Supplemental Data

Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution

D. Allan Drummond and Claus O. Wilke

Box S1: Quantifying rates of evolution

Rates of genetic change during the evolutionary divergence of two lineages can be inferred by examining differences between genes sequenced from representative organisms. First, orthologous genes (those tracing their ancestry to the same gene in the most-recent common ancestor) in each lineage are identified. Second, the encoded protein sequences are aligned and used to align the nucleotide sequence. Finally, rates of change can be estimated, as follows.

Genetic changes in coding sequences either preserve or alter the encoded amino acid, and are correspondingly called synonymous or nonsynonymous changes (**Figure B1**). Counting such changes gives the numerator of a per-site rate. Counting sites is slightly more involved; we employ a physical-sites definition (Bierne and Eyre-Walker, 2003). At some sites, such as the third position of the codon CGA (encoding arginine), all changes are synonymous, whereas at the second position, all changes are nonsynonymous, and at the first position, one change is synonymous (to AGA) and the other two are nonsynonymous—that is, the first position is 1/3 synonymous and 2/3 nonsynonymous. The sums of the (possibly fractional) synonymous and nonsynonymous sites over the length of a coding sequence then give the denominator of the per-site rates, computed as the number of nonsynonymous substitutions per nonsynonymous site (dN) and synonymous substitutions per synonymous site (dS).

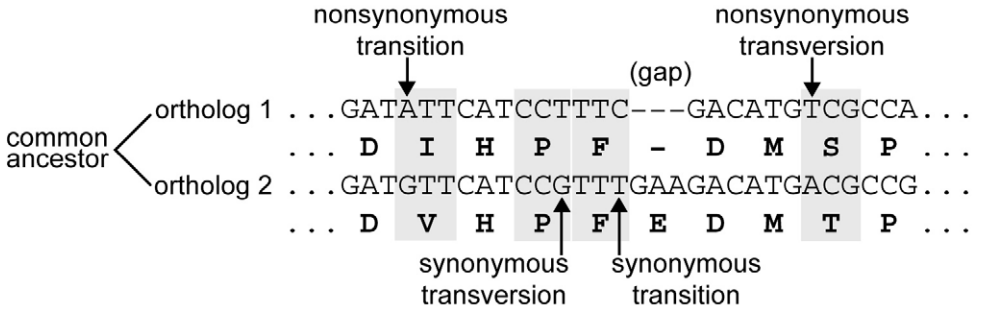

Figure B1: Anatomy of major evolutionary changes between related sequences.

The relative proportion of transitions (A–G or C–T) to transversions (all other changes) (**Figure B1**) is also used to quantify the types of mutations arising and fixing; both transition/transversion ratios (of numbers of each substitution type) and instantaneous rate ratios may be estimated (Wakeley, 1996).

After sequences have accumulated many substitutions, *e.g.* due to long periods of evolutionary divergence, multiple substitutions may have occurred at the same site. Sophisticated estimation methods attempt to infer these "multiple hit" events to yield proper estimates. Analysis of closely related organisms helps to minimize errors arising from improper inference.

Supplemental Results

The positive dS–dN/dS correlation

Here we demonstrate that a spurious dS–dN/dS correlation can be trivially constructed under the assumptions that 1) dN and dS are not independent, 2) dN and dS are roughly log-normally distributed, and 3) the relationship between dN and dS is roughly log-log linear. All three assumptions hold for all real organisms analyzed in the main text. Let $N(\mu,\sigma)$ be a normally distributed random variable with mean μ and standard deviation σ. Let dN and dS be given by:

$$dS = \exp[N(-1.5, 0.27)]$$

$$dN = dS^2 \exp[N(-1,0.7)]$$

Using the statistical program *R* (R Development Core Team, 2007), drawing 1000 genes from these distributions yields dN and dS with a mean and variance virtually identical to those observed between *S. cerevisiae* and *S. paradoxus*, a dN–dS correlation of $r = 0.54$***, and a dS–dN/dS correlation of $r = 0.28$***. *R* code to reproduce this result is as follows:

```
set.seed(9) # random number seed, for reproducibility
n <- 1000  # draw 1000 genes
ds <- exp(rnorm(n, mean=-1.5, sd=0.27))
dn <- ds^2 * exp(rnorm(n,mean=-1,sd=0.7))
print(cor.test(dn,ds,meth='s')) # spearman rank correlations
print(cor.test(dn/ds,ds,meth='s'))
```

This result shows that the non-independence of dN and dS, which is well-established and for which we provide an explanation in the main text, is sufficient to generate a strong positive dS–dN/dS correlation.

In short, the dS–dN/dS correlation is almost certainly an artifact of the dependence between dN and dS, and has no unique biological significance. The results of the simulation suggest that the non-independence of dN and dS can be explained by selection against mistranslation-induced misfolding.

Measuring codon bias in mammals

Throughout the main text, we use the fraction of optimal codons, F_{op}, to score the use of preferred codons in *E. coli*, yeast, worm, and fly, as is common practice. Use of F_{op} is questionable in human because it is clearly distorted by local nucleotide content. Most optimal codons are GC-ending in human, and GC content varies regionally across the genome; accordingly, as **Figure S7** shows, F_{op}, third-codon-position GC content, and

intronic GC content are bimodal in human. In mouse, this distortion is not evident to the eye. To control for these effects, we use the measure $F_{op}GC$ (see *Experimental Procedures*), which considers only codons which encode a four-fold or six-fold degenerate amino acid (L, V, S, P, T, A, R, or S, the only amino acids that possess both G- and C-ending codons) and end in G or C. In exactly this subset of codons, the third-position GC content is identical for all genes and synonymous codons ending in G or C are possible. As **Figure S7** shows, bimodality is largely removed from $F_{op}GC$. The measure is surely imperfect, but has the merit of simplicity. We use $F_{op}GC$ when analyzing both mouse and human, although in mouse, F_{op} is relatively well-behaved and yields mostly similar results in all our analyses (not shown). In the main text we carry out several more sophisticated tests to specifically detect selection for translational accuracy and determine whether these codons are indeed optimized for accurate translation. Use of $F_{op}GC$ in place of F_{op} for *E. coli*, yeast, worm, or fly yields essentially identical results in all our analyses (not shown), suggesting it remains a valid measure of preferential codon usage.

Confounding influences in mammals revealed by patterns of nucleotide composition

Two deviations in sign separate human from the other organisms: F_{op}–ts/tv-ratio ($r = -0.1$***) and F_{op}–dS ($r = 0.27$***) (**Figure 1B**), with the latter being a particularly radical departure. The F_{op}–dS deviation persists when evolutionary rates in human are measured relative to orthologous macaque genes rather than dog genes ($r = 0.25$***), implicating an ongoing effect in the primate lineage. The variance explained by the F_{op}–dS correlation is reduced more than 40-fold by statistically controlling for intronic guanosine and cytosine content (GC_i) using partial correlation ($r = 0.04$***), which also reverses the F_{op}–ts/tv-ratio correlation ($r = 0.06$***), suggesting both relationships are mediated by a process which acts on genomic regions, likely methylated CpG hypermutation.

