Supplementary Material for Zhou, Gu, and Wilke, Detecting positive and purifying selection at synonymous sites in yeast and worm

Counting model for estimating the proportions of differences at non-conservative and conservative synonymous sites

In the main text, we describe a maximum likelihood model for computing the rates of conservative and non-conservative synonymous substitutions. We also developed a counting model to estimate these rates based on the method developed by Nei and Gojobori (1986).

In the counting model, we compare sequences codon by codon and count the numbers and types of substitutions. We denote the number of conservative and non-conservative synonymous substitutions per codon by sc and sn, respectively. For two codons that differ by a single synonymous substitution, we simply count this substitution as either conservative or non-conservative. For example, if the codons we compare are GTT and GTA for valine in yeast, there is one synonymous difference with $sc = 0$ and $sn = 1$. When two nucleotide differences exist between the two codons, we consider the alternative paths by which the two mutations may have arisen. For example, in a comparison of CGA and AGG for arginine in yeast, the two pathways are as follows: i) CGA-CGG-AGG and ii) CGA-AGA-AGG. (In yeast, CGT and AGA are preferred codons whereas CGC, CGA, CGG, and AGG are unpreferred.) Pathway i) involves two conservative synonymous substitutions, and pathway ii) involves two non-conservative synonymous substitution. We assume that pathways i) and ii) occur with equal probability. The sc and sn then become 1 and 1, respectively. When there are three nucleotide differences between the two codons, there are six different possible pathways, and in each there are three substitution steps. Otherwise, the evaluation of sc and sn remains the same as in the case of two nucleotide differences. We obtain the total number of conservative and non-conservative synonymous subsititutions by summing up the sc and sn values over all codons.

To correctly normalize substitution counts, we also have to determine the number of conservative and non-conservative synonymous sites. We define a site as conservative synonymous if each possible nucleotide substitution will lead to a codon change that is synonymous and conservative. Likewise, a site is non-conservative synonymous if each possible nucleotide substitution is synonymous and non-conservative. In practice, at many sites some substiutions will be conservative synonymous, others will be non-conservative synonymous, and

others yet will be non-synonymous. In this case, we assign fractional values to the sites. For example, in yeast, the preferred codons for valine are GTT and GTC. For codon GTT, the third nucleotide position is a synonymous site. One of the three possible nucleotide substitutions at this site leads to conservative synonymous codon change (GTC) while two possible substitution (GTA and GTG) are non-conservative. Thus, the third nucleotide position of GTT is counted as one-third conservative synonymous site and two-thirds non-conservative synonymous site.

Once we have obtained the number of conservative and non-conservative synonymous sites and the number of conservative and non-conservative synonymous nucleotide substitutions, we compute the proportion of conservative synonymous substitutions (PS_{C}) by dividing the number of conservative substitutions by the average number of conservative sites for the two sequences; similarly, we compute the proportion of non-conservative synonymous substitutions (PS_{N}) by dividing the number of non-conservative synonymous substitutions by the average number of non-conservative synonymous sites for the two sequences.

We used the counting model to verify the results we obtained with the maximum-likelihood model. The results from these two models were very similar. For example, PS_{N} correlates strongly with expression level in both species (Spearman's $\rho = -0.416, P = 6.0 \times 10^{-160}$ for yeast and $\rho = -0.354, P = 1.6 \times 10^{-139}$ for worm) whereas the correlation between PS_{C} and expression vanishes or becomes much weaker in each species (Spearman's $\rho = 0.001, P = 0.941$ for yeast and $\rho = -0.092, P = 2.6 \times 10^{-10}$ for worm). See also Fig. S3.

An alternative maximum likelihood model

In our maximum likelihood model discussed in the main text, ψ is purely a measure for the difference in conservative and non-conservative synonymous substitutions within codon families. An alternative choice is to include a term representing synonymous selection into all substitutions (both synonymous and non-synonymous) that connect preferred with unpreferred codons. The transition matrix of this alternative model reads as follows:

$$q_{ij} = \begin{cases} 0 & \text{the two codons differ at more than one position} \\ \psi\alpha_{i_k j_k}\pi_j & \text{one non-conservative synonymous substitution} \\ \alpha_{i_k j_k}\pi_j & \text{one conservative synonymous substitution} \\ \omega\alpha_{i_k j_k}\pi_j & \text{one non-synonymous substitution without codon-property change} \\ \psi\omega\alpha_{i_k j_k}\pi_j & \text{one non-synonymous substitution with codon-property change} \end{cases}$$

Here, by codon-property change we mean a change from a preferred codon to an unpreferred one or vice versa.

We calculated correlations between all variables derived from our main model and from the alternative model (Table S3). We also calculated correlations with expression level. We found that dS_{N} correlates negatively with expression level in both species (Spearman's $\rho = -0.449, P = 9.6 \times 10^{-193}$ for yeast and $\rho = -0.366, P = 3.7 \times 10^{-182}$ for worm). The correlation between dS_{C} and expression level is weaker (Spearman's $\rho = -0.143, P = 3.5 \times 10^{-19}$ for yeast and $\rho = -0.133, P = 3.0 \times 10^{-24}$ for worm).

