Estimating Empirical Codon Hidden Markov Models

Nicola De Maio,* Ian Holmes,† Christian Schlötterer,* and Carolin Kosiol*

*Institute of Population genetics, Veterinärmedizinische Universität Wien, Veterinärplatz 1, 1210 Vienna, Austria; and †Department of Bioengineering, University of California, Berkeley

Supplementary Material
Supplementary Text

Annotation and Alignment of Exons in Different *Drosophila* Clades

We estimated our models using alignments data coming from 3 different *Drosophila* clades. Here we describe the protocol used by Palmieri et al. (personal communication) to jointly annotate and align *Dpse-Dlow*. We use this same protocol to annotate and align *Dana-Dbip*. All the following steps, unless otherwise noted, were performed with Perl scripts that were kindly provided by Nicola Palmieri. We further used this protocol to obtain *Dmel-Dsim* alignments used in Tables S4 and S5.

First, annotations were downloaded from Flybase for the best annotated genome in each pair (i.e. *Dmel* version 5.29, *Dana* version 1.3, and *Dpse* version 2.19, which we will address as the "reference" genomes). For each gene only the longest isoform was kept. Exonerate (Slater and Birney, 2005) was run (parameters -model protein2genome -bestn 1 -showtargetgtf 1) on the genes selected in the previous step in the reference genomes against the respective target genomes (*Dsim*, *Dlow* and *Dbip*). From the Exonerate output we then filtered out genes with nonsense codons, frame shifts, genes annotated on different chromosomes, sharing start or stop codon with other genes, duplicated genes (with 90% similarity or more), genes not having start or stop codon, genes with shifted boundaries between the 2 species, and genes with different number of exons in the 2 species. Finally, CDS alignments were extracted and treated as already described in Material and Methods.

Performance of Models on Different Clades

We used 99% of the *melanogaster* clade data set to train an ECM (ECM\textsubscript{mel}) and 99% of the *pseudobscura* clade data set to train a second ECM (ECM\textsubscript{pse}). The remaining 1% of the data sets are used to test these models. We found that ECM\textsubscript{mel} outperforms ECM\textsubscript{pse} on the *melanogaster* clade test data (log-likelihood -247,515 versus -248,001). In a similar way ECM\textsubscript{pse} outperforms ECM\textsubscript{mel} on the *pseudobscura* clade test data (log-likelihood -156,402 versus -156,751).

Additional Codon Models

In this study we explored the performance of various empirical and mechanistic codon models. Some of these models were not included in the main manuscript, because of their weak performance on genome-wide data, and minor relevance to the conclusions of the article. Below we describe these models, that employ various levels of statistical complexity.

Additional Non-HMM Models

Non-Reversible ECM

The non-reversible ECM is obtained from the standard ECM (eq. 1 in main text) relaxing the constraint $s_{ij} = s_{ji}$ so that 3660 exchangeability parameters need to be estimated (in addition to the codon frequencies) bringing to a total of 3720 free parameters.
Codon Extension of the Nucleotide GTR

As a representative of mechanistic models we used an extension of the nucleotide General Time Reversible (GTR) model (Rodriguez et al., 1990) to a codon model, that also includes MNSs:

\[
q_{ij} = \begin{cases} 
  s_{d(i,j)} \pi_j & i \rightarrow j \text{ syn. single nt change} \\
  \omega s_{d(i,j)} \pi_j & i \rightarrow j \text{ nonsyn. single nt change} \\
  s_{2\pi_j} & i \rightarrow j \text{ syn. double nt change} \\
  \omega s_{2\pi_j} & i \rightarrow j \text{ nonsyn. double nt change} \\
  s_{3\pi_j} & i \rightarrow j \text{ syn. triple nt change} \\
  \omega s_{3\pi_j} & i \rightarrow j \text{ nonsyn. triple nt change}, 
\end{cases} \quad (S1)
\]

where \(d(i,j) = (n_1, n_2)\) if and only if codons \(i\) and \(j\) only differ at one position, and at that position codon \(i\) has nucleotide \(n_1\) and codon \(j\) has \(n_2\). This model has 9 rate parameters: 6 for the single nucleotide substitutions since we set \(s_{n_1n_2} = s_{n_2n_1}\), 2 for MNSs, and 1 for \(\omega\). It also has 60 free parameters describing codon frequencies.

The Combined Semi-Empirical Model

As a representative of semi-empirical models we introduce one that consists of a combination of a GTR nucleotide model and an empirical amino-acid model:

\[
q_{ij} = \begin{cases} 
  A_{a_i, a_j} s_{d(i,j)} \pi_j & i \rightarrow j \text{ single nt change} \\
  s_{2\pi_j} & i \rightarrow j \text{ double syn. nt change} \\
  s_{2\text{ns}\pi_j} & i \rightarrow j \text{ double nonsyn. nt change} \\
  s_{3\pi_j} & i \rightarrow j \text{ triple syn. nt change} \\
  s_{3\text{ns}\pi_j} & i \rightarrow j \text{ triple nonsyn. nt change}, 
\end{cases} \quad (S2)
\]

where \(A_{a_i, a_j} = A_{a_j, a_i}\) is the exchangeability amino-acid parameter between the amino-acids of codons \(i\) and \(j\), and \(s_{d(i,j)}\) is the nucleotide exchangeability parameter as in eq. S1. Counting amino-acid exchangeabilities, nucleotide exchangeabilities, and codon frequencies, this model has 162 free parameters (only amino-acid pairs separated by a single nucleotide substitution are considered). Therefore, this model has an intermediate complexity between a classical mechanistic codon model and a classical ECM. Similar to the model proposed by Doron-Faigenboim and Pupko (2007) it combines empirical amino-acid and mechanistic codon features, but in our case the empirical amino-acid substitution matrix is not pre-estimated. We also model MNSs in a different way. We briefly call this model Combined.

Additional ecHMMs

We investigated several possible heterogeneity features for our ecHMMs. In the main text we only present the codon usage ecHMM (cu-ecHMM) and the nonsynonymous rate ecHMM (R-ecHMM). Yet, it is possible that other evolutionary features vary along the genome. Those presented here resulted in a significant increase in likelihood with respect to non-HMM models, but the increase in performance was smaller compared to the other ecHMMs presented in main text.