To determine whether differences in the sets of genes being analyzed explain the apparent contrast between mouse and human, we compared only the 2,955 genes in our analysis that are present in both organisms; the human (dog) F_{op}–dS remained positive, $r = 0.28$***, and the mouse (rat) correlation remained negative, $r = -0.08$***. Neither can changes in codon usage explain the difference, because comparing the human gene's F_{op} to the mouse ortholog's dS yields $r = -0.05$**, while the mouse F_{op} and human dS correlate with $r = 0.27$***.

To understand the differences, we then examined the GC content of introns, which reports on mutational bias and gene conversion. In human, the dS–GC_i correlation ($r = 0.44$***) profoundly differs from mouse ($r = -0.01$), whereas a substantial F_{op}–GC_i correlation appears in both mouse and human ($r = 0.42$*** and 0.53***, respectively). The latter is expected because intronic and exonic nucleotide content correlate in most organisms and mammalian optimal codons almost exclusively end in G or C. Together, these results suggest that mutational bias or gene conversion have accelerated synonymous changes in the human lineage, creating a positive F_{op}–dS correlation as a side-effect.

Among the dominant mechanisms of composition-biased mutation in mammals is deamination of 5-methyl-cytosine-(phosphate)-guanine (mCpG) dinucleotides to yield TpG and a T-G mismatch which is repaired incorrectly half of the time to A-C on the

opposing strand for a net loss of CpG (and thus GC) content (Bird, 1980). GC content varies regionally in mammalian genomes, likely due to GC-biased gene conversion (Duret et al., 2006). Consequently, in GC-rich regions where CpG dinucleotides occur more frequently by chance, mCpG mutations will specifically destroy them. If observed CpG abundance relative to expectation based on single-nucleotide frequencies is quantified by the ratio of the proportion of CpG to the product of proportions of C and G, a measure we denote dCpG, the hypermutation hypothesis predicts a negative correlation between GC content and dCpG in species (namely vertebrates) which methylate CpGs. Computing GC proportion and dCpG using intronic sequences for orthologous mouse and human genes, we find dramatic dinucleotide depletion in human (GC_i–dCpG $r = -0.68$***), less in mouse ($r = -0.51$***), and no depletion in invertebrates (fly $r = 0.04$, worm $r = 0.25$***). Consistent with its lesser influence in mouse than in human, mCpG hypermutation appears less consequential for synonymous-site evolution. Elevated regional hypermutation should lead to accelerated CpG depletion and elevated synonymous-site evolution, predicting a negative correlation which is observed in (human dS–dCpG $r = -0.36$***) but not mouse ($r = 0.06$**).

Such divergent results confirm the expectation that patterns of evolutionary rate variation conserved across vertebrates and invertebrates are unlikely to arise from CpG hypermutation. At the same time, they confirm more extensive analyses on smaller datasets which indicate that CpG hypermutation drives primate synonymous-site evolution (Subramanian and Kumar, 2003). We have attempted to mute this effect in human correlations by controlling for intronic GC proportion, with the understanding that such controls are necessarily imperfect.

A unique fitness function describes protein misfolding costs

Let the fitness of an organism $f(m) > 0$ be a monotonically decreasing function of the amount of protein misfolding m with a continuous first derivative f' and $f(0) = 1$. Assume that misfolded protein is non-specifically toxic, such that any change Δm in the amount of misfolded protein produces the same fitness disadvantage $s = f(m + \Delta m) / f(m) - 1$. We claim these assumptions determine $f(m)$ up to a constant. *Proof*: We consider $\Delta m > 0$ without loss of generality. Consider two genes expressing amounts m_1 and m_2 of misfolded protein. Then:

$$s_1 = s_2$$
$$\frac{f(m_1 + \Delta m)}{f(m_1)} - 1 = \frac{f(m_2 + \Delta m)}{f(m_2)} - 1$$
$$\ln \frac{f(m_1 + \Delta m)}{f(m_1)} = \ln \frac{f(m_2 + \Delta m)}{f(m_2)}$$
$$g(m_1 + \Delta m) - g(m_1) = g(m_2 + \Delta m) - g(m_2)$$
$$g'(x_1) = g'(x_2)$$
$$g'(m) = c$$
$$g(m) = cm + d$$
$$f(m) = e^{cm+d}$$
$$f(m) = e^{-cm}$$

Note that in this model, polypeptides have no production cost, and misfolding does not impede the synthesis of a full complement of properly folded proteins. The only cost is the toxicity of misfolded proteins produced during synthesis.

Supplemental Experimental Procedures

Coding DNA sequences were built from coding exon sequences which were extracted, along with intronic sequences, from chromosomal DNA sequences, except for *E. coli* and yeast which were downloaded in pre-annotated gene format. Ortholog assignments were obtained from TIGR (Peterson et al., 2001) (*E. coli* and *S. typhimurium*, using reciprocal best BLAST hits with $P < 10^{-20}$), Ensembl's (Birney et al., 2006) BioMart homology track (human, dog, mouse, rat), WormBase's (Rogers et al., 2007) WormMart (worm), the *Saccharomyces* Genome Database (Hong et al., 2006) (yeast), and the *Drosophila* 12 Genomes Consortium AAAWiki website (Clark et al., 2007) (fly). Protein alignments were generated with MUSCLE 3.6 (Edgar, 2004) and used to align gene sequences, except for fly where alignments were downloaded from the AAAWiki website. A single cDNA per gene was randomly chosen from each gene that showed evidence of alternative splicing. Only cDNAs with 80% alignment to their ortholog, dS < 2 except as noted, and at least 30 codons were retained; final cDNA-ortholog pair counts (and those with mRNA expression data in parentheses) were: *E. coli* vs. *S. typhimurium*, 2,786 (2,229); *S. cerevisiae* vs. *S. paradoxus*, 4,616 (4,292); *C. elegans* vs. *C. briggsae*, 4,173 (2,386) (genes with dS < 4 were retained); *D. melanogaster* vs. *D. yakuba*, 7,070 (6,649); *M. musculus* vs. *R. norvegicus*, 9,061 (6,167); *H. sapiens* vs. *C. familiaris*, 5,939 (3,180).