A model with four synonymous rates

In the main model, we categorized synonymous substitutions into two groups: conservative and non-conservative. To investigate the details within each groups, we added two additional parameters (η and θ) into our main model. η and θ represent codon change difference within non-conservative and conservative synonymous substitutions, respectively. The transition matrix of this alternative model reads as follows:

$$q_{ij} = \begin{cases} 0 & \text{the two codons differ at more than one position} \\ \psi\alpha_{i_k j_k}\pi_j & \text{one synonymous substitution from preferred codon } i \text{ to unpreferred codon } j \\ \eta\psi\alpha_{i_k j_k}\pi_j & \text{one synonymous substitution from unpreferred codon } i \text{ to preferred codon } j \\ \alpha_{i_k j_k}\pi_j & \text{one synonymous substitution between unpreferred codon } i \text{ and } j \\ \theta\alpha_{i_k j_k}\pi_j & \text{one synonymous substitution between preferred codon } i \text{ and } j \\ \omega\alpha_{i_k j_k}\pi_j & \text{one non-synonymous substitution between codon } i \text{ and } j \end{cases} \quad (1)$$

Test on fly data

To test whether our main model still works in fly, we redid our analysis between *Drosophila melanogaster* and *Drosophila yakuba*. The genomic and expression data for fly were obtained from the Eisen Lab (http://rana.lbl.gov/drosophila/) and Stolc et al. (2004), respectively.

The fly results differed in an important point from the results for yeast and worm. The two rates dS_{C} and dS_{N} were nearly identical for fly, and both had a similar correlation with expression level (Spearman's $\rho = -0.116, P = 7.4 \times 10^{-8}$ for dS_{C} and Spearman's $\rho = -0.170, P = 2.2 \times 10^{-15}$ for dS_{N}, Fig. S11). Consequently, the ratio $dS_{\mathrm{N}}/dS_{\mathrm{C}}$ was very close to 1 and did not change much with expression level (Spearman's $\rho = -0.054, P = 1.3 \times 10^{-2}$, Fig. S11). Yet, the distributions of ψ in fly was significantly, if ever so slightly, shifted to the left of 1 (t-test: $P \ll 10^{-100}$, Fig. S12). Results based on the physical-sites definition and the mutational opportunity definition of evolutionary rates were comparable (Fig. S12).

Table S1: List of preferred codons.

Amino acid	Yeast	Worm
Ala	GCT, GCC	GCT, GCC
Arg	AGA	CGT, CGC
Asn	AAC	AAC
Asp	GAC	GAC
Cys	TGT	TGC
Gln	CAA	CAA
Glu	GAA	GAG
Gly	GGT	GGA
His	CAC	CAC
Ile	ATT, ATC	ATC
Leu	TTG	CTT, CTC
Lys	AAG	AAG
Phe	TTC	TTC
Pro	CCA	CCA
Ser	TCT, TCC	TCT, TCC
Thr	ACT, ACC	ACT, ACC
Tyr	TAC	TAC
Val	GTT, GTC	GTT, GTC

Table S2: Spearman correlations between results from the model in the main text and results from the alternative model.

	Yeast		Worm	
Variable	ρ	P	ρ	P
ψ	0.715	$\ll 10^{-100}$	0.694	$\ll 10^{-100}$
ω	0.948	$\ll 10^{-100}$	0.944	$\ll 10^{-100}$
dS_{C}	0.852	$\ll 10^{-100}$	0.874	$\ll 10^{-100}$
dS_{N}	0.958	$\ll 10^{-100}$	0.965	$\ll 10^{-100}$
dS	0.977	$\ll 10^{-100}$	0.971	$\ll 10^{-100}$
dN	0.978	$\ll 10^{-100}$	0.977	$\ll 10^{-100}$
$dS_{\mathrm{N}}/dS_{\mathrm{C}}$	0.808	$\ll 10^{-100}$	0.785	$\ll 10^{-100}$
dN/dS_{C}	0.931	$\ll 10^{-100}$	0.926	$\ll 10^{-100}$

