bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Natural selection on gene-specific codon usage bias is common across eukaryotes

Zhen Peng[1], Hani Zaher[1], Yehuda Ben-Shahar[1,*]

[1]Department of Biology, Washington University in St. Louis, MO 63130, USA.

[*]Corresponding Author: Dr. Yehuda Ben-Shahar, Department of Biology, Washington University in St. Louis, MO 63130, USA. benshahary@wustl.edu

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SUMMARY

Although the actual molecular evolutionary forces that shape differences in codon usage across species remain poorly understood, majority of synonymous mutations are assumed to be functionally neutral because they do not affect protein sequences. However, empirical studies suggest that some synonymous mutations can have phenotypic consequences. Here we show that in contrast to the current dogma, natural selection on gene-specific codon usage bias is common across Eukaryota. Furthermore, by using bioinformatic and experimental approaches, we demonstrate that specific combinations of rare codons contribute to the spatial and sex-related regulation of some protein-coding genes in *Drosophila melanogaster*. Together, these data indicate that natural selection can shape gene-specific codon usage bias, which therefore, represents an overlooked genomic feature that is likely to play an important role in the spatial and temporal regulation of gene functions. Hence, the broadly accepted dogma that synonymous mutations are in general functionally neutral should be reconsidered.

KEYWORDS: *Drosophila melanogaster;* fruit fly; genome evolution; synonymous mutation; neutral mutation

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

Most amino acids are encoded by at least two interchangeable synonymous DNA codons (Nirenberg and Leder, 1964; Crick, 1966; Söll et al., 1966). Although synonymous codons do not impact the amino-acid sequences of proteins, analyses of genomic DNA indicate that protein-coding genes often exhibit codon usage bias, which is defined as usage frequencies of specific synonymous codons that deviate from expected random usage frequencies (Ikemura, 1981; Sharp et al., 1988; Plotkin and Kudla, 2011; Doherty and McInerney, 2013). While the specific patterns of codon usage bias often vary across species, this phenomenon seems to be a universal feature of most protein-coding DNA sequences (Sharp et al., 1988; Kanaya et al., 1999; Wang and Hickey, 2007; Quax et al., 2015; Pouyet et al., 2016). Nevertheless, the actual molecular forces that shape gene-specific codon usage bias at the evolutionary timescale remain poorly understood.

Previous theoretical models derived from population genetics suggested that observed codon usage frequencies are the result of an interplay between three forces: 1) Mutational bias, which refers to the asymmetric occurrence of mutations; 2) Genetic drift, whereby specific synonymous codons are randomly fixed; 3) Natural selection for favored specific synonymous codon usage patterns associated with specific gene functions. This theoretical framework was mathematically formalized as the “selection-mutation-drift model” (Bulmer, 1991; Yang and Nielsen, 2008). Although the “selection-mutation-drift model” does not assign different mathematical weights to these three forces, it is generally assumed that mutational bias is likely the dominant force in shaping codon usage bias across diverse species (Stenico, Lloyd, and Sharp, 1994; Chen et al., 2004; Fu, 2010; Sharp, Emery, and Zeng, 2010; Rao et al., 2011; Doherty and McInerney, 2013; Yang et al., 2015). In contrast, whether genetic drift and natural

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

selection play equally important roles in driving gene-specific codon usage patterns remains controversial (Akashi, 1995; Rao et al., 2011; Doherty and McInerney, 2013; Kessler and Dean, 2014). Specifically, since synonymous mutations do not affect protein sequences, and therefore are usually assumed to have no impact on their structure or biochemical function, the effect of natural selection on gene-specific codon usage bias is often dismissed as weak (King and Jukes, 1969; Kimura, 1968; Kimura and Ohta, 1974; Echave, Spielman, and Wilke, 2016).

Nonetheless, a series of biochemical and molecular studies of specific protein-coding genes seem to present mechanistic challenges to the broadly accepted dogma that synonymous mutations are generally functionally neutral. First, gene-specific synonymous codon usage patterns have been shown to be important for the function of some protein-coding genes by affecting features such as translation efficacy (Ikemura, 1981; Akashi, 1994; Ko et al., 2005), mRNA secondary structures (Hasegawa, Yasunaga, and Miyata, 1979), mRNA splicing patterns (Pagani, Raponi, and Baralle, 2005; Parmley and Hurst, 2007), miRNA binding accessibility (Gu et al., 2012), protein folding (Zhang, Hubalewska, and Ignatova, 2009; Zhou et al., 2013; Fu et al., 2016), and post-translational modifications (Kimchi-Sarfaty et al., 2007; Zhou et al., 2013; Quax et al., 2015; Fu et al., 2016). Second, the temporal regulation of protein synthesis in several viruses in the *Herpesviridae* group has been shown to depend on viral-specific codon usage patterns that are different from those of the host cell (Shin et al., 2015). Third, high protein levels of some housekeeping genes depend on the biased usage of common synonymous codons (Stenico, Lloyd, and Sharp, 1994; Ma et al., 2014). Some of these anecdotal examples could be explained by the “translational selection theory”, which stipulates that “optimal” protein translation rate and accuracy partly depend on the availability and affinity of cognate tRNAs, and therefore, necessitates the evolution of

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

codon usage patterns that match the relative abundance of specific tRNAs in the relevant cellular pools (Ikemura, 1981; Shields and Sharp, 1987; Slimko and Lester, 2003; dos Reis, Savva, and Wernisch, 2004). Together, these studies suggest that gene-specific codon usage biases might be playing a general, yet underappreciated, role in the spatial and temporal regulation of gene activity, and therefore should be affected by natural selection.

Here we combine theoretical and experimental approaches to test the hypothesis that natural selection has played a significant and important role in shaping gene-specific codon usage patterns. To test this hypothesis, we developed a statistical approach for identifying signatures of natural selection on gene-specific codon usage patterns. Using this approach, we demonstrate that a considerable proportion of protein-coding genes from diverse eukaryotic species display signatures of natural selection on biased gene-specific codon usage patterns. Additionally, we use the well-annotated genome of *Drosophila melanogaster* to identify genome-wide signatures of selection on gene-specific codon usage patterns and demonstrate that a specific combination of unique rare synonymous codons are associated with specific biological functions, and specifically enhance protein expression in the male reproductive system. Together, our results indicate that in contrast to the broadly accepted dogma, natural selection plays a significant role in determining gene-specific codon usage biases in diverse eukaryotic species, and that gene-specific codon usage biases are important for the spatial and temporal regulation of some protein-coding genes. Together, these data indicate that the broadly accepted assumption that synonymous mutations are in general functionally and evolutionarily neutral should be empirically evaluated on gene-specific bases.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RESULTS

Signatures of natural selection on gene-specific codon usage are common across eukaryotes

Testing the hypothesis that gene-specific codon usage bias plays a general role in regulating gene function requires a broadly applicable method for statistically identifying signatures of selection across protein-coding genes throughout genomes. Mathematically, the previously published "selection-mutation-drift model" could be theoretically used for identifying such signatures (Bulmer, 1991). However, applying this specific model to diverse species in order to identify signatures of natural selection on codon usage bias requires *a priori* empirical information about population-level genetic variations, temporal and spatial gene expression patterns, and related estimates of fitness (Yang and Nielsen, 2008; Rodrigue, Philippe and Lartillot, 2010; Tamuri, dos Reis and Goldstein, 2012; Gilchrist et al., 2015; Kubatko et al. 2016). Since such data are not available for most species, we developed an alternative statistical approach that is solely based on the analyses of DNA sequences of protein-coding genes from publicly available genomic data. In contrast to the previously published "selection-mutation-drift" model, which evaluates the relative contribution of natural selection to the observed codon usage bias by estimating selection coefficients for each possible nucleotide substitution, our approach is based on the statistical rejection of the null hypothesis that "Synonymous mutations are neutral". Although this approach does not provide specific quantitative estimates for the level of selection on each individual gene, it serves as a robust method for identifying specific genes whose biased codon usage patterns have been impacted by natural selection.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Our statistical model is based on several key assumptions: 1) Reference genomic data represent a "wild type" genome; 2) Observed gene-specific codon usage patterns are at equilibrium; 3) For each codon, no more than one nucleotide substitution per generation is possible; 4) Nonsynonymous mutations are mostly deleterious, and therefore, the probabilities of observed nonsynonymous substitutions are negligible relative to those of synonymous ones; 5) The contributions of individual alleles of a single gene to fitness are additive; 6) Gene-specific mutational biases are independent, which allows us to ignore the possible effects of background nucleotide composition of nearby chromosomal regions on the mutational bias of focal nucleotides (Oliver et al, 2001); 7) For each individual open reading frame, the mutation rate for each of the 12 possible nucleotide substitutions (A-to-T, T-to-A, *etc.*) is constant (Rodriguez et al., 1990), which also indicates that the possible effects of adjacent nucleotides on mutation rates of each focal nucleotide could be omitted (Bérard and Guéguen, 2012); 8) The lowest mutation rate is at least 1/100 of the highest one (Drake et al., 1998).