Evolutionary rates and transition/transversion rate ratios were computed by maximum likelihood with PAML (Yang, 2006; Yang, 1997) using a physical-sites definition (Bierne and Eyre-Walker, 2003; Yang, 2006) operating on codons (codeml program) with the F3×4 codon frequency model, one dN/dS ratio per branch (model 0), and an arbitrary seed ts/tv rate ratio of 3.4. Ts/tv ratios were computed by counting transitions and transversions separating orthologous sequences, adding 1 to each (Laplace estimation), and taking their ratio. Distributions of dN and dS for all organisms are shown in **Supplementary Figure S5**.

We used previously reported mRNA levels for *E. coli* (Covert et al., 2004), yeast (Holstege et al., 1998), worm (Hill et al., 2000), fly (Chintapalli et al., 2007), mouse and human (Su et al., 2002). For *E. coli*, the geometric mean of four expression measurements under aerobic growth (ec_aer_wild_nO_[a-d]) was used. For worm, expression levels were as reported ((Hill et al., 2000), their Table 2A). For fly, mean levels in eleven adult tissues for probes with unambiguous FlyBase-gene-ID-to-probe matches were used. For human, the U133A and GNF1H array signals were merged and unambiguous Ensembl-peptide-ID-to-probe matches were retained; for mouse, the GNF1M array signal was used and unambiguous Ensembl-transcript-ID-to-probe matches were retained. Multiple signals for the same transcript were averaged. For fly, mouse and human, aggregate mRNA level was quantified as the geometric mean signal across all normal adult tissues.

Breadth of expression was computed using presence/absence calls. Tissue specificity was computed as described (Liao et al., 2006), except that all measurements were divided by a minimum expression value (0.1 for all species), zero values were set to this minimum value, and log-transformed expression values were multiplied by 1 if the gene was called present in that tissue and by 0 otherwise. Without the minimum-value adjustment, tissue specificity is not guaranteed to fall between 0 and 1 as asserted (Liao et al., 2006) because logarithms of values below 1 are negative.

We used published optimal codons for *E. coli* (Sharp and Li, 1987), yeast (Sharp and Cowe, 1991), worm (Sharp and Bradnam, 1997), fly (Duret and Mouchiroud, 1999) and human (Comeron, 2004). For mouse, optimal codons were defined as those corresponding to tRNAs with the highest gene counts in the set of 335 high-confidence tRNA genes identified by Waterston *et al*. (Waterston et al., 2002) (**Supplementary Table S2**). 5′ adenine in the anticodon was presumed to be quantitatively modified to inosine, which prefers to bind 3′ cytosine. The fraction of optimal codons, F_{op}, was calculated as described (Duret and Mouchiroud, 1999). In human and mouse we used the related measure $F_{op}GC$ (see **Supplementary Results** and **Supplementary Figure S7**).

Translational accuracy selection was first tested exactly as described using Akashi's test (Akashi, 1994); the resulting *Z*-score, when squared, yields the standard Mantel-Haenszel χ^2 statistic ((Sokal and Rohlf, 1995), p. 766). Sites with the same amino acid at the aligned position in the orthologous gene (or for the simulation, in all ancestral proteins on the line of descent) were designated conserved. In a second test, significance of the optimal–conserved association, randomized over the choice of optimal codon set, was assessed by computing the odds ratio for all possible alternate optimal codon sets which preserve the number of optimal codons per synonymous family in the naturally occurring set.

Gene dispensability in yeast was quantified using high-throughput growth-rate data from gene deletion strains grown under reference conditions (Warringer et al., 2003); because essential genes were not captured in these data, they were supplemented with essentiality data from an earlier study (Giaever et al., 2002). Fitness defect values *s* for each gene were computed as $s_i = \ln(r_i/r_{max})$, where r_i is the growth rate (fitness) of the strain having gene *i* knocked out and r_{max} is the maximum observed growth rate. This additive formulation for *s* yields values between negative infinity (for essential genes) and zero (complete dispensability). For the dispensability-fitness simulation, in which

fitness = $e^{-s(1-f)}$, s for essential genes was set to −10. Under this definition, the additive fitness effect of the knockout strain ($f = 0$) relative to wild type (assumed to have $f = 1$) is s, the dispensability. During the simulation timeframe, although some genes were almost completely dispensable, no genes were actually lost in the sense of accumulating mutations which critically destabilized the encoded protein.

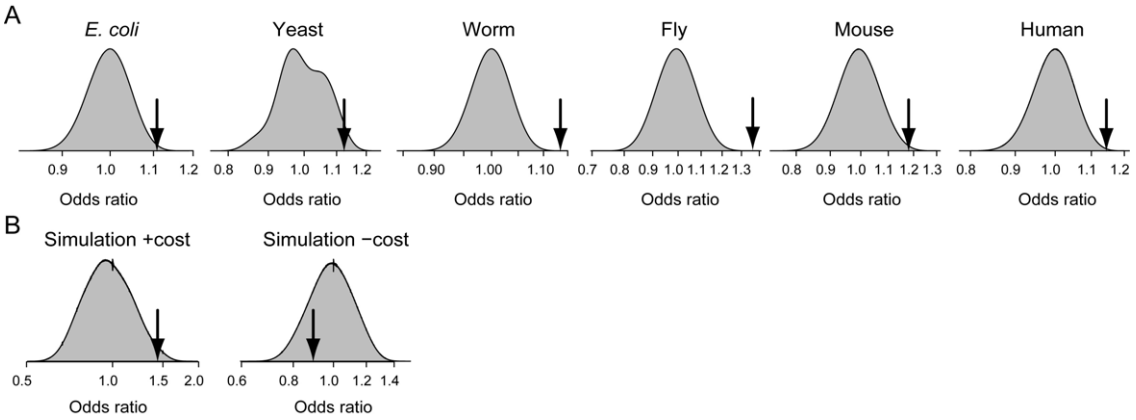

Figure S1. Distributions of the odds ratio for finding optimal codons at conserved sites, evaluated over all possible sets of codons that are synonymous with the established set of preferred codons. The odds ratio for the established set is indicated with an arrow. **A**, All organisms. **B**, The simulation, evolved with (left) and without (right) misfolding costs.