The Transition/Transversion ecHMM

In the \(\kappa\)-ecHMM each class has a specific transition/transversion rate ratio \(\kappa\). The substitution rates of class \(k\) will therefore be:

\[
q_{ij}^{(k)} = \begin{cases} 
  s_{1\pi_j} & \text{if } i \rightarrow j \text{ transversion} \\
  \kappa^{(k)} s_{1\pi_j} & \text{if } i \rightarrow j \text{ transition}. 
\end{cases} \quad (S3)
\]
The Total Substitution Rate ecHMM

In the $T$-ecHMM each class has a specific total substitution rate $T$. The substitution rates of class $k$ are:

$$q^{(k)}_{ij} = T^{(k)} s_{ij} \pi_j.$$  \hfill (S4)

The MNS Rate ecHMM

In the MNS-ecHMM each class has specific total MNS rate, that is, each class $k$ has the following substitution rates:

$$q^{(k)}_{ij} = \begin{cases} 
  s_{ij} \pi_j & \text{if } i \to j \text{ single nt change} \\
  R^{(k)}_{MNS} s_{ij} \pi_j & \text{if } i \to j \text{ MNS} 
\end{cases}$$  \hfill (S5)

Description of Parameters for Hidden Classes

When an ecHMM with $K$ hidden classes (all of the same type) is defined, we add some free parameters in the model describing the hidden class structure. For each class $k \in \{0, ..., K-1\}$ we have a probability parameter $\tau_k$ representing the probability that a CDS starts at class $k$. Since these sum up to 1 they give $K-1$ free parameters. Then for each class $k$ we introduce $K-1$ parameters describing the change rates among classes: for any $l \neq k$ we have that $\tau_{kl}$ is the probability of moving from class $k$ to class $l$. They do not sum up to one since $1 - \sum_{l \neq k} \tau_{kl} = \tau_{kk} + \eta_k$ where $\tau_{kk}$ is the probability to remain at class $k$ and $\eta_k$ is the probability that the alignment ends at site $k$. So we have in total $K$ free parameters for $\eta_k$ and $K(K-1)$ for $\tau_{kl}$.

Finally we have to count the parameters describing class-specific evolution. For the first class no additional parameter is needed, but for any other cu-class we need to add 60 free parameters (so in total the Kcu-ecHMM has $K-1 + K + K(K-1) + 60(K-1)$ class parameters) while for all other HMM models we need to add 1 free parameter for each additional class (bringing to a total of $K-1 + K + K(K-1) + K-1$ class parameters). Note that even when $K = 1$, i.e. when no HMM structure is present, there is 1 free parameter ($\eta_0$) describing the length of CDSs. We do not count this parameter for any model (ECM or ecHMM) in order to avoid confusion, and since neglecting a parameter does not affect substitution rates, or BIC and AIC score differences.

The $2cu-2R$-ecHMM

When more than 1 type of class is used (i.e. the $2cu-2R$-ecHMM), we allow, at each codon, changes of class of any class type, but not of more than 1 class type at the same time. For example if we are in class $(k_1, k_2)$, or equivalently in cu-class $k_1$ and R-class $k_2$, we can move to class $(l_1, k_2)$ or $(k_1, l_2)$, but not to class $(l_1, l_2)$ in one step (where of course $l_1 \neq k_1$ and $l_2 \neq k_2$). So if we have 2 cu-classes and 2 R-classes we need to add: 3 free parameters for the probabilities to start a CDS in any class combination, 12 class change probabilities, 1 R-value for the second R-class and 60 free parameters for the codon frequencies in the second cu-class (in total we have 76 free parameters for hidden classes).

Comparison of HMM and non-HMM Site Heterogeneity

When we compare HMM to non-HMM models of site heterogeneity (i.e. Tables S6 and S7), in order to make the comparison meaningful we constrain ourselves to the essential class parameters, and discard the ones that have negligible effect on likelihood improvement. So, assuming we only have 2 classes, for all models we set $\eta_0 = \eta_1$. Furthermore for ecHMMs we set the constraint $\tau_0 = \frac{\tau_{00}}{\tau_{01} + \tau_{10}}$. For the models with constant class on CDSs we set $\tau_{01} = \tau_{10} = 0$ and we leave $\tau_0$ as a free parameter. Finally for the models with independent sites we set $\tau_{01} = \tau_1$ and $\tau_{10} = \tau_0$, with again $\tau_0$ as a free parameter.
In ecHMMs used for positive selection tests we use the constraints $\tau_k = \tau_0$ and $\eta_k = 0.002$ (the true value) for every class $k$. Note that for the comparison between ecHMM1a and ecHMM2a a $\chi^2_4$ distribution needs to be used instead of a $\chi^2_2$. 
Likelihood surface for 2R-ecHMM

An important property for a good model is the possibility to recover, for example via maximum likelihood estimation on simulated data, its true parameter values. As mentioned in main text, this is not the case for 2R-ecHMM. When estimating the 2R-ecHMM from simulated data, we obtained very large errors for class parameters, but we also found that the likelihood of estimated parameters were very similar to the likelihood of the simulated ones, suggesting a flat likelihood surface. We wanted to explore the likelihood surface near the simulated points to find evidence of local maxima or very flat surface. Since this was too computationally demanding, we explored a similar but simpler problem. Instead of a sequence of aligned sites, we simulated a sequence of samples from a mixture Poisson distribution. The idea is the following: the number of nonsynonymous substitutions in a site can be used as a summary statistic for inferring $R$. We expected the distribution of nonsynonymous substitutions to be nearly Poisson (assuming that multiple non-synonymous substitutions in the same site are detectable) with mean proportional to the value of $R$ for the class where the site lies.

We simulated 2 sequences of samples from 2 different Poisson processes, and mixed them according to an HMM structure so that neighboring sites tend to come from the same Poisson distribution (see description of fig. S1 for further details). When looking at the likelihood shape, we found it problematic when the expected number of nonsynonymous substitutions per site is far below 1 for both classes. In fact, all pairs $(R^0, R^1)$ with $R^0 + R^1 = \text{const.}$ have very similar likelihood, making it very hard to distinguish the 2 classes (fig. S1), at least using the Expectation-Maximization algorithm in Xrate.

Therefore, we concluded that 2R-ecHMM correct estimation with Xrate is difficult at least when alignments of only 2 species are considered (as in our case). Yet, if several species are included, the number of nonsynonymous substitutions per site increases, and this might make inferences more reliable.
Figure S1: Likelihood surfaces of 2 mixed Poisson processes, mimicking the 2R-ecHMM
We simulated 100,000 observations according to 2 independent Poisson processes of mean respectively $\lambda_1$ and $\lambda_2$ and with $\lambda_2 = \lambda_1 \times 30$. We mixed the samples from the 2 distribution so that nearby sites likely came from the same source. We plotted the likelihood surfaces calculated correctly assuming 2 independent Poisson processes. We varied $\lambda_1$ and $\lambda_2$ ($x$ and $y$ axis in the plots) nearby the correct (simulated) values. Plot (a) shows that the likelihood surface is problematic for $\lambda_1=0.005$ (almost constant values along a diagonal), while plot (b) shows that for $\lambda_1=0.03$ the problem disappears.
Methods for Positive Selection Tests

Comparison of our $\hat{\omega}$ with $\omega_E$ from Kosiol, Holmes and Goldman (2007)

Estimating the nonsynonymous/synonymous ratio $\omega$ from genome-wide empirical codon rate estimates is an important and, we think, non-trivial task. Kosiol, Holmes and Goldman (2007) proposed a method that compares the total nonsynonymous and synonymous rates in the ECM with a constant expected under neutrality. They call their parameter $\omega_E$ (empirical $\omega$). We propose an alternative method (equations 8-11 in main text) that accounts for possible different mutation rates in different data sets. We also aim at obtaining values comparable to the $\omega$ of the model M0. We call this new parameter $\hat{\omega}$.