Next, we used the previously published “selection-mutation-drift model” (Bulmer, 1991; Yang and Nielsen, 2008) to describe the relationships between the various evolutionary molecular processes that may have shaped observed gene-specific codon usage biases (Equation (1)).

$$\sum_i \frac{2s_{pqi\to pqr}}{1-e^{-2Ns_{pqi\to pqr}}} \times x_{pqi} \times \mu_{i\to r} + \sum_j \frac{2s_{jqr\to pqr}}{1-e^{-2Ns_{jqr\to pqr}}} \times x_{jqr} \times \mu_{j\to p} = \sum_i \frac{2s_{pqr\to pqi}}{1-e^{-2Ns_{pqr\to pqi}}} \times x_{pqr} \times \mu_{r\to i} + \sum_j \frac{2s_{pqr\to jqr}}{1-e^{-2Ns_{pqr\to jqr}}} \times x_{pqr} \times \mu_{p\to j} \quad (1)$$

Variables *p*, *q*, and *r* denote the specific nucleotide identities (A, C, G, or T) of the first, second, and third positions respectively in each codon. Codons synonymous to codon *pqr* that vary at either the first or third positions are denoted by codons *jqr* or *pqi* respectively. Therefore, $s_{pqi\to pqr}$, for example, denotes the selection coefficient of the *pqi*-

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

to-*pqr* mutation. *N* denotes effective chromosomal population size (Yang and Nielsen, 2008). x_{pqr} denotes the count of codon *pqr* in a focal gene across the entire population, and $\mu_{i\to r}$ denotes an estimate for the *i*-to-*r* mutation rate. Consequently, the probability that an *i*-to-*r* mutation is fixed in a population is $\frac{2s_{pqi\to pqr}}{1-e^{-2Ns_{pqi\to pqr}}}$ (Yang and Nielsen, 2008). The amino acid serine represents a unique case because it is encoded by six synonymous codons that belong to two independent codon groups, which are not interchangeable via single synonymous mutations. Therefore, in our model we treat the Ser codon groups AGC, AGT and TCA, TCC, TCG, TCT as if they are coding two independent amino acids (Gilchrist et al., 2015).

Since the null hypothesis of our model specifies that all synonymous mutations are neutral, the selection coefficient of each synonymous mutation should have a zero value, which simplifies Equation (1) as follows:

$$\sum_i \lim_{s_{pqi\to pqr}\to 0} \frac{2s_{pqi\to pqr}}{1-e^{-2Ns_{pqi\to pqr}}} \times x_{pqi} \times \mu_{i\to r} + \sum_j \lim_{s_{jqr\to pqr}\to 0} \frac{2s_{jqr\to pqr}}{1-e^{-2Ns_{jqr\to pqr}}} \times x_{jqr} \times \mu_{j\to p}$$
$$= \sum_i \lim_{s_{pqr\to pqi}\to 0} \frac{2s_{pqr\to pqi}}{1-e^{-2Ns_{pqr\to pqi}}} \times x_{pqr} \times \mu_{r\to i} + \sum_j \lim_{s_{pqr\to jqr}\to 0} \frac{2s_{pqr\to jqr}}{1-e^{-2Ns_{pqr\to jqr}}} \times x_{pqr} \times \mu_{p\to j}$$

$$\Rightarrow$$

$$\sum_i \frac{1}{N} \times x_{pqi} \times \mu_{i\to r} + \sum_j \frac{1}{N} \times x_{jqr} \times \mu_{j\to p} = \sum_i \frac{1}{N} \times x_{pqr} \times \mu_{r\to i} + \sum_j \frac{1}{N} \times x_{pqr} \times \mu_{p\to j} \quad (2)$$

Furthermore, since each term in Equation (2) includes the factor $\frac{1}{N}$, Equation (2) could be further simplified as:

$$\sum_i x_{pqi} \times \mu_{i\to r} + \sum_j x_{jqr} \times \mu_{j\to p} = \sum_i x_{pqr} \times \mu_{r\to i} + \sum_j x_{pqr} \times \mu_{p\to j} \quad (3)$$

Subsequently, if the null hypothesis that all synonymous mutations are neutral is true then the impacts of genetic drift, represented by the effective chromosomal population

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

size N, and natural selection, represented by selection coefficient s, are canceled as shown above, and therefore, the observed codon usage pattern could be explained solely by mutational bias. In contrast, if the null hypothesis is rejected then we must accept the alternative hypothesis, which indicates that natural selection has had a significant impact on the observed gene-specific codon usage bias.

To apply our derived statistical model to actual genomic data, we initially generated a dataset of expected values for gene-specific codon usage patterns across all protein-coding genes in each analyzed genome, assuming they are driven solely by mutational bias. To achieve this, Equation (3) was first used to estimate gene-specific μ values by using available reference genomic sequences. We then calculated expected codon counts for each gene based on the μ values estimated by the model. We then tested whether the model-generated (expected) and the actual (observed) gene-specific codon usage patterns are similar by using a χ^2 test (df = 40) (Davis and Olsen, 2009; Xiang et al., 2015; Bera et al., 2017). Subsequently, all genes with observed codon usage patterns that were significantly different from the expected patterns generated by the model were classified as genes that carry signatures of natural selection on synonymous codon usage. As this approach only requires genetic code and the nucleotide sequences of open reading frames as input, in principle, it can be applied to any native protein-coding genes in any species whose genetic code is known.

Next, we used our approach to analyze the well-annotated genomes of five representative eukaryotic species, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Homo sapiens*. Initial analyses of the overall relative synonymous codon usage (RSCU) (Sharp et al., 1988) revealed that these species utilize markedly different genome-level patterns of synonymous codon usage biases (Fig. 1A). Next, we searched for protein-coding genes that exhibit

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

signatures of selection on gene-specific codon usage bias by utilizing our new statistical approach. We found that after correcting for false discovery rate (FDR = 0.05) (Benjamini and Hochberg, 1995), the percentage of protein-coding genes that carry signatures of selection on gene-specific codon usage bias varied between 14% and 64% across these five representative genomes (Fig. 1B). Additional analyses of available eukaryotic genomes revealed that in 35 out of 40 eukaryotic species analyzed, at least 10% of the protein-coding genes carried signatures of natural selection on synonymous codon usage patterns, independent of whether they are unicellular or multicellular (Fig. S1). These results indicate that the overall impact of natural selection on gene-specific patterns of codon usage is broader than would be expected if synonymous substitutions were mostly functionally neutral. Our results also suggest that the relative impact of selection on gene-specific codon usage bias is not constant across the eukaryotic phylogeny, and therefore, might be the result of diverse, clade-specific selective forces.

As currently framed, our quantitative approach is based on a set of assumptions that enable the broad application of our statistical model to as many genomes as possible. However, we acknowledge that some of these assumptions might not represent the correct biological context for all species under all possible ecological conditions. We also anticipate that the assumption that adjacent nucleotides have no effect on the mutation rates of focal nucleotides could increase false-positive rates (Bérard and Guéguen, 2012). However, we anticipate that this possible source for false-positives is compensated for by the assumption that background nucleotide composition of nearby chromosomal regions (Oliver et al, 2001) is negligible, which is further alleviated by the conservative use of relaxed constraints for estimating μ values (Drake et al., 1998). Nevertheless, false positive rates produced by our model could be further reduced by having more precise estimates of the μ parameters, which could be achieved by using

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

population-level genetic variation data, and more specific information about species-specific mutation rates, and how they might be impacted by the identities of adjacent nucleotides.

Selection on gene-specific codon usage bias in *D. melanogaster* is associated with specific biological functions

Our statistical approach for identifying genes that carry signatures of natural selection on synonymous codon usage cannot directly address which specific synonymous codons might be under selection in each selected gene. Therefore, we next used a clustering approach to determine whether the genes we have identified in our initial screen share similar codon usage patterns, or alternatively, represent a heterogeneous population comprised of multiple gene clusters, each defined by a unique pattern of biased codon usage.

Due to its well-annotated genome and wealth of available genetic and phenotypic data, we chose to use the genetically tractable species *D. melanogaster* for functional analyses of genes that are under selection for biased codon usage patterns. Hierarchical gene-clustering analysis revealed that genes with identified signatures of selection on biased codon usage could be assigned to one of two major clusters (Clusters A and B; Fig. 2). Further analysis of Cluster A revealed that synonymous codon usage in these genes is enriched for the rare codons Lys-AAA, Glu-GAA, Gln-CAA, Phe-TTT, Tyr-TAT, and His-CAT, relative to their usage in Cluster B genes. Similar clustering analyses of the genomes of *A. thaliana*, *S. cerevisiae*, *C. elegans*, and *H. sapiens* revealed gene-by-codon clustering in all five species, likely representing species-specific patterns of selection (Figs. S2-5).