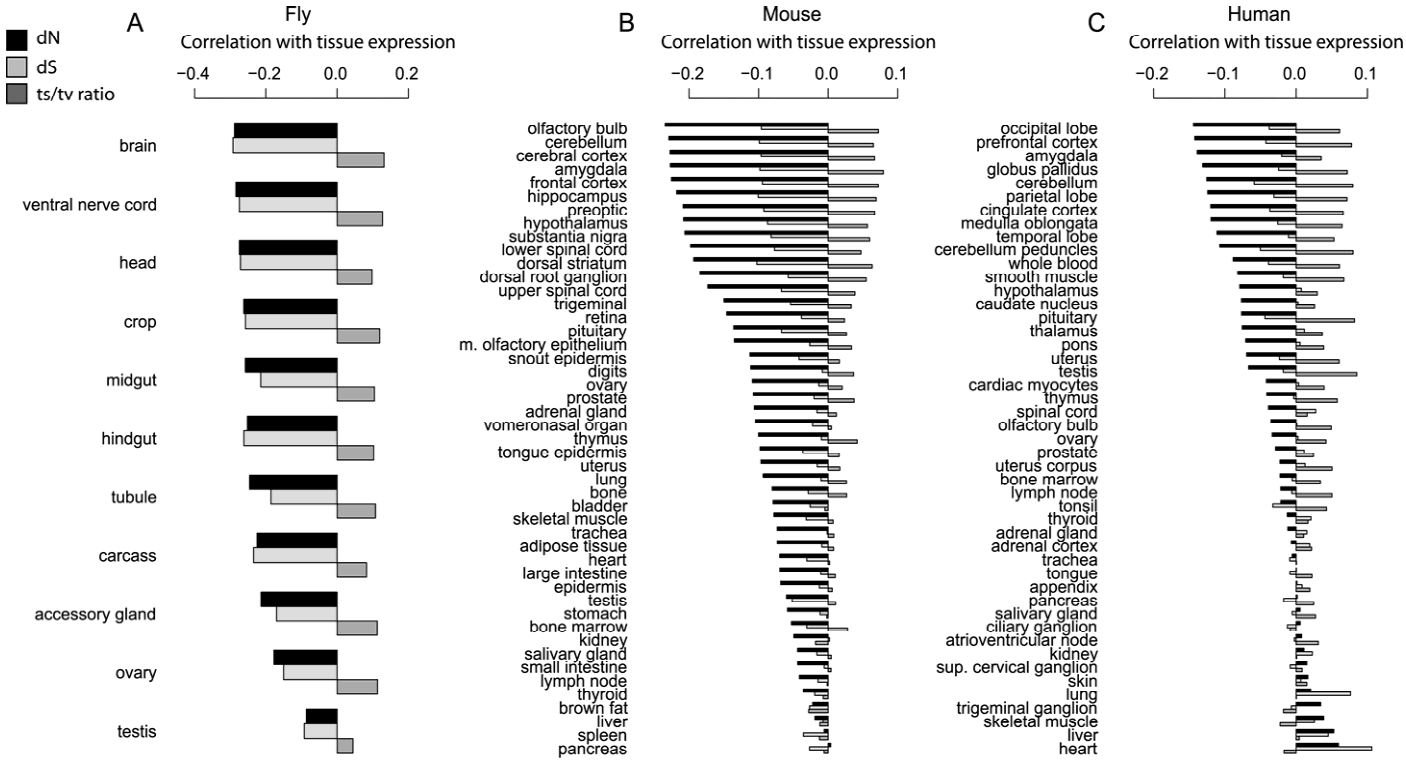

Figure S2: Correlations of per-tissue mRNA levels with dN, dS, and ts/tv ratio for fly (**A**), mouse (**B**), and human (**C**, controlled for intronic GC content) vary systematically across tissues when only genes with below-median tissue specificity of expression are considered; cf. **Figure 4**.

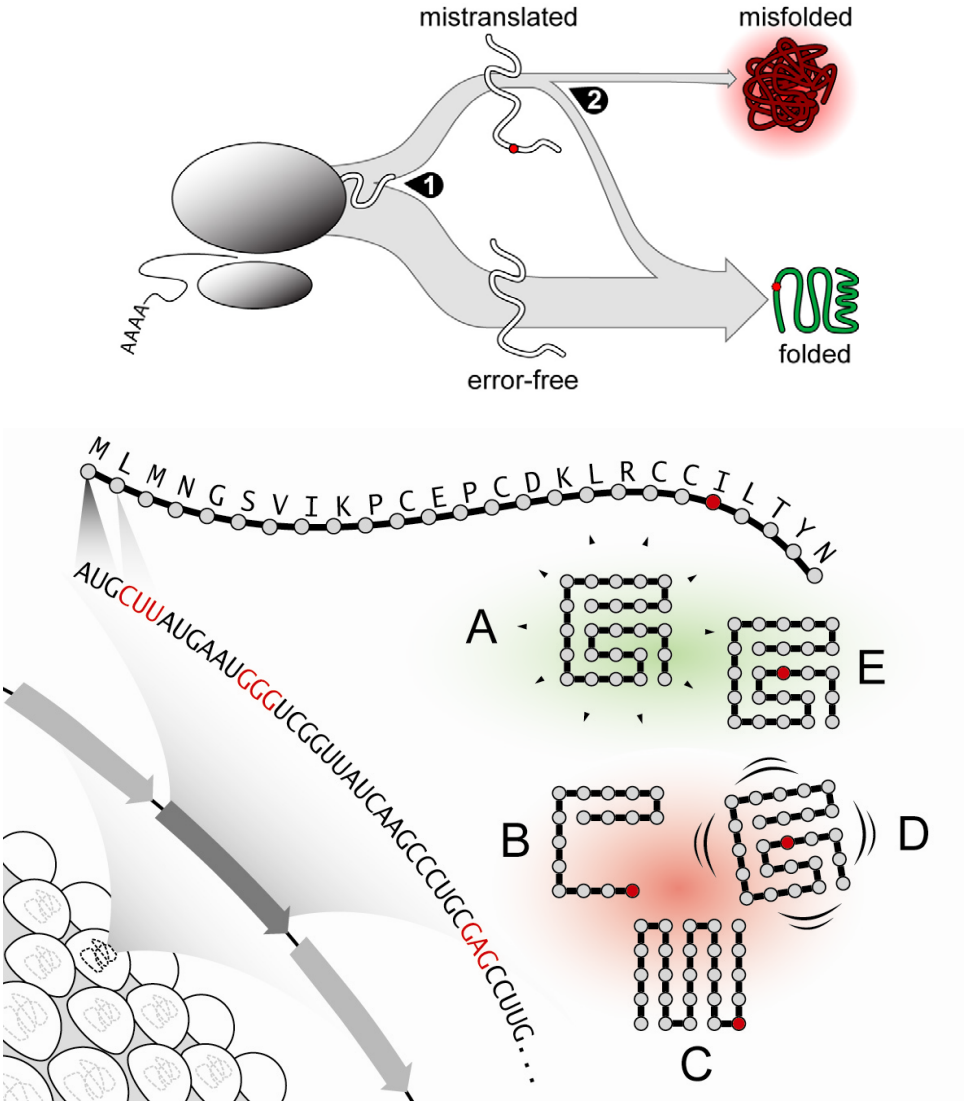

Figure S3. Overview of a large-scale simulation of genome evolution under selective pressure to avoid protein misfolding induced by mistranslation. Organisms (left) possess genomes consisting of 500 nucleotide sequences encoding 25-amino-acid polypeptides which fold according to a simple thermodynamic model. Proteins may translate and fold properly (**A**), or translate with at least one error, causing truncation (**B**), adoption of a non-native structure (**C**), folding to the native structure but with insufficient stability (**D**), or folding stably despite the error (**E**). Outcomes **B**, **C** and **D** are designated misfolded and impose a fitness cost. Certain codons are translated with higher error rates (red codons).