Here we show that our method is robust to variation in mutation rates. On the other hand, $\omega_E$ is affected by mutational biases. To prove this, we considered the model M0 (eq. 5, main text) as a particular instance of an ECM. From this M0 we calculated both $\hat{\omega}$ and $\omega_E$. In M0 we set homogeneous codon frequencies, a constant $\omega = 0.1$ (which is the true parameter to recover) and we vary $\kappa$ between 1.0 and 10.0. We observed that $\hat{\omega}$ accurately recovers $\omega$, while $\omega_E$ is affected by variation in $\kappa$ and is largely below the true value (table S1).

Table S1: Comparison Between $\hat{\omega}$ and $\omega_E$.

| $\kappa$ in M0 | $\hat{\omega}^a$ | $\omega_E^b$ |
|---------------|------------------|--------------|
| 1.0           | 0.1              | 0.0778       |
| 2.0           | 0.1              | 0.0689       |
| 3.0           | 0.1              | 0.0643       |
| 4.0           | 0.1              | 0.0615       |
| 5.0           | 0.1              | 0.0596       |
| 6.0           | 0.1              | 0.0582       |
| 7.0           | 0.1              | 0.0572       |
| 8.0           | 0.1              | 0.0563       |
| 9.0           | 0.1              | 0.0557       |
| 10.0          | 0.1              | 0.0552       |

Given a model M0 with $\kappa$ as specified in the first column, $\omega = 0.1$, and homogeneous codon frequencies, we compare the performance of 2 methods to estimate $\omega$.

a The estimation of $\omega$ based on equations 8-11, main text.
b The estimation of $\omega$ based on the method proposed by Kosiol, Holmes and Goldman (2007).

Modifying MNS Rates in Models for Positive Selection

When we used empirical models to test positive selection, we found them more conservative than classical tests. In order to investigate whether this is due to the generally improved modeling of substitution rates, or in particular to MNSs, we also tested mechanistic models with added MNSs (M1a+MNS, M2a+MNS, M7+MNS, M8+MNS) and empirical models with MNS rate fixed to 0 (restricted ecM1a, restricted ecM2a, restricted ecM7, restricted ecM8). When we add MNSs to a mechanistic model, we include the MNS rates of the ECM used in simulations, after scaling them so that the proportion of MNSs in the new model is the same as in the simulated ECM.

Adjusting False Positives for Model Comparison

We have shown that, under short phylogenetic trees, and in the presence of MNSs, classical positive selection tests have high false positive rates, higher than the nominal significance cutoff. As a solution we proposed the use of models explicitly accounting for MNSs, like ECMs. The new tests show, indeed, acceptable false positive rates, but also greatly reduced power (table S11). Therefore they are not strictly preferable to classical ones, but only better with respect to type I error. Here we test whether, after calibrating the significance cutoff for different tests in order
to have similar maximum false positive rates among all scenarios, the new tests become more powerful. We found that, first, in order to bring false positives below 5% in all scenarios with classical tests, we had to move the significance cutoff to 0.998. Second, after bringing false positive rates to similar levels, the power of different models seems comparable (table S13).
List of Supplementary Files

**Supplementary File S1:** Text file describing the standard ECM as a phylo-grammar. This file (as also the other supplementary files) can be used as input grammar for Xrate. Given this file, genome-wide CDS alignments, and a phylogenetic tree, Xrate can estimate the exchangeability parameters and codon frequencies of the ECM.

**Supplementary File S2:** Text file describing the simplified ECM as a phylo-grammar, usable as Xrate input.

**Supplementary File S3:** Text file describing the 2R-2cu-ecHMM as a phylo-grammar, usable as Xrate input.

References

Doron-Faigenboim A, Pupko T. 2007. A combined empirical and mechanistic codon model. Molecular Biology and Evolution. 24:388.

Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. Molecular Biology and Evolution. 24:1464.

Rodriguez F, Oliver J, Marin A, Medina J. 1990. The general stochastic model of nucleotide substitution. Journal of Theoretical Biology. 142:485–501.

Sackton T, Lazzaro B, Schlenke T, Evans J, Hultmark D, Clark A. 2007. Dynamic evolution of the innate immune system in Drosophila. Nature Genetics. 39:1461–1468.

Slater G, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 6:31.
## Supplementary Tables

Table S2: Numbers and Types of Exchangeability Parameters in ECM and Simplified ECM.

|                  | Single syn. | Single nonsyn. | Double syn. | Double nonsyn. | Triple syn. | Triple nonsyn. |
|------------------|-------------|----------------|-------------|----------------|-------------|----------------|
| Codon pairs      | 134         | 392            | 28          | 1540           | 12          | 1554           |
| ECM exch. Params.| 67          | 196            | 14          | 770            | 6           | 777            |
| simplified ECM exch. params. | 67          | 196            | 1           | 1              | 1           | 1              |

Number of exchangeability parameters used by ECM and simplified ECM for different classes of substitutions, and number of ordered codon pairs belonging to those classes.
| Model name | Number of parameters | BIC score\(^a\) | AIC score\(^b\) | MNS proportion\(^c\) |
|------------|----------------------|-----------------|-----------------|----------------------|
| **2Dmel-Dsim**: 46,197 CDSs, 5,403,560 codons | | | | |
| Non-rev. ECM\(^d\) | 3720 | - | - | 2.8% |
| ECM | 1890 | -5,000 | +20,000 | 2.7% |
| Simpl. ECM\(^e\) | 323 | -23,000 | +22,000 | 2.3% |
| Combined | 162 | +68,000 | +116,000 | 2.4% |
| Nucl. GTR\(^f\) | 69 | +283,000 | +332,000 | 4.0% |
| **2Dmel-Dsim-Dyak**: 44,788 CDSs, 5,162,718 codons | | | | |
| Non-rev. ECM | 3720 | - | - | 3.8% |
| ECM | 1890 | +53,000 | +78,000 | 3.7% |
| Simpl. ECM | 323 | +44,000 | +90,000 | 2.9% |
| Combined | 162 | +260,000 | +308,000 | 3.0% |
| Nucl. GTR | 69 | +786,000 | +835,000 | 5.6% |
| **2Dmel-Dsim-Dyak-Dana**: 25,012 CDSs, 2,267,923 codons | | | | |
| Non-rev. ECM | 3720 | - | - | 4.3% |
| ECM | 1890 | +57,000 | +82,000 | 3.5% |
| Simpl. ECM | 323 | +46,000 | +92,000 | 2.8% |
| Combined | 162 | +259,000 | +307,000 | 2.8% |
| Nucl. GTR | 69 | +769,000 | +818,000 | 5.3% |
| **2Dmel-Dsim-Dyak-Dana**: 24,331 CDSs, 2,176,111 codons | | | | |
| Non-rev. ECM | 3720 | - | - | 4.1% |
| ECM | 1890 | +126,000 | +145,000 | 4.0% |
| Simpl. ECM | 323 | +121,000 | +158,000 | 2.7% |
| Combined | 162 | +455,000 | +490,000 | 2.9% |
| Nucl. GTR | 69 | +648,000 | +694,000 | 3.3% |

The best model according to BIC score is shown in **bold**.

\(^a\) BIC score difference between the current model and the non-reversible ECM trained on the same data set (models with smaller BIC score are considered preferable).