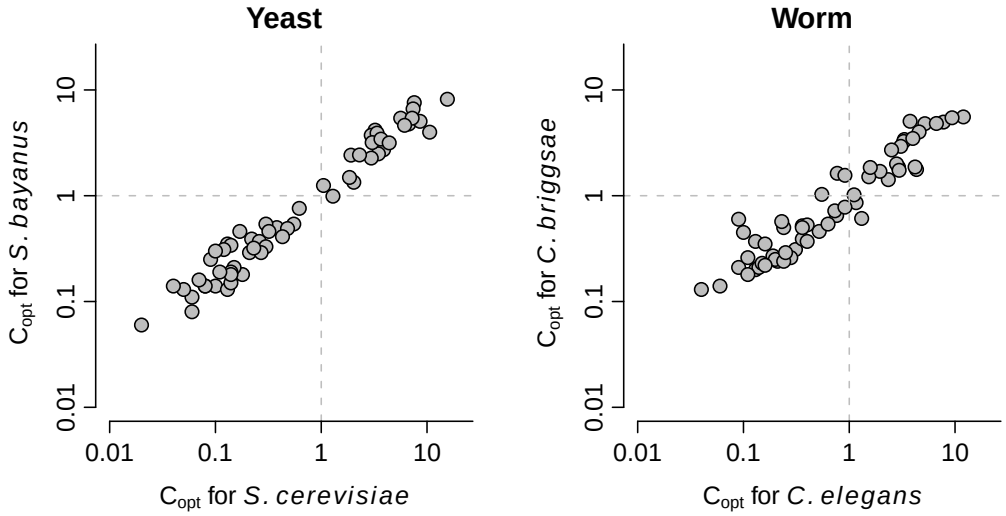

Figure S1: Comparison of codon optimality between orthologous species. Codon optimality (C_{opt}) is defined as the odds ratio of codon usage between the gene groups showing the lowest 5% and highest 5% ENC': $C_{opt} = [n_{low}/(N_{low} - n_{low})]/[n_{high}/(N_{high} - n_{high})]$. Here, n_{low} and n_{high} are the observed numbers of the codon in the lowest 5% and highest 5% ENC' groups, respectively, and N_{low} and N_{high} are the observed numbers of the corresponding amino acid in the lowest 5% and highest 5% ENC' groups.

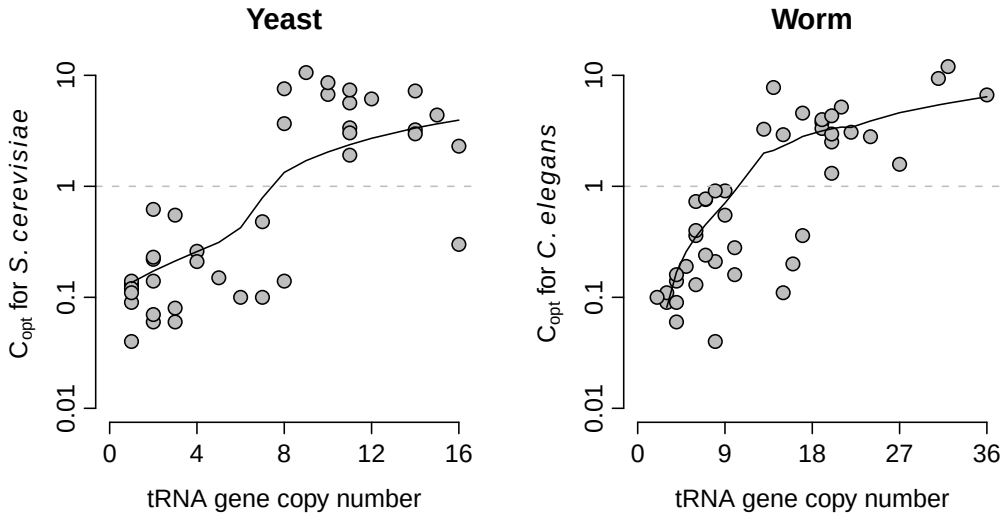

Figure S2: tRNA gene copy number versus codon optimality. The tRNA gene copy number for each codon was obtained from the Genomic tRNA Database (Chan and Lowe, 2009). Cognate codons were assigned by assuming that each DNA-encoded anticodon was matched by its reverse complement, except for anticodons with 3' adenine (ANN), which were assumed to be quantitatively modified to inosine (INN) and to prefer NNC codons rather than NNU. The codons with tRNA gene copy number equal to 0 were excluded because it is not clear which tRNA can be assigned to these codons. Codon optimality (C_{opt}) is defined as the odds ratio of codon usage between the gene groups showing the lowest 5% and highest 5% ENC': $C_{opt} = [n_{low}/(N_{low} - n_{low})]/[n_{high}/(N_{high} - n_{high})]$. Here, n_{low} and n_{high} are the observed numbers of the codon in the lowest 5% and highest 5% ENC' groups, respectively, and N_{low} and N_{high} are the observed numbers of the corresponding amino acid in the lowest 5% and highest 5% ENC' groups. The Spearman correlations between tRNA gene copy number and codon optimality are 0.722 ($P = 2.2 \times 10^{-7}$) for yeast and 0.798 ($P = 1.4 \times 10^{-10}$) for worm. Solid lines show lowess-smoothed data.

