (Gossmann et al., 2012; Romiguier et al., 2014). Thus the effect of nucleotide rate variation appears quite general, and will probably be found in many other species.

The strongest determinant of the relative support of the model with codon rate variation is GC content. It is well known (Fullerton et al., 2001; Marais et al., 2003; Chamary et al., 2006), that high-GC regions have higher recombination rates as a result of GC-BGC, notably in the species studied here. It has also been shown that models accounting for rate variation show significantly better performance in the presence of recombination (Scheffler et al., 2006), even if the true tree topology is used.

The effect of recombination rate as measured in Comeron et al. (2012) on the relative support of codon rate variation is weaker than the effect of GC content alone. In mammals the higher CpG dinucleotides mutation rates (Kong et al., 2012) can increase the substitution rate disparitiy and therefore contribute to the dependence of the GC content and the relative model support.

Yet we see a similar dependence of Drosophila, where there is no significant neighboring base contextual effects on the mutation rate (Keightley et al., 2009). Here we hypothesize that the GC content can be viewed as a proxy for the average recombination rate over time via GC-BGC. Considering the rapid evolution of recombination hotspots (Ptak et al., 2005), GC content is probably better capturing historical recombination rates, while the direct measurement of recombination rate captures only the current state.

In both datasets we observe a signficant positive association of rate variation with the maximal expression level. Pressure for translational robustness increases with expression levels (Drummond et al., 2005), and codon choice affects expression level (Bentele et al., 2013). One of the main causes of selection on the codon sequence of highly expressed genes is protein misfolding avoidance (Yang et al., 2010), but selection for efficient translation initiation is also a cause of selection (Pop et al., 2014). It is reasonable to assume that only certain parts of protein coding genes will be affected by strong nucleotide sequence selection, and that this selection will be stronger on more expressed genes. Indeed our results show strong negative selection acting on the coding sequence of translation initiation regions, and the relative selection strength is higher for the 25% highest expressed genes (Supplementary Fig. S7). This can lead to the overall increase in the substitution rate variation.

Given the stronger negative selection on the coding sequence and the stronger variation in the substitution rate in highly expressed genes, it is especially important to take this variation into account. Having a few quickly evolving codons in combination with a low average ω of highly expressed genes can be interpreted as positive selection by models without rate variation. Indeed, in Drosophila the rate of false positives, i.e. genes identified to evolve under positive selection only by the model without rate variation, is strongly correlated with the maximum expression levels of the gene (Supplementary Table S8).

III. Codon models and rate variation

Widely used mechanistic codon models rely on the assumption of constant synonymous substitution rates. This assumption is often violated due to factors such as mutation bias or nucleotide selection, which vary across the gene. While substitution rate variation can be caused by multiple factors, we use a single compound rate parameter to model this variation.

Here we demonstrate that a simple model captures such rate variation, and that it both detects new biological signal, and substantially decreases the false positive rate in positive selection detection. Not only do we observe this effect in simulations (Fig. 1, 2, 3), but inconsistency between models is even higher when applied to the vertebrate and fly datasets. Up to 50% of the positive selection predictions performed using models without rate variation can be explained by the nucleotide rate variation (Tables 3, 4), and thus can be considered as probable false positives.

This indicates that the underlying biological process is highly variable across positions, and that a model selection procedure is able to capture this once enough information is available.

An important question is why accounting for rate variation changes the statistical properties of the test. Indeed it has been argued (Yang, 2014) that comparison between d_N and d_S is a contrast between the rates before and after the action of selection on the protein, and should not be biased by nucleotide rate variation. We hypothesise that d_N/d_S overestimation is caused not only by the variation in d_S, but also by codon-specific substitution rates. Indeed, having a small percentage of rapidly evolving codons in the gene would not be captured by an overall rate for d_S, and therefore would be interpreted as positive selection by models with protein level but without nucleotide level rate variation. Whereas fully accounting for rate variation allows to detect these codons as rapidly evolving by the signatures of both synonymous and nonsynonymous substitutions.

There is recent evidence that double mutations in coding sequences increase the branch-site model false positive rate from 1.1% to 8.6% in similar datasets to those investigated here (Venkat et al., 2017). The interaction between this effect and rate variation along the gene is worth investigating.