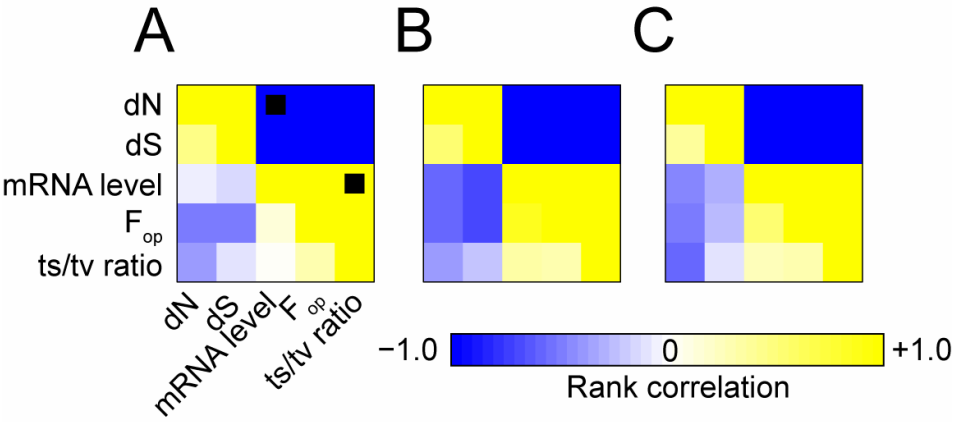

Figure S4. Results of evolutionary simulation when genes have different dispensabilities (growth-rate defects upon knocking out the gene) suggests that loss of functional molecules, rather than gain of costly molecules, is unlikely to explain the observed organismal patterns. **A**, Correlations between the five analyzed variables after evolution under a fitness function in which each gene has an individual growth-rate defect of knocking out the gene, s, and fitness = $\exp[s(1-f)]$ with f the fraction of folded proteins. As in Figure 1A, lower triangles show correlation coefficients, upper triangles show signs, and black squares indicate insignificant correlations ($P > 0.05$). **B**, The simulation carried out from identical initial conditions as in **A**, but with the original translational fitness cost. **C**, Results from yeast (cf. **Figure 1A**) for comparison.

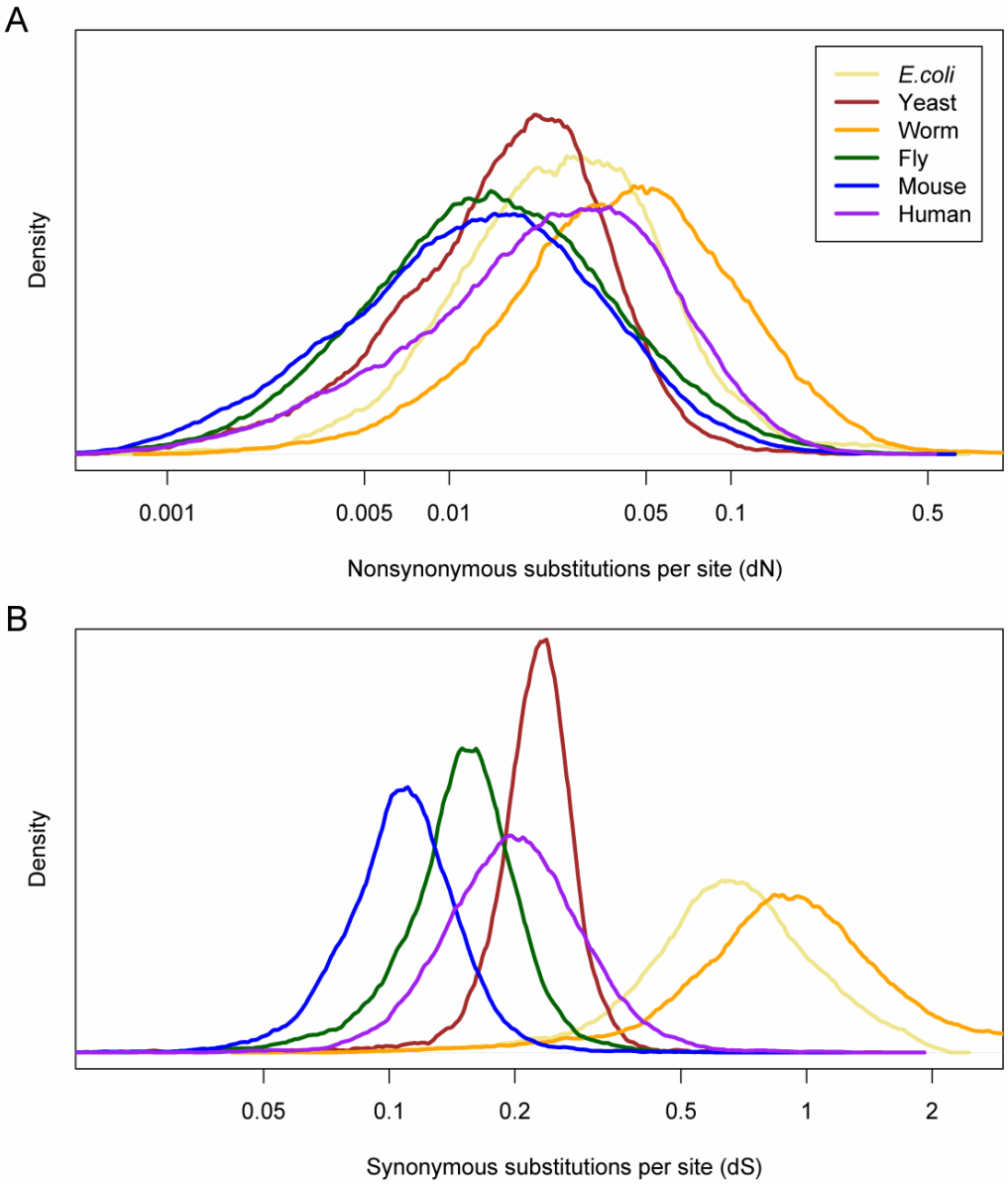

Figure S5: Evolutionary rate distributions for the six organisms analyzed. **A**, Nonsynonymous substitutions per nonsynonymous physical site (dN); **B**, Synonymous substitutions per synonymous physical site (dS).