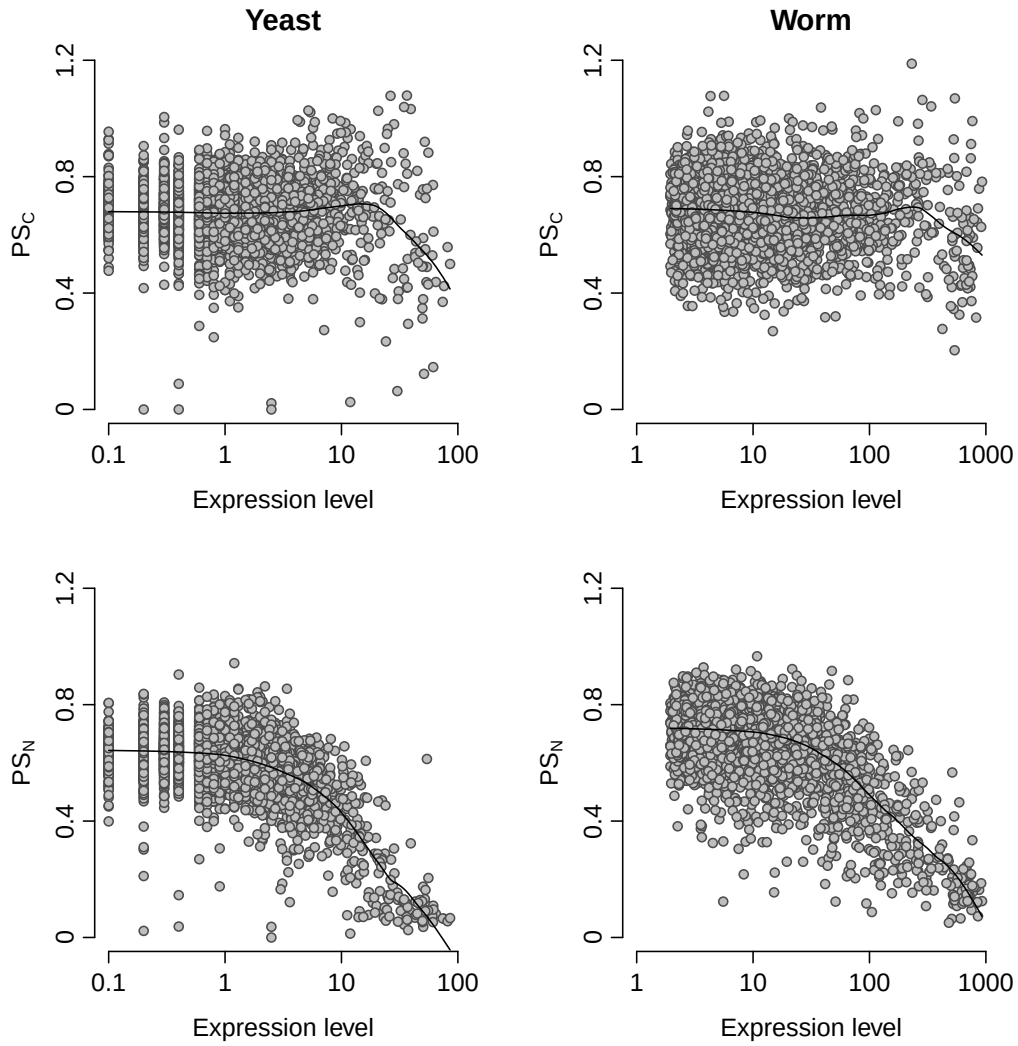

Figure S3: Correlation between gene expression level and PS_{C} and PS_{N}, for yeast (left) and worm (right). Solid lines show lowess-smoothed data.

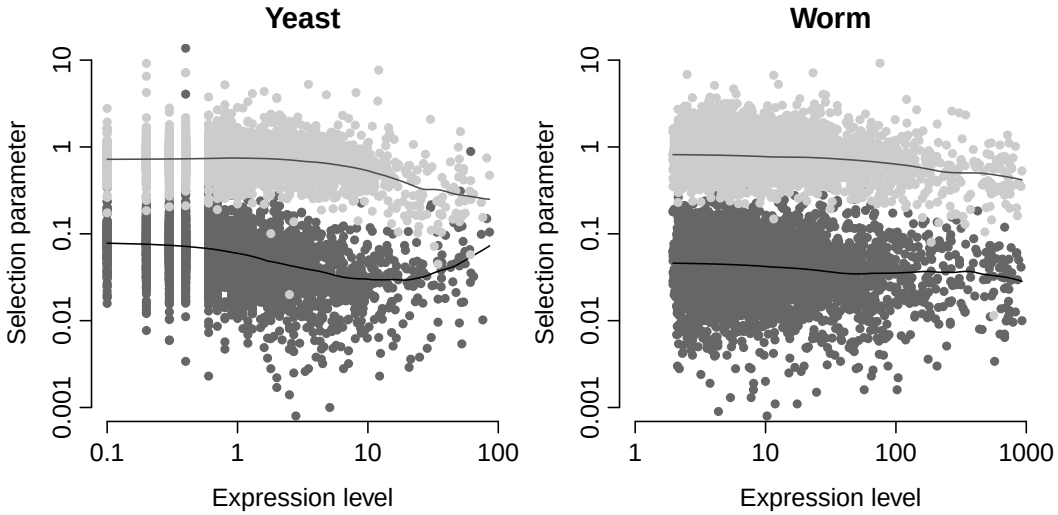

Figure S4: ω (dark points) and ψ (light points) versus expression level, for yeast (left) and worm (right). Both values decline with increasing expression levels. The Spearman correlations of expression level with ω and ψ are -0.411 ($P = 1.0 \times 10^{-155}$) and -0.119 ($P = 1.6 \times 10^{-13}$) for yeast respectively and -0.125 ($P = 8.6 \times 10^{-18}$) and -0.153 ($P = 5.7 \times 10^{-26}$) for worm respectively. Solid lines show lowess-smoothed data.