We compared two different models accounting for rate variation: the site variation model of Rubinstein et al. (2011) and our new codon variation model which extends Scheffler et al. (2006). The codon rate variation model can be informally thought of as a special case of the site rate variation model. Despite that, the codon rate variation performs better both in the simulations (Table 2, Supplementary Table S2) and on the vertebrate dataset (Supplementary Table S4). There are probably two reasons for that. First, the fact that we can assign a rate to a particular nucleotide position does not necessary mean that we can reliably estimate it. Only two amino acids allow single nucleotide synonymous substitution associated with the first or second codon positions. This means that individual position rates can be estimated mostly through non-synonymous substitutions, which are typically rare compared to synonymous ones. Moreover, branch-site and M8 models allow variation in the nonsynonymous rate over codon positions, which means estimates of ω and site rates are not independent.

Secondly, we expect site rates to be autocorrelated along the sequence since many factors, such as GC content, recombination rate, or chromatin state change slowly over the gene. Indeed we see a weak signal of such autocorrelation in our data (not shown). Therefore having an independent rate for every site is probably redundant.

One of the key advantages of codon variation relative to site variation is computational performance. Having a distinct rate for every position increases the number of site classes for which likelihood computations have to be performed by a factor of k^3, where k is the number of discrete categories for gamma distribution. Whereas having a rate only for each codon increases the number of site classes by a factor of k. Which means that even for four discrete categories, the slowdown of likelihood computation for site rate gamma model will be about 64 times, vs. only 4 times for codon rate variation model. In practice this ratio between the two models was respected in simulated and vertebrate data. This makes codon rate variation model usable in large real-world datasets, as we demonstrate including on the large 12 Drosophila genomes set.

Unlike traditional mechanistic codon models, our new models allow independent estimations of substitution rate at the nucleotide level and of selective pressure on amino acid sequences. It should be noted that individual site rates estimates may be still noisy because of the amount of data available. But given enough data it is possible to have accurate estimates of selection acting on specific regions, e.g. splicing motifs, within coding sequences (Fig. 4).

V. Conclusions

We present here a new codon rate variation model family. These mechanistic codon models relax an unrealistic assumption that the only source of substitution rate variation over the gene sequence is selection on the protein. Failure to account for this leads to both type I and type II errors. We demonstrate that our model has a good statistical performance both in the presence and in the absence of rate variation. Rate variation is strongly supported by homologous genes both from species with larger (flies) and smaller (vertebrates) effective population sizes. We are able to capture differences in substitution rate caused by nucleotide selection. Importantly, while being more complex these model remain computationally tractable and therefore can be applied to large-scale datasets. These models open the opportunity of simultaneous analysis of different layers of selection.

VI. Methods

I. Sequence simulations

We simulated six datasets (Table 1) that include either no rate variation across sites (corresponding to the GY94 model), variation between sites (corresponding to the Rubinstein et al. (2011) model) and variation between codons (corresponding to our new approach). Each dataset contains 1,000 alignments simulated under the null hypothesis H0 with no positive selection (all $\omega \leq 1$) as well as 1,000 alignments under the alternative hypothesis H1 with positive selection (some $\omega > 1$). All the datasets had between 8 and 12 sequences composed of sequences of 100 to 400 codons and were simulated using the software `cosim`. The parameters of each simulation, including the alignment length and the number of species, were generated at random from their respective distribution (Supplementary Table S9, Supplementary Fig. S8). Values of α were within the range of values estimated from the real data (Supplementary Fig. S9), with an emphasis on smaller values where the variation is stronger. For the simulations including rate variation, we used four discrete gamma categories that we assigned either to sites or to codons. The M8 model assumes that the neutral sites and those under purifying selection have an ω drawn from a beta distribution and we represented this distribution using five discrete categories. Finally, to simulate evolution under the branch-site model, we randomly selected one 'foreground' branch of the phylogenetic tree (either internal or terminal) for every simulated alignment.

II. Vertebrate and Drosophila datasets

We analyzed two biological datasets. Our goals were to compare the fit of the different models on real data, and to study which gene features are contributing to the variation of the substitution rate. First, we used a vertebrate one-to-one orthologs dataset (Studer et al. (2008), available at `http://bioinfo.unil.ch/supdata/positiveselection/Singleton.html`) consisting of 767 genes (singleton dataset). This dataset was already used in previous studies of codon models (Fletcher and Yang, 2010; Gharib and Robinson-Rechavi, 2013; Davydov et al., 2017).

We also used a subset of one-to-one orthologs from 12 Drosophila species from the Selectome database (release 6, `http://selectome.unil.ch/`). This dataset consists of 8,606 genes, and the alignments are filtered to remove unreliably aligned codons (Moretti et al., 2014).