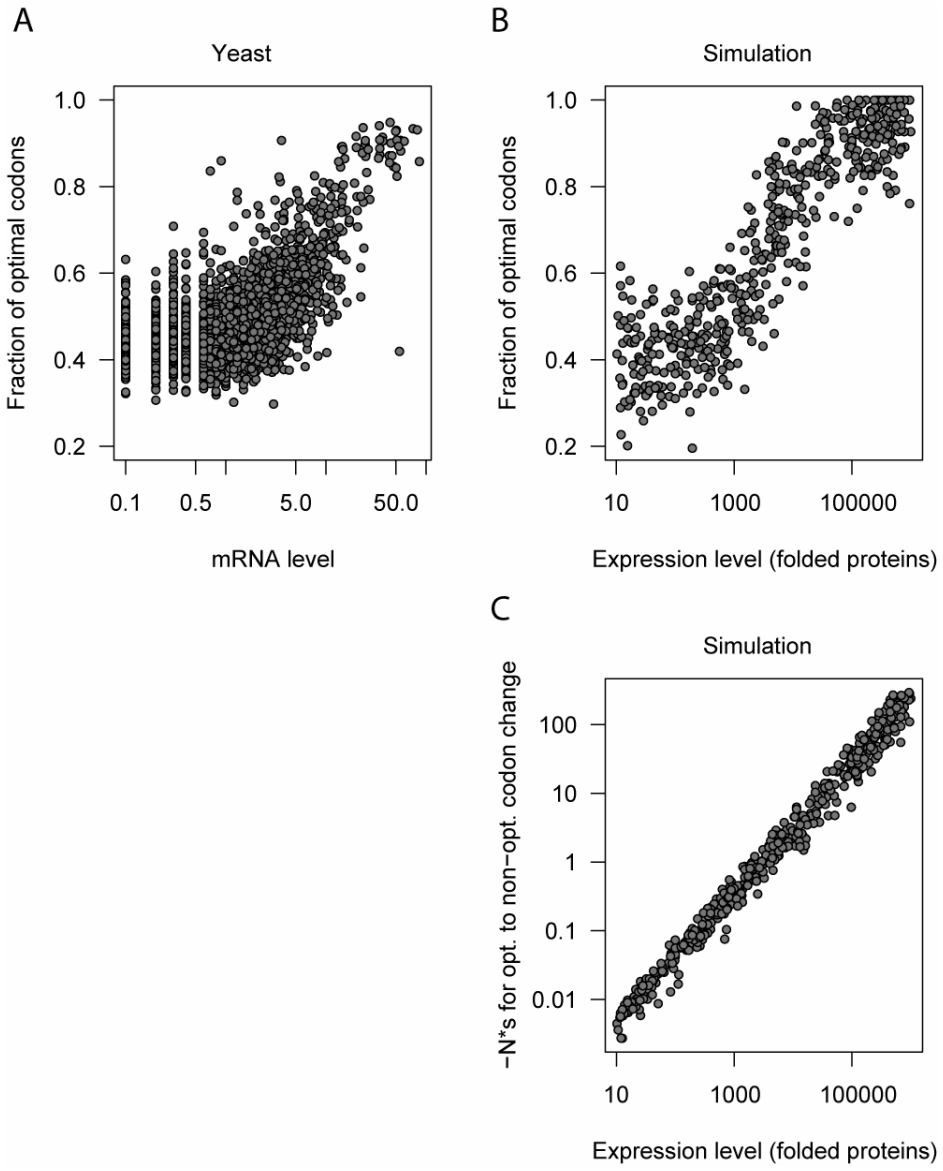

Figure S6. How the selective disadvantage of switching an optimal codon to a non-optimal codon changes as a function of expression level in the simulation. **A**, Expression level versus the fraction of optimal codons in yeast. **B**, Expression level versus the fraction of optimal codons in the simulation. **C**, The average selective disadvantage $s = 1 - f_{\text{unopt}}/f_{\text{opt}}$ (where f is the growth rate), multiplied by the population size (N = 1,000), induced by the change of one codon from optimal to non-optimal in each simulated gene, as a function of expression level. If we infer from the qualitative agreement of the simulation and yeast mRNA-level–F_{op} curves that the scaled selective disadvantages Ns are roughly equivalent, then the selective disadvantage of a single codon change in yeast may be estimated. The estimated effective population size for yeast is $N = 10^7$–10^8 (Lynch and Conery, 2003), and the highest-expressed gene should have $Ns \sim -100$. This yields a disadvantage of $s \sim -10^{-4}$ to -10^{-5} (*i.e.*, a mutant strain bearing a non-optimal codon in one of the most highly expressed genes would be expected to grow 0.01–0.001% slower than wild type).