\(^b\) AIC score difference between the current model and the non-reversible ECM trained on the same data set (models with smaller AIC score are considered preferable).

\(^c\) Estimated proportion of Multiple Nucleotide Substitutions.

\(^d\) Non reversible Empirical Codon Model.

\(^e\) Simplified Empirical Codon Model.

\(^f\) Nucleotide General Time Reversible model.
Table S4: Models Performance on Drosophila Clades (All Aligned with Same Procedure).

| Model name | Number of parameters | BIC score\(^a\) | AIC score\(^b\) | MNS proportion\(^c\) |
|------------|----------------------|-----------------|-----------------|---------------------|
| **Dmel-Dsim**: 25,941 CDSs and 3,179,704 codons | | | | |
| Non-rev. ECM\(^d\) | 3720 | - | - | 2.8% |
| ECM | 1890 | -6,000 | +18,000 | 2.7% |
| Simpl. ECM\(^e\) | 323 | -25,000 | +19,000 | 2.2% |
| 2R-ecHMM | 328 | -65,000 | -21,000 | 0.7% |
| 2cu-ecHMM | 387 | -124,000 | -81,000 | 2.1% |
| **2R-2cu-ecHMM**\(^f\) | 398 | -184,000 | -141,000 | 1.5% |
| **Dpse-Dlow**: 29,483 CDSs and 3,796,335 codons | | | | |
| Non-rev. ECM | 3720 | - | - | 2.7% |
| ECM | 1890 | -18,000 | +6,000 | 2.7% |
| Simpl. ECM | 323 | -37,000 | +8,000 | 2.2% |
| 2R-ecHMM | 328 | -84,000 | -39,000 | 0.6% |
| 2cu-ecHMM | 387 | -140,000 | -96,000 | 2.1% |
| **2R-2cu-ecHMM** | 398 | -218,000 | -174,000 | 1.4% |
| **Dana-Dbip**: 32,962 CDSs and 4,306,332 codons | | | | |
| Non-rev. ECM | 3720 | - | - | 4.5% |
| ECM | 1890 | +28,000 | +52,000 | 4.5% |
| Simpl. ECM | 323 | +9,000 | +54,000 | 3.2% |
| 2R-ecHMM | 328 | -109,000 | -64,000 | 1.5% |
| 2cu-ecHMM | 387 | -156,000 | -112,000 | 2.7% |
| **2R-2cu-ecHMM** | 398 | -277,000 | -233,000 | 2.5% |

The best model according to BIC score is shown in **bold**.

\(^a\) BIC score difference between the current model and the non reversible ECM trained on the same data set.

\(^b\) AIC score difference between the current model and the non reversible ECM trained on the same data set.

\(^c\) Proportion of Multiple Nucleotide Substitutions estimated by the model.

\(^d\) Non reversible Empirical Codon Model.

\(^e\) Simplified Empirical Codon Model.

\(^f\) The ecHMM having 2 classes for nonsynonymous/synonymous rate ratio variation and 2 classes for codon usage variation.

Table S5: Comparison of ECM Estimates from Different Clades (All Aligned with Same Procedure).

| Feature\(^a\) | Dmel-Dsim vs. Dana-Dbip | Dmel-Dsim vs. Dpse-Dlow | Dana-Dbip vs. Dpse-Dlow |
|--------------|-------------------------|-------------------------|------------------------|
| ECM Q        | 16.6%                   | 18.9%                   | 23.8%                  |
| simpl. ECM Q | 16.0%                   | 18.9%                   | 23.1%                  |
| 2-R & 2-cu-ecHMM Q | 14.5%                   | 16.9%                   | 22.1%                  |
| ECM π        | 7.0%                    | 12.0%                   | 12.7%                  |
| ECM Nucleotide | 7.3%                    | 10.9%                   | 15.4%                  |

\(^a\) Model feature that is compared between clades: “Q” is the instantaneous rate matrix, “π” is the codon frequencies vector and “Nucleotide” stands for the nucleotide substitution rates matrix extracted from the ECM averaging the single-nucleotide synonymous substitution rates for each ordered pair of nucleotides. The values shown are the euclidean distances of the parameters vectors, normalized by the average of the norm of the 2 vectors.
Table S6: HMM Versus non-HMM Modeling of Codon Usage Heterogeneity.

| Model name          | Number of parameters | BIC score $^a$ |
|---------------------|----------------------|----------------|
| **Dmel-Dsim**: 25,941 CDSs and 3,179,704 codons |                     |                |
| 2cu-ecHMM           | 385                  | -              |
| fixed-cu$^b$        | 384                  | +6,055         |
| independent-cu$^c$  | 384                  | +103,080       |

| **Dpse-Dlow**: 29,483 CDSs and 3,796,335 codons |                     |                |
| 2cu-ecHMM           | 385                  | -              |
| fixed-cu$^b$        | 384                  | +2,589         |
| independent-cu$^c$  | 384                  | +101,693       |

| **Dana-Dbi**: 32,962 CDSs and 4,306,332 codons |                     |                |
| 2cu-ecHMM           | 385                  | -              |
| fixed-cu$^b$        | 384                  | +14,372        |
| independent-cu$^c$  | 384                  | +163,522       |

| **Dmel-Dsim** (downloaded): 47,689 CDSs and 5,578,031 codons |                     |                |
| 2cu-ecHMM           | 385                  | -              |
| fixed-cu$^b$        | 384                  | +5,914         |
| independent-cu$^c$  | 384                  | +182,928       |

| **Dmel-Dsim-Dyak**: 44,788 CDSs and 5,162,718 codons |                     |                |
| 2cu-ecHMM           | 385                  | -              |
| fixed-cu$^b$        | 384                  | +21,025        |
| independent-cu$^c$  | 384                  | +174,472       |

| **Dmel-Dsim-Dyak-Dana** : 25,012 CDSs and 2,267,923 codons |                     |                |
| 2cu-ecHMM           | 385                  | -              |
| fixed-cu$^b$        | 384                  | -3,248         |
| independent-cu$^c$  | 384                  | +45,412        |

The best model according to BIC score is shown in **bold**.

$^a$ BIC score difference between the current model and the 2cu-ecHMM (with the constraints $\tau_0 = \frac{\tau_{10}}{\tau_{10} + \tau_{01}}$ and $\eta_0 = \eta_1$ for this table) trained on the same data set.

$^b$ Model with 2 cu-classes but constant codon usage along exons.