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Next, we tested the hypothesis that the identified gene clusters in *D. melanogaster* represent specific biological functions. Analyses of previously published tissue-specific gene expression data (Chintapalli, Wang, and Dow, 2007) revealed that genes that are preferentially expressed in the male accessory glands are significantly overrepresented in Cluster A and underrepresented in Cluster B (Table 1). Similarly, analysis of previously published data on sexually dimorphic genes (Graveley et al., 2010; Gelbart and Emmert, 2013) revealed that genes with male-enriched expression are significantly overrepresented in Cluster A and underrepresented in Cluster B (Table 2). These data suggest that a specific combination of rare codons contributes to the function of some genes associated with the male reproductive system.

In addition, analyses of Gene Ontology terms (Gene Ontology Consortium, 2015) revealed that genes that encode extracellular matrix proteins are overrepresented, and genes that encode cytoplasmic proteins are underrepresented, in Cluster A (Table S1). We also found that genes annotated as encoding odorant-binding proteins, a class of secreted proteins, as well as extracellular space proteins, are underrepresented in Cluster B. Together, these analyses suggest that in *D.* melanogaster, natural selection on the increased usage of specific combinations of rare synonymous codons might be important for male-specific functions of some genes.

Biased gene-specific codon usage can modulate spatial protein expression patterns

One possible explanation for the association between certain biased gene-specific codon usage patterns and tissue-specific expression is that the selected specific rare synonymous codons match tissue-specific tRNA pools, and therefore, enhances the

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

translation efficiency of some protein in specific tissues or cell types (Zalucki, Gittins and Jennings, 2008; Pechmann, Chartron and Frydman, 2014). Similar associations between gene-specific codon usage bias and organ-enriched gene expression patterns have been previously reported in humans (Plotkin, Robins, and Levine, 2004; Dittmar, Goodenbour and Pan, 2006). Therefore, we next tested the hypothesis that the specific pattern of codon usage bias exhibited by Cluster A genes contributes to increased translation efficiency in the male reproductive system.

Our clustering analysis revealed that male-enriched genes in Cluster A preferentially use the rare codons Lys-AAA, Gln-CAA, Glu-GAA, Phe-TTT, Tyr-TAT, and His-CAT (Fig. 2). However, only the first three codons are recognized by exactly matched tRNA anticodons, while the other three rare codons share the same tRNAs with their more commonly used synonymous codons, with a mismatch (wobble) at the third codon position (Crick, 1966; Chan and Lowe, 2009), which suggests that these codons are less likely to contribute to tissue-specific translation efficiency. Therefore, we next analyzed the specific usage of rare codons AAA, CAA, and GAA in all genes that show enriched expression in the male reproductive system, independent of whether these genes have passed the initial statistical threshold for the detection of selection on gene-specific codon usage patterns. We found that these specific rare synonymous codons are indeed overrepresented in male-reproductive-system-specific genes (Table 3). Thus, not only are genes preferring AAA, CAA, and GAA more likely to be male-reproductive-system-specific, but also genes specifically transcribed in male reproductive system use AAA, CAA, and GAA more often than their expected genome average usage.

If the selection on an increased relative usage of rare codons AAA, CAA, and GAA serves a biological function that is specific to gene function in the male reproductive system then it should depend, at least in part, on the increased expression of the

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

specific rare tRNAs that bind these specific codons. In accordance with this hypothesis, northern blot analyses of tissue-specific tRNA expression revealed that $tRNA^{Lys}_{TTT}$ is enriched in the male reproductive organs, further supporting the hypothesis that spatial regulation of some tRNA genes that correspond to rare codons could contribute to tissue-specific increase in protein translation efficiency (Fig. 3A-C). The expression patterns of the tRNAs that match Glu-GAA and Gln-CAA were not investigated because their sequences are almost identical to their common tRNA counterparts, which does not allow their independent detection by hybridization probes (de Muro, 2008). Nevertheless, these data indicate that some *Drosophila* genes have evolved a specific codon usage bias that specifically drives their levels and/or activity in the male reproductive system.

Next, we empirically tested whether specific combinations of rare codons are sufficient to drive tissue-specific translation efficiency by generating transgenic flies that express different alleles of the EGFP reporter under the control of the UAS-GAL4 binary expression system (Brand and Perrimon, 1993). One allele was comprised of common codons for all amino acids ($EGFP^{Common}$), and the second allele was comprised of rare codons AAA, GAA, and CAA for amino acids Lys, Glu, and Gln respectively ($EGFP^{RareKEQ}$), which represent about 18% of the total residues in EGFP. Each *EGFP* allele was also co-expressed with an *RFP* reporter encoded by common codons, which served as a transgene expression control. Both transgenes were driven by the ubiquitous *Act5C*-GAL4 driver (Kopp, Blackman, and Duncan, 1999). To evaluate the effect of codon usage bias on the spatial pattern of both EGFP alleles *in vivo*, we compared the two genotypes by measuring the RFP-normalized EGFP signals in the accessory gland secondary cells (AGSC) and in the hindgut proliferation zone (HPZ) (Takashima et al., 2013). We found that the signal from the $EGFP^{RareKEQ}$ allele in the AGSCs was significantly higher than that of the $EGFP^{Common}$ allele. In contrast, both

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

alleles produced similar signals in the HPZ (Fig. 3D-I). Because the overall transcript levels of both alleles are similar (Fig. 3J), these results suggest that some specific combinations of synonymous rare codons could enhance protein translation in the *Drosophila* male reproductive system.

DISCUSSION

In contrast to a broadly accepted dogma in evolutionary biology and population genetics, our theoretical and empirical studies indicate that synonymous mutations should not be considered neutral in general. By using multiple complementary approaches, we demonstrate here that the impact of natural selection on gene-specific codon usage bias is common across the eukaryotic phylogeny and is likely to play a much more prominent role in defining gene-specific codon usage bias than was previously assumed. Furthermore, our data indicate that in multicellular organisms such as the fruit fly, natural selection is likely to optimize codon usage patterns to match specific spatial and temporal constraints on gene function, and therefore, what might represent an optimal codon usage for individual genes does not necessarily mean common codons should be expected at every position. Our findings also indicate that previous reports about the phenotypic consequences of synonymous mutations are not anecdotal rare cases. Instead, these cases likely represent a fundamental biological principle associated with the spatial and temporal gene activity landscapes in eukaryotes. In addition, our analyses of gene-specific codon usage biases in *D. melanogaster* suggest that gene-specific codon usage bias is heterogeneous across protein-coding genes in a single species, and therefore, likely to be the outcome of diverse selective pressures associated with different biological functions. Furthermore, our data suggest that

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

different classes of rare synonymous codons exhibit differential sensitivities to natural selection.

Consequently, our discovery that gene-specific codon usage patterns are often shaped by natural selection indicates that alternative synonymous codons serve important, yet often overlooked, biological functions. Although several previous studies have used statistical approaches to suggest that synonymous mutations in specific genes and species might not be neutral (Berger, 1977; Eyre-Walker, 1991; Stenico, Lloyd, and Sharp, 1994; Lawrie et al., 2013), our current study provides the first genome-scale comprehensive view of the overall impact of natural selection on gene-specific synonymous codon usage patterns across diverse eukaryotic taxa. Thus, our unbiased and broadly applicable statistical approach has uncovered a fundamental genomic organizational principle that is sensitive to natural selection, and likely represents an important component of the genotype-phenotype axis in diverse eukaryotic species and biological contexts. Therefore, it is likely that investigations, which rely on the assumption that synonymous mutations are neutral, often overestimate or underestimate the prevalence of selection on specific protein-coding genes, which could impact the interpretations of the role of these genes in genotype-phenotype relationships at the level of populations and the evolutionary timescale in particular (Chen et al., 2010), and the genome evolution in general (Campos et al., 2012; Matsumoto et al., 2016). Specifically, the prevailing assumption that most synonymous mutations are neutral, and therefore represent a robust and quantitative measure of selection-independent evolutionary rates of nucleotide substitutions, needs to be revised. This is especially important in the context of the commonly used tests for identifying molecular signatures of selection in specific protein-coding genes, and for classifying the modes of natural

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

selection (*e.g.*, “positive” versus “negative” selection) (McDonald and Kreitman, 1991; Nielsen, 2001; Echave, Spielman, and Wilke, 2016).

Our data also indicate that, similarly to *Drosophila*, many human protein-coding genes carry signatures of selection on their biased codon usage patterns (Fig. 1). This finding could have broad implications for studies of genetic variants underlying quantitative phenotypes in human populations. Consequently, assuming that synonymous mutations are functionally neutral could lead to a biased view that associations between nonsynonymous SNPs and specific phenotypic values are the only ones important for understanding overall trait variations (Hampe et al., 2007; WTCCC and TASC, 2007; Xu and Taylor, 2009; Fong et al., 2010; Huang et al., 2012). This is supported by a recent data survey of human GWAS data from over 2000 independent studies, which suggested that synonymous and nonsynonymous SNPs have similar likelihoods and effect sizes in terms of association with disease phenotypes (Chen et al., 2010). Although our analysis of human gene-specific codon usage bias was based on data from a single reference genome, the simplest interpretation of our results is that both synonymous and nonsynonymous mutations are likely influenced by natural selection in human populations, and therefore, many synonymous SNPs could still have significant quantitative and/or qualitative impacts on gene functions and associated phenotypes in health and disease.