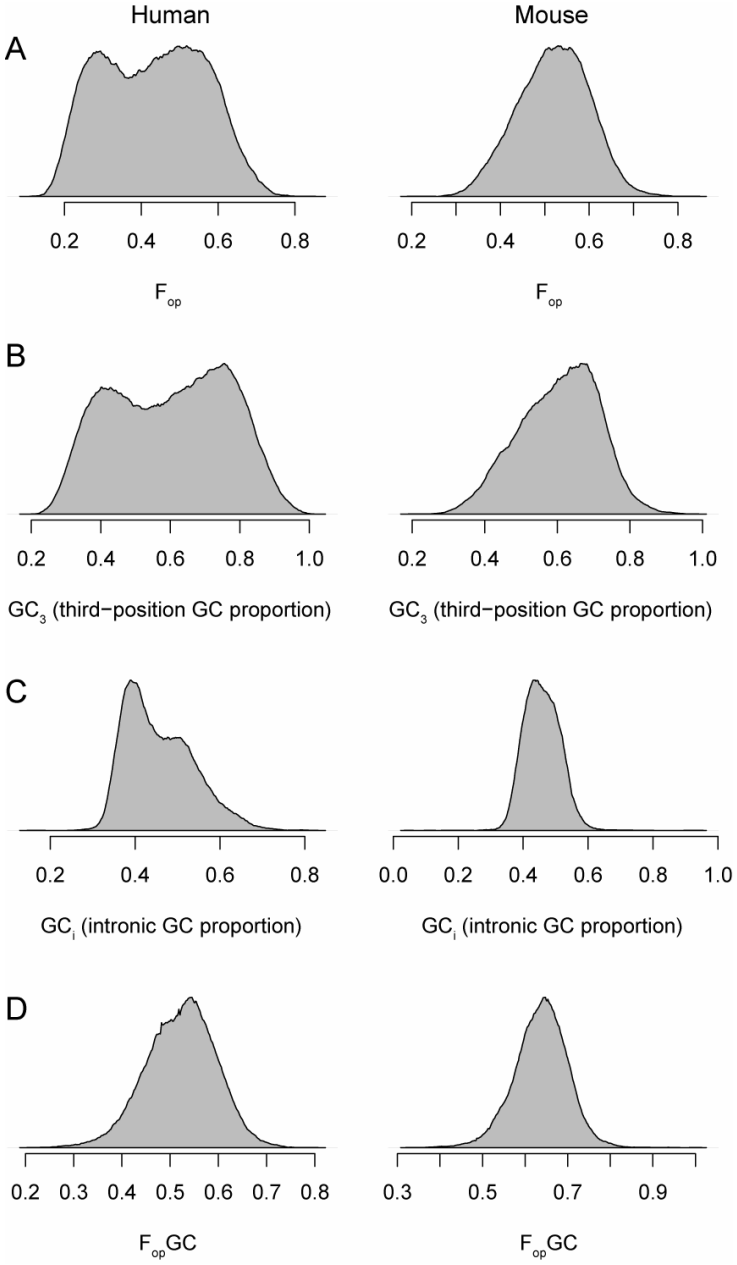

Figure S7: $F_{op}GC$ (fraction of optimal codons, controlled for guanine+cytosine content) largely eliminates major problems with F_{op} as a measure of mammalian codon bias. **A**, Measures in human of fraction of optimal codons show bimodal behavior. **B**, Third-position guanine+cytosine (GC) content mirrors the bimodality, as does intronic GC content (**C**), suggesting that the bimodality of the previous two measures reflects regional biases in nucleotide content. **D**, The $F_{op}GC$ measure largely eliminates bimodality due to nucleotide-content biases. Left, human; right, the same measures in mouse.

Table S1. Spearman rank correlations between nonsynonymous- and synonymous-site substitution rates (dN and dS), mRNA expression level, fraction of optimal codons (F_{op}), and transition/transversion (ts/tv) ratio and codon bias for yeast, fly, mouse, human and simulated organisms.

Correlation	*E. coli*	Yeast	Worm	Fly	Mouse	Human
dN–dS	0.641***	0.372***	0.538***	0.475***	0.329***	0.473***
dN–mRNA-level	−0.389***	−0.462***	−0.278***	−0.384***	−0.190***	−0.163***
dN–F_{op}	−0.542***	−0.511***	−0.277***	−0.444***	−0.090***	0.017
dN–ts/tv−ratio	−0.694***	−0.571***	−0.768***	−0.583***	−0.564***	−0.591***
dS–mRNA-level	−0.441***	−0.297***	−0.352***	−0.198***	−0.043***	−0.016
dS– F_{op}	−0.555***	−0.252***	−0.506***	−0.297***	−0.107***	0.269***
dS–ts/tv−ratio	−0.505***	−0.114***	−0.442***	−0.181***	−0.136***	−0.373***
mRNA-level–F_{op}	0.398***	0.530***	0.553***	0.281***	0.020	0.085***
mRNA-level–ts/tv−ratio	0.299***	0.248***	0.299***	0.296***	0.095***	0.076***
F_{op}–ts/tv−ratio	0.483***	0.284***	0.299***	0.201***	0.069***	−0.103***

(cont.)

Correlation	Human | GC	Sim. +cost	Sim. −cost	Sim. +noise
dN–dS	0.460***	0.668***	−0.079	0.668***
dN–mRNA-level	−0.173***	−0.682***	0.053	−0.070
dN–F_{op}	−0.065***	−0.711***	0.002	−0.140**
dN–ts/tv−ratio	−0.591***	−0.518***	−0.176***	−0.518***
dS–mRNA-level	−0.059**	−0.789***	−0.037	−0.100*
dS–F_{op}	0.042**	−0.817***	−0.038	−0.108*
dS–ts/tv−ratio	−0.293***	−0.423***	0.238***	−0.423***
mRNA-level–F_{op}	0.051**	0.897***	0	0.060
mRNA-level–ts/tv−ratio	0.095***	0.463***	0.010	0.099*
F_{op}–ts/tv−ratio	0.060***	0.479***	0.075	0.083

* = $P < 0.05$; ** = $P < 0.01$; *** = $P < 0.001$; all significance levels after false-discovery-rate correction for multiple testing.

Table S2. Optimal codons identified from tRNA gene copy numbers in mouse (*Mus musculus*) tabulated by Waterston *et al*. (Waterston et al., 2002) Codons corresponding to the most-abundant tRNA species per family were designated optimal (*). 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 .

Amino acid	Anti-codon	Cognate codon	tRNA gene copy number	Amino acid	Anti-codon	Cognate codon	tRNA gene copy number
A	IGC	GCC*	12	N	GTT	AAC*	11
	TGC	GCA	4		ITT	AAC	0
	CGC	GCG	3	P	IGG	CCC*	5
	GGC	GCC	0		TGG	CCA	4
C	GCA	UGC*	50		CGG	CCG	2
	ICA	UGC	0		GGG	CCC	0
D	GTC	GAC*	14	Q	CTG	CAG*	8
	ITC	GAC	0		TTG	CAA	5
E	CTC	GAG*	8	R	ICG	CGC*	6
	TTC	GAA*	8		TCG	CGA	5
F	GAA	UUC*	7		CCT	AGG	5
	IAA	UUC	0		TCT	AGA	5
G	GCC	GGC*	12		CCG	CGG	2
	TCC	GGA	7		GCG	CGC	0
	CCC	GGG	2	S	IGA	UCC*	7
	ICC	GGC	0		GCT	AGC*	7
H	GTG	CAC*	9		CGA	UCG	3
	ITG	CAC	0		TGA	UCA	3
I	IAT	AUC*	11		GGA	UCC	0
	TAT	AUA	4		ICT	AGC	0
	GAT	AUC	0	T	IGT	ACC*	8
K	CTT	AAG*	9		CGT	ACG	4
	TTT	AAA*	9		TGT	ACA	4
L	CAG	CUG*	8		GGT	ACC	0
	IAG	CUC	5	V	CAC	GUG*	7
	CAA	UUG	4		IAC	GUC	6
	TAG	CUA	3		TAC	GUA	3
	TAA	UUA	2		GAC	GUC	0
	GAG	CUC	0	W	CCA	UGG	8
M	CAT	AUG	15	Y	GTA	UAC*	10
					ITA	UAC	0

Table S3. Probabilities of a translation error for all 61 sense codons in the simulation (structure 699), along with optimal codon designations (*). Probability of error was estimated by the fraction of 10,000 translation events yielding anything other than the encoded amino acid. Fold-differences in error frequency are computed relative to all synonymous codons (syn) or all other codons (all).