$^c$ Model in which codons can independently belong to any of the 2 cu-classes.
Table S7: Comparison of ecHMMs with Different Class Features.

| Model name | Number of parameters | BIC score<sup>a</sup> |
|------------|---------------------|----------------------|
| Dmel-Dsim: 25,941 CDSs and 3,179,704 codons | | |
| 2R-ecHMM | 326 | - |
| 2fixed-R<sup>b</sup> | 325 | +386 |
| 2independent-R<sup>c</sup> | 325 | +42,514 |
| 2κ-ecHMM | 326 | +603,038 |
| 2MNM-ecHMM | 326 | +38,678 |
| 2T-ecHMM | 326 | +44,777 |
| Dyse-Dlow: 29,483 CDSs and 3,796,335 codons | | |
| 2R-ecHMM | 326 | - |
| 2fixed-R | 325 | -1,416 |
| 2independent-R | 325 | +44,593 |
| 2κ-ecHMM | 326 | +382,479 |
| 2MNM-ecHMM | 326 | +46,995 |
| 2T-ecHMM | 326 | +48,936 |
| Dana-Dhp: 32,962 CDSs and 4,306,332 codons | | |
| 2R-ecHMM | 326 | - |
| 2fixed-R | 325 | +38,369 |
| 2independent-R | 325 | +115,709 |
| 2κ-ecHMM | 326 | +1,764,475 |
| 2MNM-ecHMM | 326 | +113,815 |
| 2T-ecHMM | 326 | +996,330 |
| Dmel-Dsim (downloaded): 47,689 CDSs and 5,578,031 codons | | |
| 2R-ecHMM | 326 | - |
| 2fixed-R | 325 | +2,048 |
| 2independent-R | 325 | +90,063 |
| 2κ-ecHMM | 326 | +1,043,186 |
| 2MNM-ecHMM | 326 | +90,386 |
| 2T-ecHMM | 326 | +98,288 |
| Dmel-Dsim-Dyak: 44,788 CDSs and 5,162,718 codons | | |
| 2R-ecHMM | 326 | - |
| 2fixed-R | 325 | +53,203 |
| 2independent-R | 325 | +218,850 |
| 2κ-ecHMM | 326 | +2,033,278 |
| 2MNM-ecHMM | 326 | +238,858 |
| 2T-ecHMM | 326 | +274,625 |
| Dmel-Dsim-Dyak-Dana: 25,012 CDSs and 2,267,923 codons | | |
| 2R-ecHMM | 326 | - |
| 2fixed-R | 325 | +40,451 |
| 2independent-R | 326 | +52,358 |
| 2κ-ecHMM | 326 | +2,104,118 |
| 2MNM-ecHMM | 326 | +65,155 |
| 2T-ecHMM | 326 | +1,057,188 |

The best model according to BIC score is shown in **bold**.

<sup>a</sup> BIC score difference between the current model and the 2R-ecHMM (all the ecHMMs have the constraints $\tau_0 = \tau_{00}$ and $\eta_0 = \eta_1$ in this table) trained on the same data set.

<sup>b</sup> Model with 2 R-classes but constant codon usage along exons.

<sup>c</sup> Model in which codons can independently belong to any of the 2 R-classes.
Table S8: Full Table of ecHMM Performance on Real Data.

| Model name | Number of parameters | BIC score<br>a | AIC score<br>b | MNS proportion<br>c |
|------------|----------------------|----------------|----------------|---------------------|
| 2Dmel-Dsim |                      |                |                |                     |
| 2cu-ecHMM  | 387                  | -182,000       | -183,000       | 2.3%                |
| 3cu-ecHMM  | 453                  | -218,000       | -220,000       | 2.3%                |
| 4cu-ecHMM  | 521                  | -231,000       | -234,000       | 2.3%                |
| 2R-ecHMM   | 328                  | -100,000       | -100,000       | 2.0%                |
| 3R-ecHMM   | 335                  | -106,000       | -106,000       | 2.0%                |
| 4R-ecHMM   | 344                  | -112,000       | -112,000       | 2.0%                |
| 2R-2cu-ecHMM | 398             | -297,000       | -298,000       | 1.7%                |

| Model name | Number of parameters | BIC score<br>a | AIC score<br>b | MNS proportion<br>c |
|------------|----------------------|----------------|----------------|---------------------|
| Dmel-Dsim-Dyak |                      |                |                |                     |
| 2cu-ecHMM  | 387                  | -178,000       | -179,000       | 2.5%                |
| 3cu-ecHMM  | 453                  | -218,000       | -220,000       | 2.5%                |
| 4cu-ecHMM  | 521                  | -240,000       | -243,000       | 2.5%                |
| 2R-ecHMM   | 328                  | -253,000       | -253,000       | 2.3%                |
| 3R-ecHMM   | 335                  | -264,000       | -264,000       | 2.2%                |
| 4R-ecHMM   | 344                  | -290,000       | -290,000       | 2.3%                |
| 2R-2cu-ecHMM | 398            | -429,000       | -430,000       | 2.2%                |

| Model name | Number of parameters | BIC score<br>a | AIC score<br>b | MNS proportion<br>c |
|------------|----------------------|----------------|----------------|---------------------|
| Dmel-Dsim-Dyak-Dana |                      |                |                |                     |
| 2cu-ecHMM  | 387                  | -171,000       | -172,000       | 2.4%                |
| 3cu-ecHMM  | 453                  | -209,000       | -211,000       | 2.4%                |
| 4cu-ecHMM  | 521                  | -228,000       | -231,000       | 2.4%                |
| 2R-ecHMM   | 328                  | -245,000       | -245,000       | 2.1%                |
| 3R-ecHMM   | 335                  | -256,000       | -256,000       | 2.1%                |
| 4R-ecHMM   | 344                  | -284,000       | -284,000       | 2.1%                |
| 2R-2cu-ecHMM | 398            | -413,000       | -414,000       | 2.1%                |

| Model name | Number of parameters | BIC score<br>a | AIC score<br>b | MNS proportion<br>c |
|------------|----------------------|----------------|----------------|---------------------|
| Dmel-Dsim-Dyak-Dana |                      |                |                |                     |
| 2cu-ecHMM  | 387                  | -46,000        | -47,000        | 2.7%                |
| 3cu-ecHMM  | 453                  | -59,000        | -61,000        | 2.7%                |
| 4cu-ecHMM  | 521                  | -62,000        | -64,000        | 2.7%                |
| 2R-ecHMM   | 328                  | -78,000        | -78,000        | 2.7%                |
| 3R-ecHMM   | 335                  | -106,000       | -106,000       | 2.7%                |
| 4R-ecHMM   | 344                  | -111,000       | -111,000       | 2.7%                |
| 2R-2cu-ecHMM | 398            | -131,000       | -132,000       | 2.7%                |

| Model name | Number of parameters | BIC score<br>a | AIC score<br>b | MNS proportion<br>c |
|------------|----------------------|----------------|----------------|---------------------|
| Dmel-Dsim-Dyak-Dana |                      |                |                |                     |
| 2cu-ecHMM  | 387                  | -47,000        | -48,000        | 2.7%                |
| 3cu-ecHMM  | 453                  | -58,000        | -60,000        | 2.6%                |
| 4cu-ecHMM  | 521                  | -60,000        | -62,000        | 2.6%                |
| 2R-ecHMM   | 328                  | -76,000        | -76,000        | 2.6%                |
| 3R-ecHMM   | 335                  | -101,000       | -101,000       | 2.6%                |
| 4R-ecHMM   | 344                  | -107,000       | -107,000       | 2.6%                |
| 2R-2cu-ecHMM | 398            | -126,000       | -127,000       | 2.6%                |

The best model according to BIC score is shown in **bold**.

a BIC score difference between the current model and the simplified ECM trained on the same data set.
b AIC score difference between the current model and the simplified ECM trained on the same data set.
c Proportion of Multiple Nucleotide Substitutions estimated by the model.
d The ecHMM having 2 classes for nonsynonymous/synonymous rate ratio variation and 2 classes for Codon Usage variation.
Table S9: Computational Time required for Model Estimation.