We also found that gene-specific codon usage bias is not uniform across the genome. Instead, our data suggest that different groups of genes in the same species often exhibit enrichment for specific combinations of rare or common codons, which suggests that different combinations of synonymous codons may have been selected for independent biological reasons. In this regard, we show that at least one group of genes with enriched expression in the reproductive system of male *Drosophila* preferentially

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

use the rare codons Lys-AAA, Gln-CAA, and Glu-GAA. By using allelic variants of a reporter gene, we show that the selective usage of these three rare codons is sufficient for generating spatially biased protein expression patterns. These findings are consistent with some of the assumptions about the possible roles of codon usage bias, first put forward by the “translational selection theory” (Ikemura, 1981; dos Reis, Savva, and Wernisch, 2004; Plotkin, Robins, and Levine, 2004; Dittmar, Goodenbour, and Pan, 2006). Consequently, our study provides a broader context to this fundamental evolutionary theory by emphasizing the possible role of gene-specific codon usage bias in the spatial regulation of proteins in multicellular eukaryotes. However, regulation of spatial protein expression cannot explain all cases of selection on gene-specific codon usage bias because genomes of some unicellular eukaryotes, such as the choanoflagellate *Monosiga brevicollis* and the green algae *Chlamydomonas reinhardtii*, also include a significant proportion of protein-coding genes that exhibit signatures of selection on their codon usage patterns (Fig. S1).

In addition to providing a fundamental aspect of gene regulation, our discovery that the impact of natural selection on gene-specific codon usage bias is prevalent in eukaryotic genomes also carries important practical implications for experimental systems that depend on heterologous gene expression. Specifically, we have found that for some genes, the selective usage of synonymous rare codons could increase their translation efficiency in specific tissues and cell types. Therefore, designing experiments with the assumption that codon optimization of a transgene necessitates the preferential usage of common synonymous codons could lead to unexpected phenotypic outcomes, and could affect data interpretation, especially for genes with restricted spatial and/or temporal expression patterns.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Our studies also highlight the need for a better understanding of the mechanisms that regulate the spatial expression patterns of tRNAs. Some previous studies have argued that tRNA gene copy number is likely the primary mechanism that regulates the relative levels of individual tRNAs in cellular pools (dos Reis, Savva and Wernisch, 2004). Yet, variations in gene copy number alone cannot explain the tissue-specific expression patterns of some rare tRNAs observed by us (Fig. 3A-C) and others (Dittmar, Goodenbour and Pan, 2006; Gingold et al., 2014). Therefore, although tRNAs are thought to be exclusively transcribed by the constitutively-active RNA Pol III complex, there must be additional molecular mechanisms that enable the temporal and spatial regulation of some tRNAs at the transcriptional and/or post-transcriptional levels (Yoshihisa et al., 2007; Wichtowska, Turowski, and Boguta, 2013; Turowski and Tollervey, 2016).

Although our empirical work here has mostly focused on a specific set of rare codons that bind unique tRNA anti-codons, it should be noted that genes in Cluster A (Fig. 2) also preferentially use the rare codons Phe-TTT, Tyr-TAT, and His-CAT. These rare synonymous codons share their tRNA anticodons with their more common synonymous codons via a wobble mechanism (Chan and Lowe, 2009). Theoretical models and empirical data suggest that this specific class of rare codons is likely to slow down translation due to the lower affinity of the common tRNAs to their associated rare synonymous codons (Ikemura, 1981; dos Reis, Savva, and Wernisch, 2004; Savir and Tlusty, 2013; Brule and Grayhack, 2017). Therefore, we speculate that in contrast to rare codons that bind unique perfectly-matching tRNAs, rare codons that share tRNAs with their more common synonymous codons may have evolved to broadly restrict translation efficiency of some low-abundance proteins, or proteins that require slow folding kinetics associated with specific post-translation modifications, or the binding of other protein and

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

small-molecule partners, chaperones, or co-factors (Komar, Lesnik and Reiss, 1999; Weygand-Durasevic and Ibba, 2010; Angov, 2011; Zhou et al., 2013; Fluman et al., 2014; Pechmann, Chartron and Frydman, 2014; Chaney and Clark, 2015; Quax et al., 2015; Yu et al., 2015; Fu et al., 2016; Zhao, Yu, and Liu, 2017).

In summary, the results we present here suggest that in contrast to a widely accepted dogma, synonymous mutations should not be considered functionally and evolutionarily neutral by default. Instead, as was previously shown for a few specific protein-coding genes (Hasegawa, Yasunaga, and Miyata, 1979; Ikemura, 1981; Akashi, 1994; Stenico, Lloyd, and Sharp, 1994; Ko et al., 2005; Pagani, Raponi, and Baralle, 2005; Kimchi-Sarfaty et al., 2007; Parmley and Hurst, 2007; Zhang, Hubalewska, and Ignatova, 2009; Gu et al., 2012; Zhou et al., 2013; Ma et al., 2014; Shin et al., 2015; Fu et al., 2016), our data indicate that selection on synonymous codon usage is likely to serve a more general function in diverse biological processes across Eukaryotes. Nevertheless, although our data indicate that “synonymous” and “neutral” mutations often represent different biological values in the context of the genotype-phenotype axis, we do not suggest that the concept of evolutionary neutrality, first described by Darwin in the *On the Origin of Species* as "*Variations neither useful nor injurious would not be affected by natural selection*", should be overall rejected. Instead, we argue that the commonly accepted specific assumption that all synonymous mutations are generally neutral should be reevaluated.

ACKNOWLEDGEMENTS

Z.P. dedicates this manuscript to Dr. Yang Zhong of Fudan University, a prominent botanist and evolutionary biologist who had a great influence on his choice to pursue a

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

research career in mathematical and evolutionary biology. The authors also wish to thank Alan Templeton, Allan Larson, and members of the Ben-Shahar and Zaher laboratories for helpful discussions. They also thank Paula Kiefel, Dianne Duncan, and Ian Duncan for technical assistance with generating transgenic animals and microscopic imaging. Z.P. is supported by the NIH training grant 2T32HG000045. This research has been supported in part by NIH grant NS089834, and NSF grants 1545778 and 1707221 to YB, and NIH grant 5R01GM112641 to HZ.

AUTHOR CONTRIBUTIONS

ZP and YB conceived the study; ZP performed the studies under the guidance of YB and HZ; HZ assisted with the design and execution of molecular biology studies; ZP, HZ, and YB wrote the manuscript and approved its final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

Akashi, H. (1994). Synonymous codon usage in *Drosophila melanogaster*: natural selection and translation accuracy. *Genetics, 136*(3), 927-935.

Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at" silent" sites in *Drosophila* DNA. *Genetics, 139*(2), 1067-1076.

Angov, E. (2011). Codon usage: nature's roadmap to expression and folding of proteins. *Biotechnology journal*, *6*(6), 650-659.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

Bera, B. C., Virmani, N., Kumar, N., Anand, T., Pavulraj, S., Rash, A., Elton, D., Rash, N., Bhatia, S., Sood, R., et al. (2017).Genetic and codon usage bias analyses of polymerase genes of equine influenza virus and its relation to evolution. *BMC genomics, 18*(1), 652.

Bérard, J., and Guéguen, L. (2012). Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context. *Systematic biology, 61*(3), 510-521.

Berger, E. M. (1977). Are Synonymous Mutations Adaptively Neutral?.*The American Naturalist, 111*(979), 606-607.

Bischof, J., Maeda, R. K., Hediger, M., Karch, F., and Basler, K. (2007). An optimized transgenesis system for *Drosophila* using germ-line-specific φC31 integrases. *Proceedings of the National Academy of Sciences, 104*(9), 3312-3317.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Brand, A. H., and Perrimon, N. (1993). Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development, 118*(2), 401-415.

Brown, A., Shao, S., Murray, J., Hegde, R. S., and Ramakrishnan, V. (2015).Structural basis for stop codon recognition in eukaryotes. *Nature*, *524*(7566), 493-496.

Brule, C. E., and Grayhack, E. J. (2017). Synonymous codons: choose wisely for expression. *Trends in Genetics, 33*(4), 283-297.

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

Campos, J. L., Zeng, K., Parker, D. J., Charlesworth, B., and Haddrill, P. R. (2012). Codon usage bias and effective population sizes on the X chromosome versus the autosomes in *Drosophila melanogaster*. *Molecular biology and evolution, 30*(4), 811-823.

Chan, P. P., and Lowe, T. M. (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic acids research, 37*(suppl 1), D93-D97.

Chaney, J. L., and Clark, P. L. (2015). Roles for synonymous codon usage in protein biogenesis. *Annual review of biophysics*, *44*, 143-166.

Chen, R., Davydov, E. V., Sirota, M., and Butte, A. J. (2010). Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PloS one, 5*(10), e13574.

Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L., and McAdams, H. H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Proceedings of the National Academy of Sciences of the United States of America, 101(10), 3480-3485.