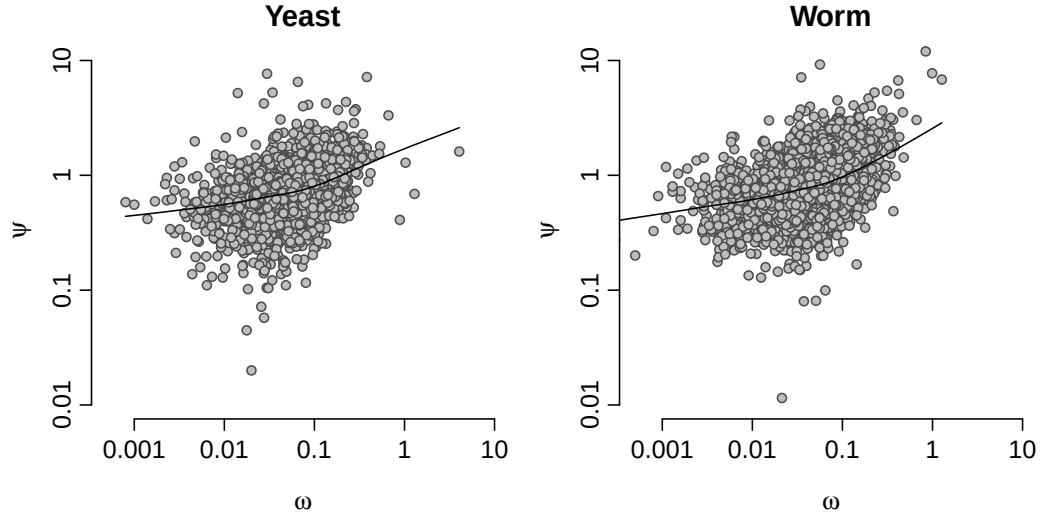

Figure S5: ω versus ψ, for yeast (left) and worm (right). Solid lines show lowess-smoothed data.

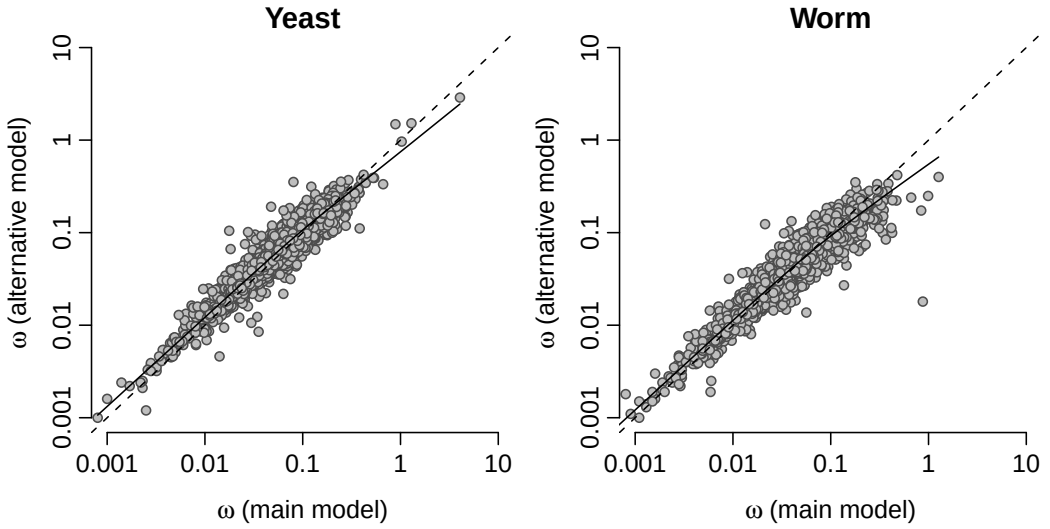

Figure S6: Evolutionary-rate ratio ω for the model described in the main text and for the alternative model described in the Supplementary Text, for yeast (left) and worm (right). The Spearman correlations for the two data sets are 0.948 ($P \ll 10^{-100}$) for yeast and 0.944 ($P \ll 10^{-100}$) for worm. The solid lines show lowess-smoothed data and the dashed lines indicate exact agreement between the two models.