| Codons   | Simplified ECM | ECM  | 2cu-ecHMM |
|----------|----------------|------|-----------|
| $10^4$   | 0m 54s         | 1m 11s | 3m 23s   |
| $2 \times 10^4$ | 2m 1s       | 1m 39s | 4m 16s   |
| $5 \times 10^4$ | 5m 11s      | 4m 24s | 12m 7s   |
| $10^5$   | 8m 23s         | 10m 49s | 34m 52s  |
| $2 \times 10^5$ | 18m 44s      | 21m 47s | 43m 59s  |
| $5 \times 10^5$ | 47m 15s     | 75m 8s  | 89m 29s  |
| $10^6$   | 94m 41s        | 109m 32s | 362m 1s |

Models have been estimated on a data set simulated according to a 2cu-ecHMM (see Materials and Methods). The following convergence criteria of Xrate have been used: `-mi 0.000001` and `-f 3`. 
Table S10: All Simulations Scenarios.

| Simulation scenario | Substitution model | Codon frequencies | Phylogenetic tree | ω | Sequence length | Number of replicates |
|---------------------|--------------------|-------------------|-------------------|---|-----------------|---------------------|
| ECM estimation      | ECM from Dmel-Dsim, Dmel-Dsim-Dyak and Dmel-Dsim-Dyak-Dana. | π of the ECM respectively used. | Estimated from Dmel-Dsim, Dmel-Dsim-Dyak and Dmel-Dsim-Dyak-Dana. | As in real data, not modified. | 5 × 10^4, 10^5, 2 × 10^5, 5 × 10^5, 10^6, 2 × 10^6, 3 × 10^6 codons. | 1 |
| 2cu-ecHMM estimation | 2cu-ecHMM estimated from Dmel-Dsim. | 2 π sets from the 2 classes of the 2cu-ecHMM estimated on Dmel-Dsim. | Estimated from Dmel-Dsim. | As in real data, not modified. | 10^4, 2 × 10^4, 5 × 10^4, 10^5, 2 × 10^5, 5 × 10^5, 10^6 codons. | 1 |

Positive selection tests

| Substitution model | Codon frequencies | Phylogenetic tree | ω | Sequence length | Number of replicates |
|--------------------|-------------------|-------------------|---|-----------------|---------------------|
| 2cu-ecHMM estimated on Dmel-Dsim. | π(0) from the 2cu-ecHMM estimated on Dmel-Dsim. | Dmel-Dsim, Dmel-Dsim-Dyak and Dmel-Dsim-Dyak-Dana. | ω_0 = 0, ω_1 = 0.9, p_0 = 0.5, p_1 = 0.5, p_2 = 0.1, p_3 = 0.1 | 500 codons | 1000 |

Positive selection tests with ecHMMs

| Substitution model | Codon frequencies | Phylogenetic tree | ω | Sequence length | Number of replicates |
|--------------------|-------------------|-------------------|---|-----------------|---------------------|
| ECM estimated on Dmel-Dsim-Dyak | π of the ECM from Dmel-Dsim-Dyak | Estimated from Dmel-Dsim-Dyak | ω_0 = 0.2, ω_1 = 0.9, ω_2 = 0.9, τ_00 = 0.95, τ_11 = 0.95 | 500 codons | 1000 |

---

*a* We have 3 trees and 3 models, and we simulated according to each of the 9 possible combinations.

*b* p_i is the frequency of sites with ω = ω_i. Classes are assigned independently for each codon.

*c* τ_ij is the probability that a codon has ω = ω_j given that the previous codon has ω = ω_i. Therefore, here ω is not assigned independently for each codon.
Table S11: Performance of Positive Selection Tests on Simulated Data.

| Model          | No Positive Selection (False Positives) | With Positive Selection (Power) |
|----------------|----------------------------------------|-------------------------------|
|                | $p_0 = 0.9, p_1 = 0.1$, $\omega_0 = 0, \omega_1 = 1$ | $p_0 = 0.45, p_1 = 0.45$, $\omega_0 = 0, \omega_1 = 1$, $\omega_2 = 5$ |
| No MNSs        |                                        |                               |
| M1a-M2a        | 11.8%(4.4%)                            | 34.3%(17.8%)                  |
| restr. ecM1a-ecM2a | 15.3%(6.6%)                             | 47.1%(29.7%)                  |
| With MNSs      |                                        |                               |
| M1a-M2a + MNS | 3.7%(0.8%)                             | 7.4%(1.9%)                    |
| ecM1a-ecM2a    | 3.1%(0.8%)                             | 3.3%(1.0%)                    |
| No MNSs        |                                        |                               |
| M7-M8          | 14.0%(5.3%)                            | 35.7%(18.5%)                  |
| restr. ecM7-ecM8 | 16.7%(7.7%)                             | 48.4%(30.8%)                  |
| With MNSs      |                                        |                               |
| M7-M8 + MNS    | 4.5%(1.6%)                             | 7.6%(1.9%)                    |
| ecM7-ecM8      | 3.5%(1.1%)                             | 3.6%(1.0%)                    |
| Dmel-Dsim-Dyak |                                        |                               |
| No MNSs        |                                        |                               |
| M1a-M2a        | 6.8%(2.3%)                             | 21.7%(10.5%)                  |
| restr. ecM1a-ecM2a | 10.5%(3.5%)                             | 42.3%(22.9%)                  |
| With MNSs      |                                        |                               |
| M1a-M2a + MNS | 1.5%(0.0%)                             | 3.9%(1.3%)                    |
| ecM1a-ecM2a    | 1.4%(0.1%)                             | 3.4%(0.9%)                    |
| No MNSs        |                                        |                               |
| M7-M8          | 8.8%(3.2%)                             | 24.2%(11.0%)                  |
| restr. ecM7-ecM8 | 12.9%(4.8%)                             | 43.5%(24.7%)                  |
| With MNSs      |                                        |                               |
| M7-M8 + MNS    | 3.3%(0.9%)                             | 4.5%(1.4%)                    |
| ecM7-ecM8      | 2.8%(0.3%)                             | 3.6%(1.1%)                    |
| Dmel-Dsim-Dsc-Dyak-Dere-Dana |                        |                               |
| No MNSs        |                                        |                               |
| M1a-M2a        | 3.8%(0.7%)                             | 9.5%(3.8%)                    |
| restr. ecM1a-ecM2a | 10.1%(3.1%)                             | 38.1%(18.7%)                  |
| With MNSs      |                                        |                               |
| M1a-M2a + MNS | 0.5%(0.3%)                             | 1.2%(0.3%)                    |
| ecM1a-ecM2a    | 0.6%(0.1%)                             | 2.3%(0.6%)                    |
| No MNSs        |                                        |                               |
| M7-M8          | 7.0%(2.2%)                             | 14.5%(6.1%)                   |
| restr. ecM7-ecM8 | 14.2%(4.9%)                             | 43.1%(22.3%)                  |
| With MNSs      |                                        |                               |
| M7-M8 + MNS    | 2.9%(0.5%)                             | 3.3%(1.1%)                    |
| ecM7-ecM8      | 1.8%(0.1%)                             | 2.4%(0.8%)                    |