Chintapalli, V. R., Wang, J., and Dow, J. A. (2007). Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nature genetics*, *39*(6), 715-720.

Crick, F. H. (1966). Codon—anticodon pairing: the wobble hypothesis. *Journal of molecular biology*, *19*(2), 548-555.

Davis, J. J., and Olsen, G. J. (2009). Modal codon usage: assessing the typical codon usage of a genome. *Molecular biology and evolution, 27*(4), 800-810.

de Muro, M. A. (2008). Probe design, production, and applications. In Molecular Biomethods Handbook, J. M. Walker and R. Rapleyeds. (Totowa, NJ, U.S.A.: Humana Press. 2nd edition), pp. 41-53.

Dittmar, K. A., Goodenbour, J. M., and Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. *PLoS Genet*, *2*(12), e221.

Doherty, A., and McInerney, J. O. (2013). Translation selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. *Molecular biology and evolution, 30*(10), 2263-2267.

dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translation selection. *Nucleic acids research, 32*(17), 5036-5044.

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics, 148*(4), 1667-1686.

Echave, J., Spielman, S. J., and Wilke, C. O. (2016). Causes of evolutionary rate variation among protein sites.*Nature Reviews. Genetics, 17*(2), 109.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Eyre-Walker, A. C. (1991). An analysis of codon usage in mammals: selection or mutation bias?.*Journal of molecular evolution, 33*(5), 442-449.

Fluman, N., Navon, S., Bibi, E., and Pilpel, Y. (2014). mRNA-programmed translation pauses in the targeting of E. coli membrane proteins. *Elife*, *3*, e03440.

Fong, C., Ko, D. C., Wasnick, M., Radey, M., Miller, S. I., and Brittnacher, M. (2010). GWAS analyzer: integrating genotype, phenotype and public annotation data for genome-wide association study analysis. *Bioinformatics, 26*(4), 560-564.

Fu, M. (2010). Codon usage bias in herpesvirus. *Archives of Virology, 155*(3), 391-396.

Fu, J., Murphy, K. A., Zhou, M., Li, Y. H., Lam, V. H., Tabuloc, C. A., Chiu, J. C., and Liu, Y. (2016). Codon usage affects the structure and function of the *Drosophila* circadian clock protein PERIOD. *Genes and Development*, *30*(15), 1761-1775.

Gelbart, W. M., and Emmert, D. B. (2013). FlyBase High Throughput Expression Pattern Data

Gene Ontology Consortium. (2015). Gene ontology consortium: going forward. *Nucleic acids research*, *43*(D1), D1049-D1056.

Gilchrist, M. A., Chen, W. C., Shah, P., Landerer, C. L., and Zaretzki, R. (2015). Estimating Gene Expression and Codon-Specific Translation Efficiencies, Mutation Biases, and Selection Coefficients from Genomic Data Alone. *Genome biology and evolution*, *7*(6), 1559-1579.

Gingold, H., Tehler, D., Christoffersen, N. R., Nielsen, M. M., Asmar, F., Kooistra, S. M., Christophersen, N. S., Christensen, L. L., Borre, M., Sørensen, K. D., et al. (2014). A

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dual program for translation regulation in cellular proliferation and differentiation. *Cell*, 158(6), 1281-1292.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Cherbas, L., Choi, J., Davis, C.A., Dobin, A., Duff, M., Eads, B., Hansen, K.D., et al. (2010.03.15). The *D. melanogaster* transcriptome: modENCODE RNA-Seq data

Gu, W., Wang, X., Zhai, C., Xie, X., and Zhou, T. (2012). Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. *Molecular biology and evolution, 29*(10), 3037-3044.

Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr, G., De La Vega, F. M., Briggs, J., et al. (2007). A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nature genetics, 39*(2), 207.

Hasegawa, M., Yasunaga, T., and Miyata, T. (1979). Secondary structure of MS2 phage RNA and bias in code word usage. *Nucleic acids research, 7*(7), 2073-2079.

Hill, A., Zheng, X., Li, X., McKinney, R., Dickman, D., and Ben-Shahar, Y. (2017). The *Drosophila* postsynaptic DEG/ENaC channel ppk29 contributes to excitatory neurotransmission. *Journal of Neuroscience, 37*(12), 3171-3180.

Huang, X., Kurata, N., Wang, Z. X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H., Li, W., Guo, Y., et al. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature, 490*(7421), 497.

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

synonymous codon choice that is optimal for the *E. coli* translation system. *Journal of molecular biology, 151*(3), 389-409.

Kanaya, S., Yamada, Y., Kudo, Y., and Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, *238*(1), 143-155.

Kessler, M. D., and Dean, M. D. (2014). Effective population size does not predict codon usage bias in mammals. *Ecology and evolution, 4*(20), 3887-3900.

Kimchi-Sarfaty, C., Oh, J. M., Kim, I. W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., and Gottesman, M. M. (2007). A" silent" polymorphism in the MDR1 gene changes substrate specificity. *Science, 315*(5811), 525-528.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature, 217*(5129), 624-626.

Kimura, M., and Ohta, T. (1974). On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences, 71*(7), 2848-2852.

King, J. L., and Jukes, T. H. (1969). Non-darwinian evolution. *Science, 164*(3881), 788-798.

Ko, H. J., Ko, S. Y., Kim, Y. J., Lee, E. G., Cho, S. N., and Kang, C. Y. (2005). Optimization of codon usage enhances the immunogenicity of a DNA vaccine encoding mycobacterial antigen Ag85B. *Infection and immunity, 73*(9), 5666-5674.

Komar, A. A., Lesnik, T., and Reiss, C. (1999). Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS letters*, *462*(3), 387-391.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Kopp, A., Blackman, R. K., and Duncan, I. (1999). Wingless, decapentaplegic and EGF receptor signaling pathways interact to specify dorso-ventral pattern in the adult abdomen of *Drosophila*. *Development, 126*(16), 3495-3507.

Kraft, D. (1988). A software package for sequential quadratic programming. Technical report DFVLR-FB 88–28. (Oberpfaffenhofen: Institut für Dynamik der Flugsysteme).

Kubatko, L., Shah, P., Herbei, R., and Gilchrist, M. A. (2016). A codon model of nucleotide substitution with selection on synonymous codon usage. *Molecular Phylogenetics and Evolution*, *94*, 290-297.

Lawrie, D. S., Messer, P. W., Hershberg, R., and Petrov, D. A. (2013). Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS genetics, 9*(5), e1003527.

Lu, B., LaMora, A., Sun, Y., Welsh, M. J., and Ben-Shahar, Y. (2012). ppk23-Dependent chemosensory functions contribute to courtship behavior in *Drosophila melanogaster*. *PLoS Genet*, *8*(3), e1002587.

Ma, L., Cui, P., Zhu, J., Zhang, Z., and Zhang, Z. (2014). Translation selection in human: more pronounced in housekeeping genes. *Biology direct*, 9(1), 17.

Matsumoto, T., John, A., Baeza-Centurion, P., Li, B., and Akashi, H. (2016). Codon usage selection can bias estimation of the fraction of adaptive amino acid fixations. *Molecular biology and evolution, 33*(6), 1580-1589.

McDonald, J. H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature, 351*(6328), 652-654.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Meurer, A., Smith, C. P., Paprocki, M., Čertík, O., Kirpichev, S. B., Rocklin, M., Kumar, A., Ivanov, S., Moore, J. K., Singh, S., et al. (2017). SymPy: symbolic computing in Python. *PeerJ Computer Science, 3*, e103.

Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity*, *86*(6), 641-647.

Nirenberg, M., and Leder, P. (1964). RNA codewords and protein synthesis.*Science, 145*(3639), 1399-1407.

Oliver, J. L., Bernaola-Galván, P., Carpena, P., and Román-Roldán, R. (2001). Isochore chromosome maps of eukaryotic genomes. *Gene, 276*(1), 47-56.

Pagani, F., Raponi, M., and Baralle, F. E. (2005). Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(18), 6368-6372.

Parmley, J. L., and Hurst, L. D. (2007). Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Molecular biology and evolution, 24*(8), 1600-1603.

Pechmann, S., Chartron, J. W., and Frydman, J. (2014). Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. *Nature structural and molecular biology, 21*(12), 1100-1105.

Plotkin, J. B., Robins, H., and Levine, A. J. (2004). Tissue-specific codon usage and the expression of human genes. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(34), 12588-12591.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Plotkin, J. B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics, 12*(1), 32.

Pouyet, F., Bailly-Bechet, M., Mouchiroud, D., and Guéguen, L. (2016). SENCA: A Multilayered Codon Model to Study the Origins and Dynamics of Codon Usage. *Genome Biology and Evolution*, *8*(8), 2427-2441.

Quax, T. E., Claassens, N. J., Söll, D., and van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. *Molecular cell*, *59*(2), 149-161.

Rao, Y., Wu, G., Wang, Z., Chai, X., Nie, Q., and Zhang, X. (2011). Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. *DNA research, 18*(6), 499-512.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*, *107*(10), 4629-4634.