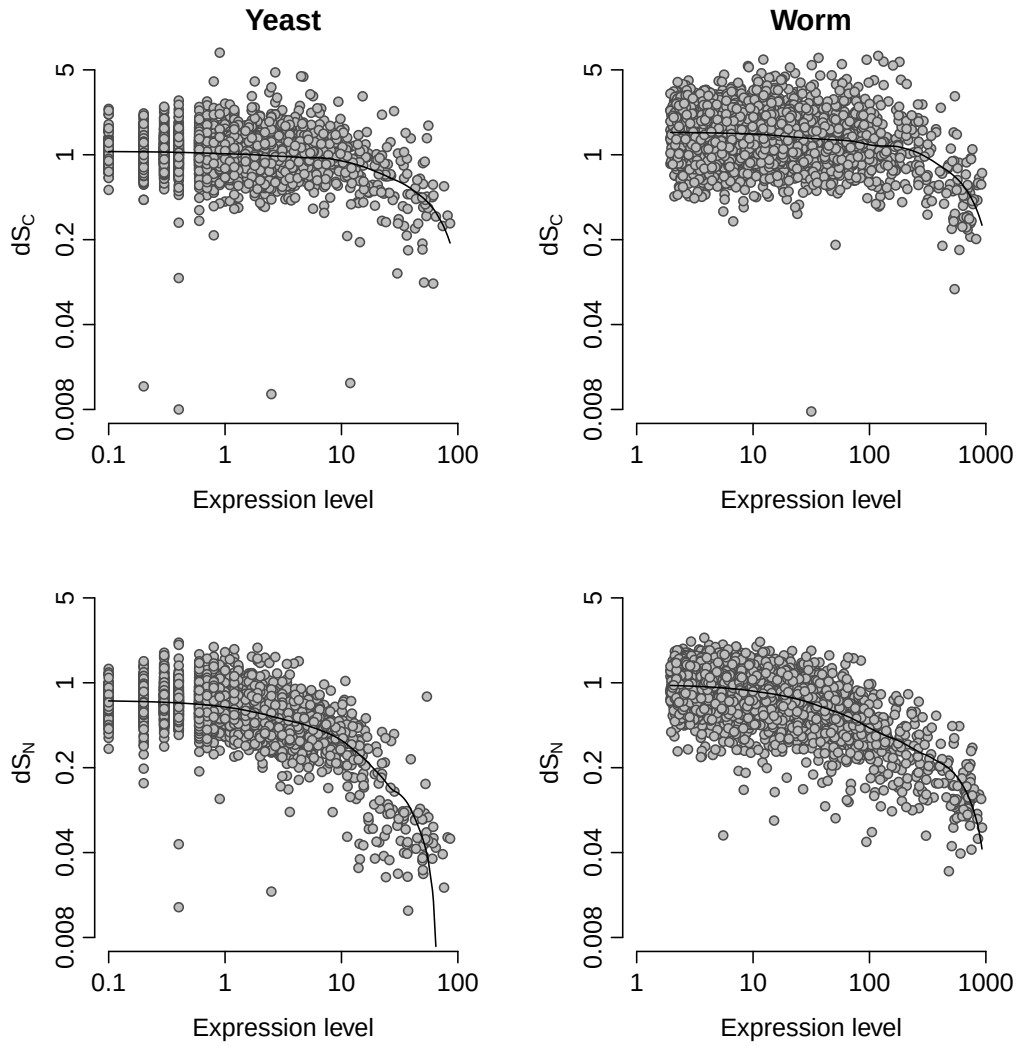

Figure S7: Evolutionary rates dS_{C} and dS_{N} calculated by the alternative model versus expression level, for yeast (left) and worm (right). Solid lines show lowess-smoothed data.

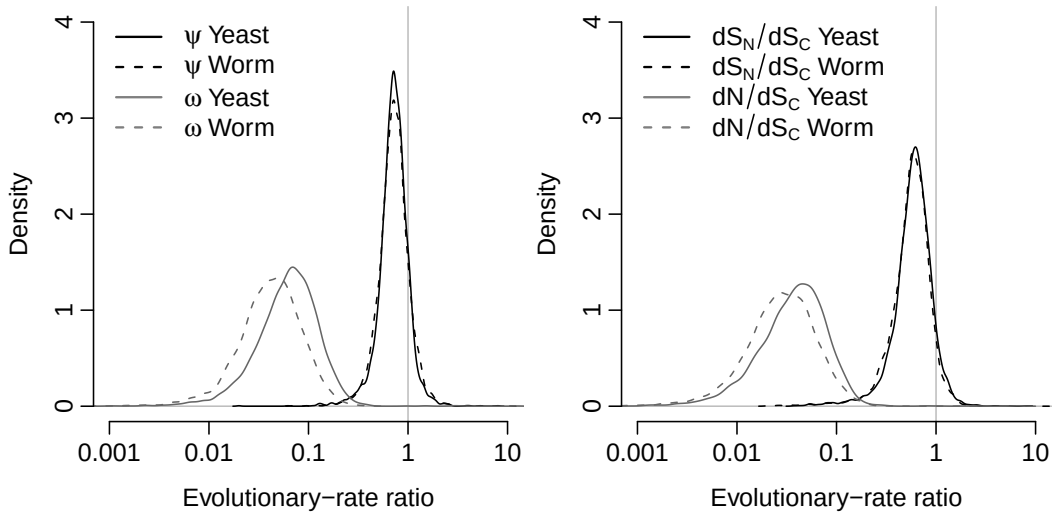

Figure S8: Distribution of evolutionary-rate ratios caculated by the alternative model. Left panel: Distribution of the ratios ω and ψ. Right panel: Distribution of the ratios dN/dS_{C} and $dS_{\mathrm{N}}/dS_{\mathrm{C}}$.