Proportion of LRTs detecting positive selection over 1000 simulations. LRTs were performed picking outliers from the 5% (1%) tail of a $\chi^2$ distribution. Alignments were simulated according to the same ECM used for ecM1a, ecM2a, ecM7 and ecM8 (see Material and Methods).

a “restr.” (restricted) means that MNS rates are set to 0.
b “+ MNS” means that multiple nucleotide substitutions are allowed in the mechanistic models used for the test.
Table S12: Performance of HMM and non-HMM Positive Selection Tests on Simulated Data.

| Data set       | ccM1a-ccM2a | ecHMM1a-ecHMM2a |
|----------------|-------------|-----------------|
| False Positives|             |                 |
| $\tau_{00} = 0.95, \tau_{11} = 0.5$ | 0%           | 0.3%            |
| $\omega_1 = 1$   |             |                 |
| $\tau_{00} = 0.9, \tau_{11} = 0.9$ | 0%           | 0%              |
| $\omega_1 = 1$   |             |                 |
| Power           |             |                 |
| $\tau_{00} = 0.95, \tau_{11} = 0.5$ | 41.6%        | 48.5%           |
| $\omega_1 = 2$   |             |                 |
| $\tau_{00} = 0.95, \tau_{11} = 0.5$ | 97.8%        | 98.3%           |
| $\omega_1 = 3$   |             |                 |

Proportion of HMM and non-HMM tests detecting positive selection on 1000 simulations. LRTs were performed picking outliers from the 5% tail of a $\chi^2_4$ distribution. Alignments were simulated according to the same ECM used for ecM1a, ecM2a, ecHMM1a and ecHMM2a, and according to an 8-species phylogenetic tree with homogeneous branch lengths (see Material and Methods and table S10).
Table S13: Performance of Positive Selection Tests on Simulated Data, after Calibration of Different Models.

| Model                  | No Positive Selection (False Positives) | With Positive Selection (Power) |
|------------------------|----------------------------------------|---------------------------------|
|                         | $p_0 = 0.9, p_1 = 0.1$                  | $p_0 = 0.45, p_1 = 0.45,$       |
|                         | $\omega_0 = 0, \omega_1 = 1$           | $p_2 = 0.1,$                    |
|                         |                                         | $\omega_0 = 0, \omega_1 = 1,$  |
|                         |                                         | $\omega_2 = 1.5$                |
|                         |                                         | $\omega_2 = 5$                  |
| **Dmel-Dsim**           |                                        |                                 |
| No MNSs                 |                                        |                                 |
| M1a-M2a                 | 3.8%(1.8%, 0.2%)                        | 14.8%(8.9%, 3.6%)               |
|                        | 9.3%(4.9%, 1.1%)                        | 70.9%(59.4%, 38.2%)             |
|                        | 1.2%(0.7%, 0.2%)                        | 12.5%(6.1%, 1.5%)               |
|                        | 2.6%(1.2%, 0.4%)                        | 11.6%(4.6%, 1.0%)               |
|                        | 4.0%(1.2%, 0.3%)                        | 9.9%(4.8%, 1.1%)                |
|                        | 10.0%(5.1%, 1.0%)                       | 15.4%(8.6%, 1.6%)               |
|                        | 12.0%(5.1%, 1.0%)                       | 71.7%(55.8%, 30.6%)             |
| With MNSs               |                                        |                                 |
| M1a-M2a + MNS           | 8.4%(4.5%, 0.4%)                        | 14.9%(8.1%, 1.4%)               |
|                        | 7.0%(3.3%, 0.6%)                        | 72.2%(59.3%, 33.5%)             |
|                        | 4.6%(1.6%, 0.5%)                        | 11.6%(4.6%, 1.0%)               |
|                        | 9.9%(4.9%, 1.2%)                        | 71.3%(60.2%, 39.3%)             |
|                        | 10.9%(5.1%, 1.0%)                       | 70.2%(54.8%, 28.9%)             |
| No MNSs                 |                                        |                                 |
| M7-M8                   | 4.2%(2.3%, 0.2%)                        | 15.3%(9.1%, 3.8%)               |
|                        | 9.8%(4.8%, 1.1%)                        | 71.5%(60.2%, 39.3%)             |
|                        | 3.0%(1.2%, 0.2%)                        | 13.1%(6.3%, 1.6%)               |
|                        | 11.6%(6.1%, 1.5%)                       | 70.2%(54.8%, 28.9%)             |
| With MNSs               |                                        |                                 |
| M7-M8 + MNS             | 9.8%(5.1%, 1.0%)                        | 15.4%(8.6%, 1.6%)               |
|                        | 7.2%(3.5%, 0.6%)                        | 72.8%(60.4%, 34.4%)             |
|                        | 5.0%(1.6%, 0.5%)                        | 71.7%(55.8%, 30.6%)             |
| **Dmel-Dsim-Dyak**      |                                        |                                 |
| No MNSs                 |                                        |                                 |
| M1a-M2a                 | 1.6%(0.4%, 0.0%)                        | 7.4%(4.2%, 1.2%)                |
|                        | 1.9%(0.7%, 0.2%)                        | 90.7%(84.9%, 71.0%)             |
|                        | 2.6%(1.2%, 0.4%)                        | 8.7%(4.2%, 0.7%)                |
|                        | 3.0%(1.1%, 0.1%)                        | 9.9%(4.6%, 0.9%)                |
|                        | 2.0%(0.8%, 0.2%)                        | 8.4%(4.4%, 1.3%)                |
|                        | 9.9%(4.6%, 0.9%)                        | 96.5%(91.2%, 75.2%)             |
| With MNSs               |                                        |                                 |
| M1a-M2a + MNS           | 4.9%(2.1%, 0.0%)                        | 8.8%(4.7%, 1.2%)                |
|                        | 3.3%(1.3%, 0.3%)                        | 91.6%(86.1%, 68.0%)             |
|                        | 5.4%(2.4%, 0.1%)                        | 96.5%(91.2%, 75.2%)             |
|                        | 2.7%(0.9%, 0.2%)                        | 9.9%(4.6%, 0.9%)                |
|                        | 2.9%(1.3%, 0.4%)                        | 9.5%(4.7%, 0.8%)                |
|                        | 1.2%(0.4%, 0.1%)                        | 95.7%(90.3%, 74.3%)             |
| With MNSs               |                                        |                                 |
| M7-M8 + MNS             | 7.5%(3.9%, 0.5%)                        | 10.5%(5.0%, 1.3%)               |
|                        | 3.7%(1.4%, 0.3%)                        | 92.6%(87.6%, 70.1%)             |
|                        | 7.8%(4.0%, 0.3%)                        | 96.9%(92.2%, 76.7%)             |
| **Dmel-Dsim-Dsec-Dyak-Dere-Dana** |                                 |                                 |
| No MNSs                 |                                        |                                 |
| M1a-M2a                 | 0.6%(0.4%, 0.1%)                        | 2.6%(1.1%, 0.4%)                |
|                        | 0.9%(0.3%, 0.2%)                        | 99.0%(98.1%, 94.5%)             |
|                        | 1.0%(0.2%, 0.0%)                        | 5.8%(2.0%, 0.7%)                |
|                        | 4.3%(1.8%, 0.2%)                        | 100%(99.8%, 98.4%)              |
| With MNSs               |                                        |                                 |
| M1a-M2a + MNS           | 1.5%(0.6%, 0.3%)                        | 4.0%(1.3%, 0.3%)                |
|                        | 4.3%(1.8%, 0.2%)                        | 99.9%(98.2%, 94.9%)             |
|                        | 2.4%(0.7%, 0.1%)                        | 5.2%(2.8%, 0.6%)                |
|                        | 1.3%(0.3%, 0.0%)                        | 100%(99.9%, 97.9%)              |
| No MNSs                 |                                        |                                 |
| M7-M8                   | 1.3%(0.3%, 0.0%)                        | 4.3%(2.1%, 0.5%)                |
|                        | 2.6%(1.2%, 0.3%)                        | 99.4%(98.6%, 96.0%)             |
|                        | 1.2%(0.2%, 0.0%)                        | 7.4%(2.9%, 1.1%)                |
|                        | 1.2%(0.2%, 0.0%)                        | 100%(99.8%, 98.3%)              |
| With MNSs               |                                        |                                 |
| M7-M8 + MNS             | 7.2%(3.6%, 0.5%)                        | 7.6%(3.8%, 1.0%)                |
|                        | 8.8%(3.9%, 0.7%)                        | 99.4%(99.0%, 96.6%)             |
|                        | 7.5%(2.6%, 0.1%)                        | 7.7%(3.3%, 0.8%)                |
|                        | 2.3%(0.9%, 0.1%)                        | 100%(99.6%, 97.9%)              |