III. Positive selection inference

For all the tests on simulated data we used the correct (i.e. simulated) tree topology, but starting branch lengths were estimated using PhyML v. 20131022 (Guindon et al., 2010) with the model HKY85 (Hasegawa et al., 1985). We did not start the optimization from the true branch lengths, by similarity to a real use-case, when only gene sequences are available, and the true branch lengths are unknown. While tree topology is also inferred in real use-cases, and wrong topology could impact the inference of positive selection (Diekmann and Pereira-Leal, 2015), investigating this is outside the scope of our study.

Optimization of all model parameters jointly with branch lengths is not practical and substantially increases the computational load. We instead first estimated branch lengths using the simpler M0 model, which assumes a constant ω across branches and sites, and optimized in a second step the model parameters of the M8 or branch-site models with or without rate variation, while fixing branch lengths. A similar approach was used in previous studies (Scheffler et al., 2006; Moretti et al., 2014).

We show that this approach at least in the case of the absence of variation does not decrease significantly the statistical properties of the positive selection inference (Supplementary Fig. S10).

The model optimization was performed in `Godon`, followed by model selection (see below).

For BUSTED we used an implementation available in HyPhy v. 2.2.6 (Pond et al., 2005).

For the biological datasets, all the internal branches were tested using the branch-site model for positive selection. Tip branches were not tested to reduce the potential effect of sequencing errors. The M8 model was also applied to estimate substitution rates and ω for individual sites.

IV. Model selection

During model selection we had six model to choose from: three rate variation approaches and, for each, the absence or presence of positive selection. Although LRT can be used to test for positive selection, it is not possible to use it to compare across all six models that we tested (i.e. any pair of codon rate variation and site rate variation models cannot be represented as a nested pair).

We thus first used the AIC on the alternative model to select one of the three approaches to model rate variation: no rate variation, site rate variation or codon rate variation.

The relative support of each model was computed as a log ratio between Akaike weights (Wagenmakers and Farrell, 2004) of the model with codon rate variation and the model without rate variation.

Once the rate variation model was selected, we performed LRT to detect positive selection on the corresponding pair of models, i.e. model with $\omega \leq 1$ and model without this constraint. A 50:50 mix of a χ^2 distribution with one degree of freedom and of a point mass of 0 was used as a null distribution (Yang and dos Reis, 2011).

V. Posterior rates inference

In order to estimate rates of synonymous substitution for individual codons we used an approach similar to Rubinstein et al. (2011). First, we estimated the probability of a codon belonging to each rate as $P = Pr(\rho^{(h)} = \rho_i | x_h, \eta)$, where $\rho^{(h)}$ is the rate of codon h, ρ_i is the i-th discrete gamma rate, x_h is the data observed at codon h, and η are the parameters of the model (e.g. for M8 $\eta = \{p_0, p, q, s\}$). In this approach, η is replaced with the maximum likelihood estimate of model parameters $\hat{\eta}$. Thus codon rates can be estimated as a weighted sum $\hat{\rho}^{(h)} = \sum_i^k Pr(\rho^{(h)} = \rho_i | x_h, \hat{\eta}) \rho_i$.

An alternative would be to use Bayes empirical bayes (BEB, Yang et al. (2005)) instead. However BEB was developed and tested for site detection in particular codon models, and we do not know how well is it applicable to rate variation. On top of that given the increased parametric space of the model, BEB would be computationally intensive. Since we are averaging rates over multiple sites, random noise should not introduce a substantial bias.

Site d_N/d_S ratios in the M8 model can be estimated using a similar approach, while replacing codon rate categories with the ω categories.

Posterior site rate and d_N/d_S estimation is implemented in `Godon`. In all cases we used an alternative model codon rate estimation. Since the null model for every pair is a special case of the alternative, we can use the later for parameter estimation without any significant loss of precision.

For the branch-site model we averaged position rate estimates from all the individual branch tests.

VI. Regression analysis

To estimate dependencies between various parameters we used linear models (lm function, R version 3.3.2). Various parameters were transformed to correspond a bell-shaped if possible (see Supplementary Table S10, Supplementary Fig. S11).

We used expression data for *H. sapiens* from Fagerberg et al. (2014), acquired from Kryuchkova-Mostacci and Robinson-Rechavi (2015). For *D. melanogaster* we used data from Li et al. (2014), available at `http://www.stat.ucla.edu/~jingyi.li/software-and-data.html`. Recombination rates for genes were computed using Recombination Rate Calculator (ver. 2.3, Fiston-Lavier et al. (2010)) using dataset from Comeron et al. (2012).