Rodriguez, F. J. L. O. J., Oliver, J. L., Marin, A., and Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *Journal of theoretical biology, 142*(4), 485-501.

Savir, Y., and Tlusty, T. (2013). The ribosome as an optimal decoder: A lesson in molecular recognition. *Cell, 153*(2), 471-479.

Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H., and Wright, F. (1988). Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic acids research*, *16*(17), 8207-8211.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Sharp, P. M., Emery, L. R., and Zeng, K. (2010). Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 365*(1544), 1203-1212.

Shields, D. C., and Sharp, P. M. (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translation selection and mutational biases. *Nucleic Acids Research, 15*(19), 8023-8040.

Shin, Y. C., Bischof, G. F., Lauer, W. A., and Desrosiers, R. C. (2015). Importance of codon usage for the temporal regulation of viral gene expression. *Proceedings of the National Academy of Sciences, 112*(45), 14030-14035.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods, 9*(7), 676.

Slimko, E. M., and Lester, H. A. (2003). Codon optimization of *Caenorhabditis elegans* GluCl ion channel genes for mammalian cells dramatically improves expression levels. *Journal of neuroscience methods*, *124*(1), 75-81.

Söll, D., Jones, D. S., Ohtsuka, E., Faulkner, R. D., Lohrmann, R., Hayatsu, H., Khorana, H.G., Cherayil, J.D., Hampel, A., and Bock, R. M. (1966). Specificity of sRNA for recognition of codons as studied by the ribosomal binding technique. *Journal of molecular biology*, *19*(2), 556-573.

Stenico, M., Lloyd, A. T., and Sharp, P. M. (1994). Codon usage in *Caenorhabditis elegans*: delineation of translation selection and mutational biases. *Nucleic Acids Research*, *22*(13), 2437-2446.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Takashima, S., Paul, M., Aghajanian, P., Younossi-Hartenstein, A., and Hartenstein, V. (2013). Migration of *Drosophila* intestinal stem cells across organ boundaries. *Development, 140*(9), 1903-1911.

Tamuri, A. U., dos Reis, M., and Goldstein, R. A. (2012). Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, *190*(3), 1101-1115.

Turowski, T. W., and Tollervey, D. (2016). Transcription by RNA polymerase III: insights into mechanism and regulation. *Biochemical Society Transactions, 44*(5), 1367-1375.

Wang, H. C., and Hickey, D. A. (2007). Rapid divergence of codon usage patterns within the rice genome. *BMC evolutionary biology, 7*(1), S6.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., and Lumley, T. (2015). gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1. *https://cran.r-project.org/web/packages/gplots/gplots.pdf.*

Wellcome Trust Case Control Consortium (WTCCC) and The Australo-Anglo-American Spondylitis Consortium (TASC). (2007). Association scan of 14,500 nsSNPs in four common diseases identifies variants involved in autoimmunity. *Nature genetics, 39*(11), 1329.

Weygand-Durasevic, I., and Ibba, M. (2010). New roles for codon usage. *Science*, *329*(5998), 1473-1474.

Wichtowska, D., Turowski, T. W., and Boguta, M. (2013). An interplay between transcription, processing, and degradation determines tRNA levels in yeast. *Wiley Interdisciplinary Reviews: RNA, 4*(6), 709-722.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Xiang, H., Zhang, R., Butler III, R. R., Liu, T., Zhang, L., Pombert, J. F., and Zhou, Z. (2015). Comparative Analysis of Codon Usage Bias Patterns in Microsporidian Genomes. *PloS one, 10*(6), e0129223.

Xu, Z., and Taylor, J. A. (2009). SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic acids research, 37*(suppl_2), W600-W605.

Yang, Z., and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular biology and evolution*, *25*(3), 568-579.

Yang, X., Ma, X., Luo, X., Ling, H., Zhang, X., and Cai, X. (2015). Codon usage bias and determining forces in *Taenia solium* genome. *The Korean journal of parasitology, 53*(6), 689.

Yoshihisa, T., Ohshima, C., Yunoki-Esaki, K., and Endo, T. (2007). Cytoplasmic splicing of tRNA in Saccharomyces cerevisiae. *Genes to Cells, 12*(3), 285-297.

Yu, C. H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M. S., and Liu, Y. (2015). Codon usage influences the local rate of translation elongation to regulate co-translation protein folding. *Molecular cell, 59*(5), 744-754.

Zalucki, Y. M., Gittins, K. L., and Jennings, M. P. (2008). Secretory signal sequence non-optimal codons are required for expression and export of β-lactamase. *Biochemical and biophysical research communications, 366*(1), 135-141.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2017). Ensembl 2018. *Nucleic acids research, 46*(D1), D754-D761.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Zhang, G., Hubalewska, M., and Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translation folding. *Nature structural & molecular biology, 16*(3), 274-280.

Zhao, F., Yu, C. H., and Liu, Y. (2017). Codon usage regulates protein structure and function by affecting translation elongation speed in *Drosophila* cells*. Nucleic Acids Research.45*(14), 8484-8492.

Zhou, M., Guo, J., Cha, J., Chae, M., Chen, S., Barral, J. M., Sachs, M. S., and Liu, Y. (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature*, *495*(7439), 111-115.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

FIGURE LEGENDS

Figure 1. Signatures of natural selection on gene-specific synonymous codon usage bias in representative eukaryotic species. (A) Analyses of the relative synonymous codon usage (RSCU) as a measure of the overall codon usage bias across five representative Eukaryotic species. (B) Total number of protein-coding genes analyzed for each species were: *D. melanogaster*, 12099; *A. thaliana*, 23694; *S. cerevisiae*, 5366; *C. elegans*, 17210; *H. sapiens*, 20261. The estimated percentages of genes carrying selection signatures are corrected by a 0.05 false discovery rate. See also Fig. S1.

Figure 2. Clustered codon usage patterns across genes carrying signatures of selection on gene-specific codon usage bias in *D. melanogaster*. Hierarchical clustering of all genes (horizontal axis) identified as carrying significant signatures of selection on gene-specific codon usage bias. The relative usage frequencies of each codon in each gene are color-coded (vertical axis). Euclidean distances were used to measure dissimilarities between codons. Spearman's correlation coefficients were used to measure similarities between genes. Complete linkage was used as the clustering criterion. Clusters A (red bar) and B (black bar) represent the two major gene groups identified. See also Figs. S2-5, Table S1.

Figure 3. Gene-specific codon usage bias affects spatial protein expression in *D. melanogaster*. (A) Representative northern blots and (B-C) summary data of the relative abundance of $tRNA^{Lys}_{CTT}$ (common) and $tRNA^{Lys}_{TTT}$ (rare) codons across male tissues. *, $p<0.05$ (n = 4, ANOVA followed by SNK *post hoc* tests). Error bars denote

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

standard deviation. †: Reproductive system is excluded. (D-E) Representative images showing EGFP and RFP expression in the HPZ (yellow lines mark tissue boundaries). (F-G) Representative images showing EGFP and RFP expression in the AGSC (yellow lines mark secondary cells). (H-I) Summary data of normalized EGFP signals in HPZ and AGSC. *, $p<0.01$ (n=5, two-tailed unpaired Student's *t*-test). Error bars denote standard deviation. (J) Real-time qRT-PCR mRNA expression data of $EGFP^{Common}$ and $EGFP^{RareKEQ}$ (two-tailed unpaired Student's *t*-test, n = 4, NS). Error bars denote SEM.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

TABLES

	Cluster A		Cluster B	
Tissue	Fold enrichment	p-value	Fold enrichment	p-value
Adult Central Nervous System	0.48	0.378	1.13	0.04
Brain	N/A	N/A	0.57	0.205
Crop	N/A	N/A	1.15	0.622
Midgut	2.31	0.066	1.19	0.003
Hindgut	N/A	N/A	1.01	0.582
Malpighian Tubules	1.11	0.599	1.1	0.205
Ovary	1.71	0.256	1.09	0.123
Testis	1.36	0.082	**0.56**	**<0.001**
Male Accessory Gland	**4.45**	**<0.001**	**0.36**	**<0.001**
Carcass	N/A	N/A	**0.34**	**<0.001**
Salivary Gland	N/A	N/A	0.49	0.024
Heart	N/A	N/A	0.86	0.493
Eye	2.53	0.187	0.8	0.05

Table 1. Tissue-specific expression patterns of genes in Clusters A and B. Genes included in Clusters A and B are as in Fig. 2. Significant over- or under-representation is shown in bold. N/A means that no tissue-specific genes were found in the entire genome or the focal gene cluster. Although present in the FlyAtlas database, we did not identify tissue-enriched genes in the thoracic-abdominal ganglion, virgin female spermatheca, inseminated female spermatheca, and adult fat body, and therefore, are not included

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

here. Fold enrichment is relative to the entire genome. Bonferroni correction is applied, $p < 0.05/20 = 0.0025$. See also Table S1.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Gene cluster	Sex	Fold enrichment	p-value
Cluster A	Male	**1.66**	**<0.001**
	Female	1.67	0.091
Cluster B	Male	**0.61**	**<0.001**
	Female	1.07	0.060

Table 2. Sex-biased expression patterns of genes in Clusters A and B. Genes included are as in Table 1. Significant over- or under-representation is shown in bold. Fold enrichment is relative to the entire genome. Bonferroni correction is applied, $p < 0.05/4 = 0.0125$. See also Table S1.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Codons	Count of codons in male-reproductive-organ-specific genes	Count of codons in all analyzed genes	Fold enrichment in male-reproductive-organ-specific genes	p-value
Lys-AAA	12936	117591	1.18	<0.001
Lys-AAG	22139	258907	0.92	
Gln-CAA	10181	108912	1.26	<0.001
Gln-CAG	15727	241374	0.88	
Glu-GAA	16093	151650	1.22	<0.001
Glu-GAG	21843	285332	0.88	

Table 3. Usage frequencies of rare codons Lys-AAA, Gln-CAA, and Glu-GAA in the male reproductive tissues relative to whole genome averages.