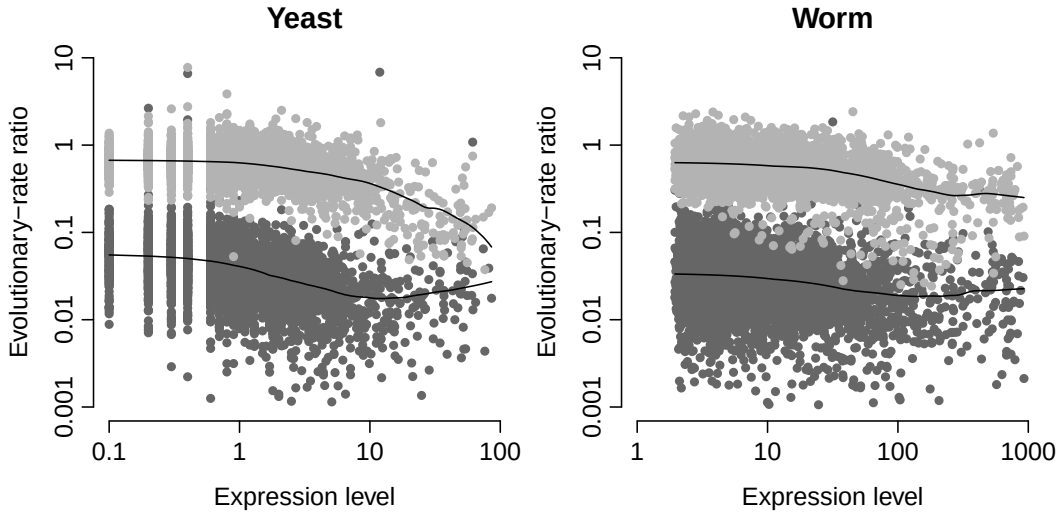

Figure S9: Evolutionary-rate ratios dN/dS_{C} (dark points) and $dS_{\mathrm{N}}/dS_{\mathrm{C}}$ (light points) caculated by the alternative model versus expression level, for yeast (left) and worm (right). Both ratios decline with increasing expression levels. The Spearman correlations of expression level with dN/dS_{C} and $dS_{\mathrm{N}}/dS_{\mathrm{C}}$ are -0.465 ($P = 7.8 \times 10^{-209}$) and -0.342 ($P = 1.2 \times 10^{-107}$) for yeast respectively and -0.215 ($P = 3.1 \times 10^{-61}$) and -0.265 ($P = 1.6 \times 10^{-93}$) for worm respectively. Solid lines show lowess-smoothed data.

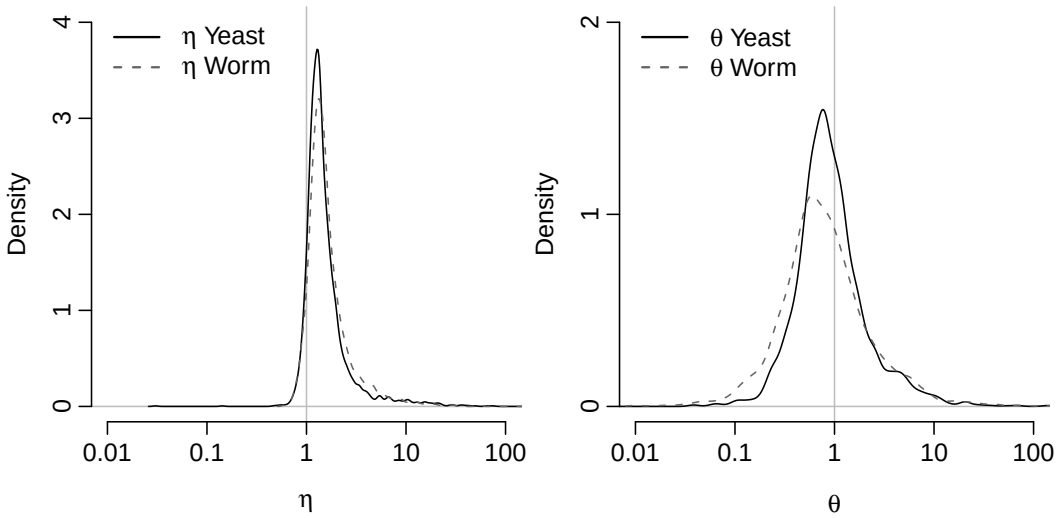

Figure S10: Distribution of η and θ caculated by the model with two more parameters. Left panel: Distribution of η. Right panel: Distribution of θ.