VII. Availability

All the code is available from `https://bitbucket.org/Davydov/codon.rate.variation`. Sequence simulator `cosim` is available from `http://bitbucket.org/Davydov/cosim`. Codon model parameter estimator `Godon` is available from `https://bitbucket.org/Davydov/godon`.

VII. Acknowledgments

This work was supported by the Swiss National Science Foundation (grant number IZLRZ3_163872). The computations were performed at the Vital-IT (`http://www.vital-it.ch`) center for high-performance computing of the Swiss Institute of Bioinformatics.

References

Bentele, K. et al. Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.*, 9:675, Jun 2013.

Betancur-R, R., Orti, G. and Pyron, R.A. Fossil-based comparative analyses reveal ancient marine ancestry erased by extinction in ray-finned fishes. *Ecol. Lett.*, 18(5):441–450, May 2015.

Boyko, A.R. et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.*, 4(5):e1000083, May 2008.

Bulmer, M. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129 (3):897–907, Nov 1991.

Campos, J.L. et al. The relation between recombination rate and patterns of molecular evolution and variation in Drosophila melanogaster. *Mol. Biol. Evol.*, 31(4):1010–1028, Apr 2014.

Carlini, D.B. and Stephan, W. In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. *Genetics*, 163 (1):239–243, Jan 2003.

Cartegni, L., Chew, S.L. and Krainer, A.R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.*, 3(4):285–298, Apr 2002.

Chamary, J.V., Parmley, J.L. and Hurst, L.D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.*, 7(2):98–108, Feb 2006.

Comeron, J.M. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics*, 167(3): 1293–1304, Jul 2004.

Comeron, J.M., Ratnappan, R. and Bailin, S. The many landscapes of recombination in Drosophila melanogaster. *PLoS Genet.*, 8(10): e1002905, 2012.

Davydov, I.I., Robinson-Rechavi, M. and Salamin, N. State aggregation for fast likelihood computations in molecular evolution. *Bioinformatics*, 33(3):354–362, 02 2017.

Diekmann, Y. and Pereira-Leal, J.B. Gene Tree Affects Inference of Sites Under Selection by the Branch-Site Test of Positive Selection. *Evol. Bioinform. Online*, 11(Suppl 2): 11–17, 2015.

Drummond, D.A. et al. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U.S.A.*, 102(40):14338–14343, Oct 2005.

Duret, L. and Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*, 10:285–311, 2009.

Fagerberg, L. et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell Proteomics*, 13(2): 397–406, Feb 2014.

Fiston-Lavier, A.S. et al. Drosophila melanogaster recombination rate calculator. *Gene*, 463(1-2):18–20, Sep 2010.

Fletcher, W. and Yang, Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.*, 27(10):2257–2267, Oct 2010.

Fryxell, K.J. and Zuckerkandl, E. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.*, 17(9):1371–1383, Sep 2000.

Fullerton, S.M., Bernardo Carvalho, A. and Clark, A.G. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.*, 18(6): 1139–1142, Jun 2001.

Gharib, W.H. and Robinson-Rechavi, M. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Mol. Biol. Evol.*, 30(7):1675–1686, Jul 2013.

Goldman, N. and Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11 (5):725–736, Sep 1994.

Gossmann, T.I., Keightley, P.D. and Eyre-Walker, A. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol Evol*, 4(5):658–667, 2012.

Guharoy, M. and Chakrabarti, P. Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl. Acad. Sci. U.S.A.*, 102(43):15447–15452, Oct 2005.

Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, 59(3):307–321, May 2010.

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

Hellmann, I. et al. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.*, 72(6): 1527–1535, Jun 2003.

Hellmann, I. et al. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res.*, 18 (7):1020–1029, Jul 2008.

Hodgkinson, A. and Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.*, 12(11):756–766, Oct 2011.

Hwang, D.G. and Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 101(39):13994–14001, Sep 2004.

Keightley, P.D., Lercher, M.J. and Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.*, 3(2):e42, Feb 2005.

Keightley, P.D. et al. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res.*, 19(7):1195–1201, Jul 2009.

Kertesz, M. et al. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107, Sep 2010.

Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412):471–475, Aug 2012.

Koonin, E.V. and Wolf, Y.I. Constraints and plasticity in genome and molecular-phenome evolution. *Nat. Rev. Genet.*, 11(7): 487–498, Jul 2010.

Kosiol, C. et al. Patterns of positive selection in six Mammalian genomes. *PLoS Genet.*, 4 (8):e1000144, Aug 2008.

Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. Tissue-Specific Evolution of Protein Coding Genes in Human and Mouse. *PLoS ONE*, 10(6):e0131673, 2015.

Kudla, G. et al. Coding-sequence determinants of gene expression in Escherichia coli. *Science*, 324(5924):255–258, Apr 2009.

Lartillot, N. and Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, 21(6):1095–1109, Jun 2004.

Leffler, E.M. et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science*, 339(6127):1578–1582, Mar 2013.

Lercher, M.J. and Hurst, L.D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.*, 18(7):337–340, Jul 2002.

Li, J.J. et al. Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome Res.*, 24(7):1086–1101, Jul 2014.

Majewski, J. and Ott, J. Distribution and characterization of regulatory elements in the human genome. *Genome Res.*, 12(12):1827–1836, Dec 2002.

Marais, G., Mouchiroud, D. and Duret, L. Neutral effect of recombination on base composition in Drosophila. *Genet. Res.*, 81(2):79–87, Apr 2003.

Mattick, J.S. and Makunin, I.V. Non-coding RNA. *Hum. Mol. Genet.*, 15 Spec No 1:17–29, Apr 2006.

Moretti, S. et al. Selectome update: quality control and computational improvements to a database of positive selection. *Nucleic Acids Res.*, 42(Database issue):D917–921, Jan 2014.

Murrell, B. et al. Gene-wide identification of episodic selection. *Mol. Biol. Evol.*, 32(5):1365–1371, May 2015.

Muse, S.V. and Gaut, B.S. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, 11(5):715–724, Sep 1994.

Plotkin, J.B. and Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.*, 12(1):32–42, Jan 2011.

Pond, S.K. and Muse, S.V. Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.*, 22(12):2375–2385, Dec 2005.

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

Pop, C. et al. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. Syst. Biol.*, 10:770, Dec 2014.

Ptak, S.E. et al. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat. Genet.*, 37(4):429–434, Apr 2005.

Ratnakumar, A. et al. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 365(1552):2571–2580, Aug 2010.

Romiguier, J. et al. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526): 261–263, Nov 2014.

Rubinstein, N.D. et al. Evolutionary models accounting for layers of selection in protein-coding genes and their impact on the inference of positive selection. *Mol. Biol. Evol.*, 28 (12):3297–3308, Dec 2011.

Rudolph, K.L. et al. Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States. *PLoS Genet.*, 12(5): e1006024, May 2016.

Russo, C.A. et al. Phylogenetic analysis and a time tree for a large drosophilid data set (diptera: Drosophilidae). *Zoological Journal of the Linnean Society*, 169(4):765–775, 2013.

Scheffler, K., Martin, D.P. and Seoighe, C. Robust inference of positive selection from recombining coding sequences. *Bioinformatics*, 22(20):2493–2499, Oct 2006.

Segurel, L., Wyman, M.J. and Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet*, 15:47–70, 2014.

Spielman, S.J., Wan, S. and Wilke, C.O. A Comparison of One-Rate and Two-Rate Inference Frameworks for Site-Specific dN/dS Estimation. *Genetics*, 204(2):499–511, Oct 2016.

Stamatoyannopoulos, J.A. et al. Human mutation rate associated with DNA replication timing. *Nat. Genet.*, 41(4):393–395, Apr 2009.

Studer, R.A. et al. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res.*, 18(9): 1393–1402, Sep 2008.

Supek, F. and Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*, 521 (7550):81–84, May 2015.

Venkat, A., Hahn, M.W. and Thornton, J.W. Multinucleotide mutations cause false inferences of positive selection. *bioRxiv*, 2017. doi: 10.1101/165969. URL `http://www.biorxiv.org/content/early/2017/07/20/165969.1`.

Wagenmakers, E.J. and Farrell, S. AIC model selection using Akaike weights. *Psychonomic bulletin & review*, 11(1):192–196, 2004.

Yang, J.R., Zhuang, S.M. and Zhang, J. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol. Syst. Biol.*, 6:421, Oct 2010.

Yang, Z. *Molecular evolution: a statistical approach*. Oxford University Press, 2014.

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

Yang, Z. and dos Reis, M. Statistical properties of the branch-site test of positive selection. *Mol. Biol. Evol.*, 28(3):1217–1228, Mar 2011.

Yang, Z. et al. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–449, May 2000.

Yang, Z., Wong, W.S. and Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.*, 22 (4):1107–1118, Apr 2005.

Zhang, G. et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, 346(6215):1311–1320, Dec 2014.

Zhang, J., Nielsen, R. and Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.*, 22(12): 2472–2479, Dec 2005.