Counts of relevant codons were calculated for the entire genome and male-reproductive-organ-specific genes (including genes with enriched expression in the testis and the accessory-glands). Hypergeometric tests were used to test whether a specific synonymous codon is over- or under-represented in male-reproductive-organ-specific genes. Fold enrichment is calculated by dividing the relative frequency of a focal codon

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in male-reproductive-organ-specific genes by the relative frequency of the focal codon usage across all protein-coding genes.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

EXPERIMENTAL PROCEDURES

Genomic and transcriptomic data

Protein-coding DNA sequences were from Ensembl (Zerbino et al., 2017) (https://www.ensembl.org/). Reference coding sequences included in the analysis of gene-specific codon usage bias were chosen according to the following criteria: 1) The sequence length is a multiple of three. 2) The sequence uses standard genetic code. 3) For each gene, only the longest mRNA isoform was used for analysis. If there were multiple isoforms of the same length, then the first record shown in the fasta file was used. 4) The encoded protein includes all 19 amino acids that have degenerate codons; for the amino acid serine, the two-fold and four-fold degenerate codon groups were treated as if they encoded two different amino acids. Transcriptomic data were from the FlyAtlas microarray database (Chintapalli, Wang, and Dow, 2007) and the modENCODE RNA-seq database (Graveley et al., 2010; Gelbart and Emmert, 2013).

Estimating μ and expected codon counts

The relationships between μ values and codon counts are described by Equation (3) (Results section). Based on the standard genetic code and all possible combinations for synonymous nucleotide substitutions, we classified all degenerate codons into six categories. For each category, we used Equation (3) to generate a homogeneous linear equation system. Subsequently, we treated one of the x variables as a known parameter and analytically solved all other x variables. Finally, these solutions were used to calculate the expected codon counts of a protein-coding gene.

Category one – Codons for Arginine:

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\begin{cases} x_{AGA} \times \mu_{A\to C} + x_{AGA} \times \mu_{A\to G} = x_{CGA} \times \mu_{C\to A} + x_{AGG} \times \mu_{G\to A} \\ x_{AGG} \times \mu_{A\to C} + x_{AGG} \times \mu_{G\to A} = x_{CGG} \times \mu_{C\to A} + x_{AGA} \times \mu_{A\to G} \\ x_{CGA} \times \mu_{C\to A} + x_{CGA} \times (\mu_{A\to C} + \mu_{A\to G} + \mu_{A\to T}) \\ \qquad = x_{AGA} \times \mu_{A\to C} + x_{CGC} \times \mu_{C\to A} + x_{CGG} \times \mu_{G\to A} + x_{CGT} \times \mu_{T\to A} \\ x_{CGG} \times \mu_{C\to A} + x_{CGG} \times (\mu_{G\to A} + \mu_{G\to C} + \mu_{G\to T}) \\ \qquad = x_{AGG} \times \mu_{A\to C} + x_{CGA} \times \mu_{A\to G} + x_{CGC} \times \mu_{C\to G} + x_{CGT} \times \mu_{T\to G} \\ x_{CGC} \times (\mu_{C\to A} + \mu_{C\to G} + \mu_{C\to T}) = x_{CGA} \times \mu_{A\to C} + x_{CGG} \times \mu_{G\to C} + x_{CGT} \times \mu_{T\to C} \\ x_{CGT} \times (\mu_{T\to A} + \mu_{T\to C} + \mu_{T\to G}) = x_{CGA} \times \mu_{A\to T} + x_{CGC} \times \mu_{C\to T} + x_{CGG} \times \mu_{G\to T} \end{cases} . \quad (4)$$

Category two – Codons for Leucine:

$$\begin{cases} x_{TTA} \times \mu_{T\to C} + x_{TTA} \times \mu_{A\to G} = x_{CTA} \times \mu_{C\to T} + x_{TTG} \times \mu_{G\to A} \\ x_{TTG} \times \mu_{T\to C} + x_{TTG} \times \mu_{G\to A} = x_{CTG} \times \mu_{C\to T} + x_{TTA} \times \mu_{A\to G} \\ x_{CTA} \times \mu_{C\to T} + x_{CTA} \times (\mu_{A\to C} + \mu_{A\to G} + \mu_{A\to T}) \\ \qquad = x_{TTA} \times \mu_{T\to C} + x_{CTC} \times \mu_{C\to A} + x_{CTG} \times \mu_{G\to A} + x_{CTT} \times \mu_{T\to A} \\ x_{CTG} \times \mu_{C\to T} + x_{CTG} \times (\mu_{G\to A} + \mu_{G\to C} + \mu_{G\to T}) \\ \qquad = x_{TTG} \times \mu_{T\to C} + x_{CTA} \times \mu_{A\to G} + x_{CTC} \times \mu_{C\to G} + x_{CTT} \times \mu_{T\to G} \\ x_{CTC} \times (\mu_{C\to A} + \mu_{C\to G} + \mu_{C\to T}) = x_{CTA} \times \mu_{A\to C} + x_{CTG} \times \mu_{G\to C} + x_{CTT} \times \mu_{T\to C} \\ x_{CTT} \times (\mu_{T\to A} + \mu_{T\to C} + \mu_{T\to G}) = x_{CTA} \times \mu_{A\to T} + x_{CTC} \times \mu_{C\to T} + x_{CTG} \times \mu_{G\to T} \end{cases} . \quad (5)$$

Category three – Four-fold degenerate codons:

$$\begin{cases} x_{pqA} \times (\mu_{A\to C} + \mu_{A\to G} + \mu_{A\to T}) = x_{pqC} \times \mu_{C\to A} + x_{pqG} \times \mu_{G\to A} + x_{pqT} \times \mu_{T\to A} \\ x_{pqC} \times (\mu_{C\to A} + \mu_{C\to G} + \mu_{C\to T}) = x_{pqA} \times \mu_{A\to C} + x_{pqG} \times \mu_{G\to C} + x_{pqT} \times \mu_{T\to C} \\ x_{pqG} \times (\mu_{G\to A} + \mu_{G\to C} + \mu_{G\to T}) = x_{pqA} \times \mu_{A\to G} + x_{pqC} \times \mu_{C\to G} + x_{pqT} \times \mu_{T\to G} \\ x_{pqT} \times (\mu_{T\to A} + \mu_{T\to C} + \mu_{T\to G}) = x_{pqA} \times \mu_{A\to T} + x_{pqC} \times \mu_{C\to T} + x_{pqG} \times \mu_{G\to T} \end{cases} . \quad (6)$$

Category four – Codons for Isoleucine:

$$\begin{cases} x_{ATA} \times (\mu_{A\to C} + \mu_{A\to T}) = x_{ATC} \times \mu_{C\to A} + x_{ATT} \times \mu_{T\to A} \\ x_{ATC} \times (\mu_{C\to A} + \mu_{C\to T}) = x_{ATA} \times \mu_{A\to C} + x_{ATT} \times \mu_{T\to C} \\ x_{ATT} \times (\mu_{T\to A} + \mu_{T\to C}) = x_{ATA} \times \mu_{A\to T} + x_{ATC} \times \mu_{C\to T} \end{cases} . \quad (7)$$

Category five – C/T-ended two-fold degenerate codons:

$$x_{pqC} \times \mu_{C\to T} = x_{pqT} \times \mu_{T\to C} . \quad (8)$$

Category six– A/G-ended two-fold degenerate codons:

$$x_{pqA} \times \mu_{A\to G} = x_{pqG} \times \mu_{G\to A} . \quad (9)$$

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Information about the analytical solutions of the above equations is shown in the Supplemental Information (Analytical Solutions of Equation Systems Used to Estimate *μ* Values). Using these solutions, with a given set of *μ* values and counts of amino acid residues, we can calculate the expected counts of synonymous codons. For example, the expected counts of codons for Lys can be calculated by

$$\begin{cases} Ex_{AAA} = y_{Lys} \times \frac{x_{AAA}}{x_{AAA}+x_{AAG}} = y_{Lys} \times \frac{x_{AAA}}{x_{AAA}+x_{AAA}\times\frac{\mu_{A\to G}}{\mu_{G\to A}}} = y_{Lys} \times \frac{1}{1+\frac{\mu_{A\to G}}{\mu_{G\to A}}} \\ Ex_{AAG} = y_{Lys} \times \frac{x_{AAG}}{x_{AAA}+x_{AAG}} = y_{Lys} \times \frac{x_{AAA}\times\frac{\mu_{A\to G}}{\mu_{G\to A}}}{x_{AAA}+x_{AAA}\times\frac{\mu_{A\to G}}{\mu_{G\to A}}} = y_{Lys} \times \frac{\frac{\mu_{A\to G}}{\mu_{G\to A}}}{1+\frac{\mu_{A\to G}}{\mu_{G\to A}}} \end{cases}, \quad (10)$$

where Ex_{AAA} is the expected count of AAA codon, and y_{Lys} is the count of Lys residues.