Amino acid	Codon	Pr(error)	Fold diff. (syn)	Fold diff. (all)	Amino acid	Codon	Pr(error)	Fold diff. (syn)	Fold diff. (all)
A	GCA	0.00714	5.0	5.5	N	AAC*	0.00256	1.0	2.0
	GCC	0.00738	5.2	5.7		AAU	0.01263	4.9	9.8
	GCG	0.00721	5.1	5.6	P	CCA*	0.00157	1.0	1.2
	GCU*	0.00142	1.0	1.1		CCC	0.00704	4.5	5.5
C	UGC	0.01352	5.9	10.5		CCG	0.00731	4.7	5.7
	UGU*	0.00229	1.0	1.8		CCU	0.00747	4.8	5.8
D	GAC*	0.00226	1.0	1.8	Q	CAA*	0.00237	1.0	1.8
	GAU	0.01265	5.6	9.8		CAG	0.01222	5.2	9.5
E	GAA*	0.00264	1.0	2.0	R	AGA*	0.00212	1.0	1.6
	GAG	0.01214	4.6	9.4		AGG	0.01143	5.4	8.9
F	UUC*	0.00232	1.0	1.8		CGA	0.00453	2.1	3.5
	UUU	0.01259	5.4	9.8		CGC	0.00699	3.3	5.4
G	GGA	0.00687	4.9	5.3		CGG	0.00560	2.6	4.3
	GGC	0.00711	5.1	5.5		CGU	0.00709	3.3	5.5
	GGG	0.00700	5.0	5.4	S	AGC	0.01246	8.9	9.7
	GGU*	0.00139	1.0	1.1		AGU	0.01257	9.0	9.7
H	CAC*	0.00243	1.0	1.9		UCA	0.00687	4.9	5.3
	CAU	0.01268	5.2	9.8		UCC*	0.00156	1.1	1.2
I	AUA	0.00981	5.7	7.6		UCG	0.00714	5.1	5.5
	AUC	0.00205	1.2	1.6		UCU*	0.00140	1.0	1.1
	AUU*	0.00172	1.0	1.3	T	ACA	0.00705	5.5	5.5
K	AAA	0.01315	5.8	10.2		ACC*	0.00129	1.0	1.0
	AAG*	0.00228	1.0	1.8		ACG	0.00718	5.6	5.6
L	CUA	0.00434	2.1	3.4		ACU	0.00138	1.1	1.1
	CUC	0.00713	3.5	5.5	V	GUA	0.00742	4.8	5.8
	CUG	0.00483	2.4	3.7		GUC*	0.00155	1.0	1.2
	CUU	0.00736	3.6	5.7		GUG	0.00765	4.9	5.9
	UUA	0.00990	4.9	7.7		GUU*	0.00159	1.0	1.2
	UUG*	0.00204	1.0	1.6	W	UGG	0.01498	1.0	11.6
M	AUG	0.01601	1.0	12.4	Y	UAC*	0.00255	1.0	2.0
						UAU	0.01153	4.5	8.9

Table S4: Rank correlations between evolutionary rates and aggregate levels or patterns of expression.

Organism	Dependent variable	Expression mean (arithmetic)	Expression mean (geometric)	Expression breadth	Tissue specificity
Fly	dN	−0.246***	−0.384***	−0.357***	0.399***
	dS	−0.119***	−0.198***	−0.157***	0.216***
Mouse	dN	−0.169***	−0.190***	−0.178***	0.209***
	dS	−0.039**	−0.043***	−0.016	0.020
Human	dN	−0.126***	−0.163***	−0.224***	0.235***
	dS	0.018	−0.017	−0.127***	0.149***
Human | GC_i[a]	dN	−0.135***	−0.173***	−0.209***	0.221***
	dS	−0.029	−0.060**	−0.084***	0.108***

* = $P < 0.05$; ** = $P < 0.01$; *** = $P < 0.001$; all significance levels after false-discovery-rate correction for multiple testing.

[a]Partial correlations, controlling for intronic guanine and cytosine proportion.

Supplemental References

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics *136*, 927-935.

Bierne, N., and Eyre-Walker, A. (2003). The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. Genetics *165*, 1587-1597.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res *8*, 1499-1504.

Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T.*, et al.* (2006). Ensembl 2006. Nucleic Acids Res *34*, D556-561.

Chintapalli, V.R., Wang, J., and Dow, J.A. (2007). Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet *39*, 715-720.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N.*, et al.* (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203-218.

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

Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., and Palsson, B.O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. Nature *429*, 92-96.

Duret, L., Eyre-Walker, A., and Galtier, N. (2006). A new perspective on isochore evolution. Gene *385*, 71-74.

Duret, L., and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci U S A *96*, 4482-4487.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B.*, et al.* (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*, 387-391.

Hill, A.A., Hunter, C.P., Tsung, B.T., Tucker-Kellogg, G., and Brown, E.L. (2000). Genomic analysis of gene expression in C. elegans. Science *290*, 809-812.

Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. Cell *95*, 717-728.

Hong, E., Balakrishnan, R., Christie, K., Costanzo, M., Dwight, S., Engel, S., Fisk, D., Hirschman, J., Livstone, M., Nash, R.*, et al.* (2006). *Saccharomyces* Genome Database, ftp://ftp.yeastgenome.org/yeast/.

Liao, B.Y., Scott, N.M., and Zhang, J. (2006). Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol *23*, 2072-2080.

Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. Science *302*, 1401-1404.

Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K., and White, O. (2001). The Comprehensive Microbial Resource. Nucleic Acids Res *29*, 123-125.

R Development Core Team (2007). R: A Language and Environment for Statistical Computing (Vienna, Austria, R Foundation for Statistical Computing).

Rogers, A., Antoshechkin, I., Bieri, T., Blasiar, D., Bastiani, C., Canaran, P., Chan, J., Chen, W.J., Davis, P., Fernandes, J.*, et al.* (2007). WormBase 2007. Nucleic Acids Res.

Sharp, P.M., and Bradnam, K.R. (1997). Codon usage in C. elegans. In C elegans II, D.L. Riddle, T. Blumenthal, B.J. Meyer, and J.R. Priess, eds. (Cold Spring Harbor Laboratory Press).

Sharp, P.M., and Cowe, E. (1991). Synonymous codon usage in Saccharomyces cerevisiae. Yeast *7*, 657-678.

Sharp, P.M., and Li, W.H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res *15*, 1281-1295.

Sokal, R.R., and Rohlf, F.J. (1995). Biometry, 3 edn (New York, W. H. Freeman and Co.).

Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A.*, et al.* (2002). Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A *99*, 4465-4470.

Subramanian, S., and Kumar, S. (2003). Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res *13*, 838-844.

Wakeley, J. (1996). The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends Ecol Evol *11*, 158-162.

Warringer, J., Ericson, E., Fernandez, L., Nerman, O., and Blomberg, A. (2003). High-resolution yeast phenomics resolves different physiological features in the saline response. Proc Natl Acad Sci U S A *100*, 15724-15729.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P.*, et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Yang, Z. (2006). Computational Molecular Evolution (Oxford, UK, Oxford University Press).

Yang, Z.H. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences *13*, 555-556.