Proportion of LRTs detecting positive selection over 1000 simulations. Here we use a different significance level for each model, so that the maximum amount of false positives on simulated data is 10% (respectively 5%, 1%). LRTs were performed picking outliers from a $\chi^2_2$ distribution. For standard mechanistic models a significance level of 0.994 (resp. 0.998, 0.9998) was used. For mechanistic models with MNSs we used 0.87 (0.94, 0.9915), for empirical models 0.835 (0.93, 0.99), and finally for empirical models without MNSs 0.9989 (0.9998, 0.999995). Alignments were simulated according to the same ECM used for ecM1a, ecM2a, ecM7 and ecM8 (see Material and Methods).

*a* “restr.” (restricted) means that MNS rates are set to 0.

*b* “+ MNS” means that multiple nucleotide substitutions are allowed in the mechanistic models used for the test.
Supplementary Figures

Figure S2: Estimation error of ECM (2).
Error in estimating ECM exchangeability parameters \( \Delta \) and substitution rates • with phylogenies: *Dmel-Dsim* (red), *Dmel-Dsim-Dyak* (green), *Dmel-Dsim-Dyak-Dana* (blue). The error is measured as the euclidean distance between the vector of the estimated parameters and the vectors of parameters used for simulations, normalized by the norm of the vector of parameters used for simulations. The plot shows the error as a percentage on the Y axes. On the X axes is the number of codons simulated. The vertical purple line represents the amount of codons in the smallest real data set used. The ECM used for these simulations is the one estimated on *Dmel-Dsim-Dyak*. 
Figure S3: Estimation error of ECM (3).
Error in estimating ECM exchangeability parameters $\triangle$ and substitution rates $\bullet$ with phylogenies: $Dmel-Dsim$ (red), $Dmel-Dsim-Dyak$ (green), $Dmel-Dsim-Dyak-Dana$ (blue). The error is measured as the euclidean distance between the vector of the estimated parameters and the vectors of parameters used for simulations, normalized by the norm of the vector of parameters used for simulations. The plot shows the error as a percentage on the Y axes. On the X axes is the number of codons simulated. The vertical purple line represents the amount of codons in the smallest real data set used. The ECM used for these simulations is the one estimated on $Dmel-Dsim$. 
Figure S4: Bubbleplot of the ECM from Dmel-Dsim data. Bubbleplot representing the substitution rates of the ECM estimated from the alignments of Dmel and Dsim. The area of a circle is proportional to the substitution rate from the codon on the Y axes to the codon on the X axes. Red circles represent single-nucleotide synonymous substitutions, green circles single-nucleotide non-synonymous substitutions, blue circles (inside the matrix) synonymous MNSs and yellow circles non-synonymous MNSs. Codon frequencies are shown in the bottom vector, each codon frequency is proportional to the area of its blue circle.
Figure S5: Bubbleplot of the ECM from Dana-Dbip data.

Bubbleplot representing the substitution rates of the ECM estimated from the alignments of Dana and Dbip. The area of a circle is proportional to the substitution rate from the codon on the Y axes to the codon on the X axes. Red circles represent single-nucleotide synonymous substitutions, green circles single-nucleotide non-synonymous substitutions, blue circles (inside the matrix) synonymous MNSs and yellow circles non-synonymous MNSs. Codon frequencies are shown in the bottom vector, each codon frequency is proportional to the area of its blue circle.
Figure S6: Bubbleplot of the ECM from $Dpse$-$Dlow$ data. Bubbleplot representing the substitution rates of the ECM estimated from the alignments of $Dpse$ and $Dlow$. The area of a circle is proportional to the substitution rate from the codon on the Y axes to the codon on the X axes. Red circles represent single-nucleotide synonymous substitutions, green circles single-nucleotide non-synonymous substitutions, blue circles (inside the matrix) synonymous MNSs and yellow circles non-synonymous MNSs. Codon frequencies are shown in the bottom vector, each codon frequency is proportional to the area of its blue circle.
Figure S7: Estimation error of exchangeability parameters with 2cu-ecHMM. Error in estimation of the exchangeability matrix when data is simulated under variable codon usage, that is, under a 2cu-ecHMM. △ shows the error of the 2cu-ecHMM, • the error of the simplified ECM, and □ of the ECM. On the X axes is the number of codons simulated, on the Y axes is the error. Error is measured as the euclidean distance between the vector of the estimated parameters and the vector of parameters used for simulations, normalized by the norm of the vector of parameters used for simulations, and plotted as a percentage.
Figure S8: Estimation error of MNS parameters with 2cu-ecHMM.
Error in estimation of the MNS parameters ($s_{2s}$, $s_{2ns}$, $s_{3s}$ and $s_{3ns}$) when data is simulated under variable codon usage (2cu-ecHMM). 2cu-ecHMM is also fit to the data. ○ shows the error in estimating $s_{2s}$, □ the error in $s_{2ns}$, ● in $s_{3s}$, and △ in $s_{3ns}$. On the X axes is the number of codons simulated, on the Y axes is the error. Error is measured as the absolute value of the difference between the estimated parameter and true parameter, normalized by the absolute value of true parameter, and plotted as a percentage.
Figure S9: Application of positive selection tests to real data.
Number of genes detected to be under positive selection by different tests, from a total of 181 genes used in Sackton et al. (2007). Numbers shared by the circles represent genes detected by multiple tests. The test ECM1a-ECM2a detected no gene to be under positive selection, and therefore is not included in the plot.