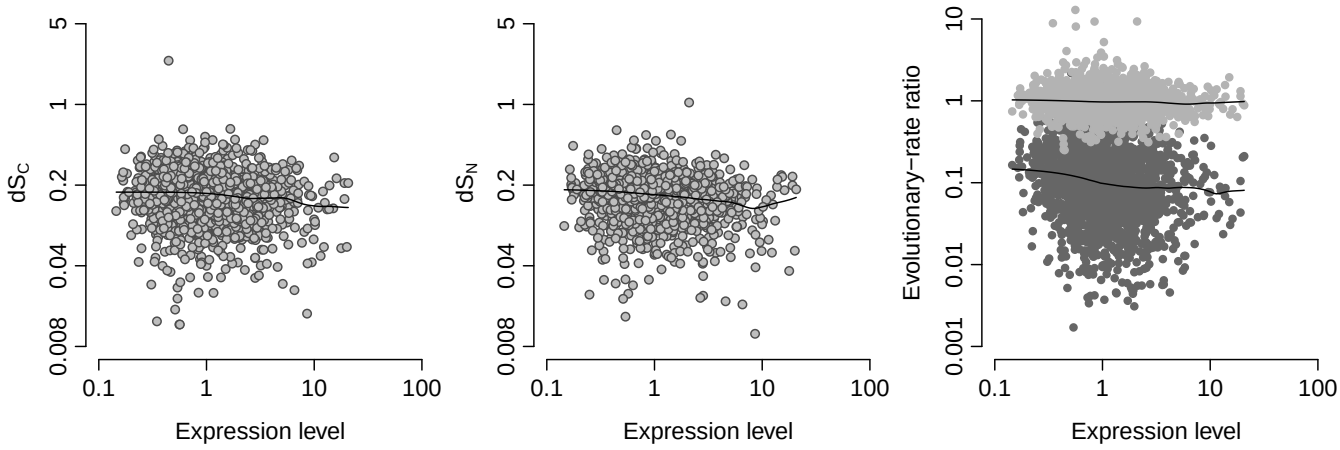

Figure S11: Synonymous evolutionary rates ($dS_{\rm C}$ and $dS_{\rm N}$) and evolutionary-rate ratios ($dN/dS_{\rm C}$ and $dS_{\rm N}/dS_{\rm C}$) calculated by the main model versus expression level for fly. Left panel: $dS_{\rm C}$ versus expression level. Middle panel: $dS_{\rm N}$ versus expression level. Right panel: evolutionary-rate ratios versus expression level. The Spearman correlations of expression level with $dS_{\rm C}$, $dS_{\rm N}$, $dN/dS_{\rm C}$, and $dS_{\rm N}/dS_{\rm C}$ are -0.116 ($P = 7.4 \times 10^{-8}$), -0.170 ($P = 2.2 \times 10^{-15}$), -0.203 ($P = 1.6 \times 10^{-21}$), and -0.054 ($P = 1.3 \times 10^{-2}$), respectively. Solid lines show lowess-smoothed data.

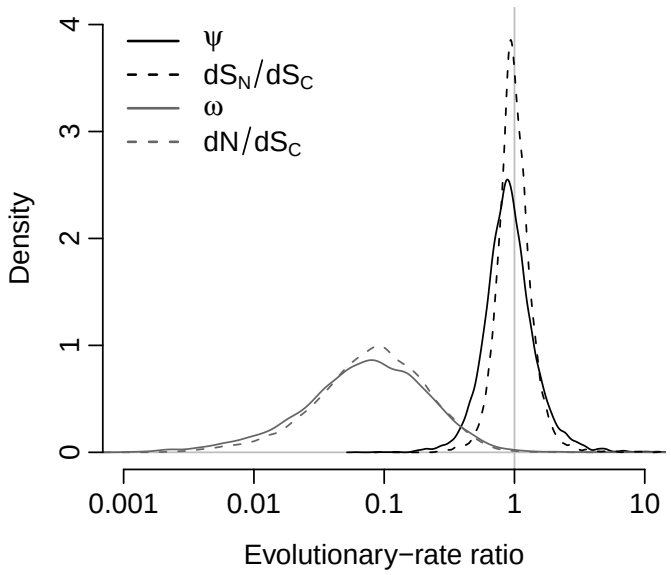

Figure S12: Distribution of evolutionary-rate ratios in fly.

References

Chan P P, Lowe T M. 2009. GtRNAdb: A database of transfer RNA genes detected in genomic sequence. Nucl. Acids Res. 37:D93–D97.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418–426.

Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg M F, Rifkin S A, Hua S, Herreman T, Tongprasit W, Barbano P E, Bussemaker H J, White K P. 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science 306:655–660.