Then we can use the expected counts and the observed real counts of all degenerate codons to calculate a χ^2 value. Since for a given protein-coding sequence, the χ^2 value is a function of *μ* values, we can define this function as $\chi^2(\Theta)$, where Θ is a vector describing all *μ* values,

$$\Theta = (\mu_{A\to C}, \mu_{C\to A}, \mu_{A\to G}, \mu_{G\to A}, \mu_{A\to T}, \mu_{T\to A}, \mu_{C\to G}, \mu_{G\to C}, \mu_{C\to T}, \mu_{T\to C}, \mu_{G\to T}, \mu_{T\to G}). \quad (11)$$

For a gene, to estimate *μ* values, we try to minimize the value of $\chi^2(\Theta)$ by changing the elements in Θ, using the sequential least squares programming (SLSQP) algorithm (Kraft, 1988).

Since equations generated from Equation (3) form a system of homogeneous linear equations, it is meaningless to infer exact *μ* values from the minimization of $\chi^2(\Theta)$; rather, only the relative magnitudes of different *μ* values are important for calculating the expected codon counts later. As we used the assumption that the lowest *μ* value is at least 1/100 of the highest, we set the range of *μ* values between 0.001 and 0.1 during the minimization of $\chi^2(\Theta)$. For each gene, the minimum $\chi^2(\Theta)$ is used to calculate p-

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

value. Since we assumed that reference genomic data represent a "wild type" genome, the procedure mentioned above was applied to each individual reference gene.

Codon usage heatmaps

For a protein-coding gene *g*, the relative usage frequency f_{gd} of a codon *d* is defined as

$$f_{gd} = \frac{n_{gd}}{y_{ga}}, \quad (12)$$

where n_{gd} is the count of *d* in *g*, and y_{ga} is the count of amino acid *a* encoded by *d* and its synonymous codons in *g*. It should be noted that codons for Ser are not treated as two codon groups in codon usage heatmaps. As the mechanism of recognizing stop codons is fairly different from recognizing other codons (Brown et al., 2015), and methionine and tryptophan respectively have only one codon, the analysis is restricted to the other 59 codons. Therefore, a 59-dimension vector B_g is used to describe the codon usage pattern of *g*,

$$B_g = \left(f_{g1}, f_{g2}, f_{g3}, \cdots, f_{gd}, \cdots, f_{g59}\right)^T. \quad (13)$$

For a genome containing *M* protein-coding genes, an *M*-dimension vector H_d is used to describe how often a codon *d* is used across these genes,

$$H_d = \left(f_{1d}, f_{2d}, f_{3d}, \cdots, f_{gd}, \cdots, f_{Md}\right). \quad (14)$$

Both B_g's and H_d's are hierarchically clustered. The values of all f_{gd}'s are then color-coded to generate a codon usage heatmap.

The "gplots" (Warnes et al., 2015) package in R was used to generate heatmaps. The method of hierarchical clustering was complete linkage with Euclidean distance

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

measuring the dissimilarities between codons and Spearman's correlation coefficient measuring the similarities between genes.

Identifying genes with tissue-specific expression patterns

Mean adult gene expression data from the FlyAtlas database were used to identify genes with tissue-specific expression patterns. A gene was classified as tissue-specific if its average mRNA level in a specific tissue was at least ten-fold to the tissue with the second highest mRNA level. To reduce redundancy, "head" expression values were excluded from the analysis. Specifically, since "brain" and "thoracic-abdominal ganglion" both belong to the central nervous system (CNS) and many CNS-specific genes are expressed in both places, "adult central nervous system"-specific mRNA level is represented by the higher mRNA level between "brain" and "thoracic-abdominal ganglion". When "adult central nervous system" was involved in the analysis, the "brain" and "thoracic-abdominal ganglion" were excluded, and *vice versa*.

Identifying genes with sex-biased expression patterns

Adult fly expression data from the modENCODE RNA-seq database were used to identify genes with either male- or female-biased expression patterns. Only genes that showed at least ten-fold expression in one sex relative to the other were defined as sex-biased.

Hypergeometric tests for gene set enrichment

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The hypergeometric tests were performed using the online tool at http://www.rothsteinlab.com/tools/apps/hyper_geometric_calculator.

Gene ontology analysis

Gene ontology (GO) analysis (Gene Ontology Consortium, 2015) of *Drosophila* genes was performed by using online tools (http://geneontology.org/). Category enrichments were determined by comparing term frequencies between each gene cluster and the whole genome, followed by a Bonferroni correction.

Animals

Fruit flies (*D. melanogaster*) were raised on corn syrup-soy food (Archon Scientific) at 25°C and 60% relative humidity with a 12-hour light/dark cycle. Synthetic cDNAs encoding EGFP and mCherry fluorescent proteins were from IDT Inc. Full sequences of *mCherry*^{Common}, *EGFP*^{Common}, and *EGFP*^{RareKEQ} alleles are in the Supplemental Information (Sequences of Fluorescent Reporter cDNAs). To generate animals that express each transgenic allele under UAS control, synthetic cDNAs were cloned into the EcoRI/NotI cloning sites of the pUASTattB plasmid (Bischof et al., 2007). Since the *EGFP*^{RareKEQ} allele contains one EcoRI site in the middle, digestion time was shortened to less than 20 minutes to allow incomplete digestion. The UAS-*RFP*^{Common} transgene was inserted into a chromosome II landing site (Bloomington #24483), and the UAS-*EGFP*^{Common} and UAS-*EGFP*^{RareKEQ} transgenes were inserted into the same chromosome III landing site (Bloomington #24749). Double homozygotic lines UAS-*RFP*^{Common}; UAS-*EGFP*^{Common} and UAS-*RFP*^{Common}; UAS-*EGFP*^{RareKEQ} were generated and then crossed to

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

*Act*5*C*-GAL4/*SM*6 (Duncan lab, Washington University). Canton-S isoA flies were used for tRNA northern blot.

Analyses of tRNA gene expression analyses

Northern blots were used to measure relative tRNA abundance in different body parts. Total RNA was extracted from four pools of 10 dissected male tissues (head, reproductive system, and remaining thorax and abdominal parts) and 10 whole male flies with the TRIzol reagent (Invitrogen Catalog # 15596-026). Probe sequences were: $tRNA^{Lys}{}_{CTT}$, 5'-AACGTGGGGCTCGAACCCACGACCCTGA-3'; $tRNA^{Lys}{}_{TTT}$, 5'-GAACAGGGACTTGAACCCTGGACCCTTG-3'. Probes were labeled by ^{32}P using T4 PNK (NEB Catalog # M0201S). Signals were measured and normalized to total tRNA signals using the BIO-RAD Quantity One 1-D analysis software. ANOVA followed by SNK *post hoc* test was used to compare tRNA levels between samples.

Real-time qRT-PCR

The mRNA expression levels of reporter genes in whole four-day-old male flies were quantified by using real-time qRT-PCR, following previously published methods (Lu et al., 2012; Hill et al., 2017). For *RpL*32, the forward primer was 5'-CACCAAGCACTTCATCCG-3', and the reverse primer was 5'-TCGATCCGTAACCGATGT-3'. For $EGFP^{Common}$ and $EGFP^{RareKEQ}$, the forward primers were respectively 5'-AACTTCAAGATCCGCCACAAC-3' and 5'-AACTTCAAAATCCGCCACAAC-3', while these alleles shared the same reverse primer 5'-GTGCTCAGGTAGTGGTTATCG-3'.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Quantitative reporter gene imaging

Male reproductive and gut tissues from four-day-old adult male flies that express either the *EGFP*^{Common} or *EGFP*^{RareKEQ} were dissected in chilled PBS and mounted for imaging on a Nikon A1Si laser scanning confocal microscope with a 20X oil objective (n=5 samples per genotype). All images were taken within 10 minutes of dissection. Single plane fluorescent images of the AGSC were used to estimate EGFP expression levels of each allele. Similar images of the HPZ were used as generic tissue controls. The NIS-Element Ar software was used to capture EGFP and RFP signals and generate channel-merged images. Normalized EGFP signals from each image were quantified using the Fiji image processing software (Schindelin et al., 2012). The tissue-specific effect of the EGFP codon usage was analyzed by comparing the normalized EGFP signals in either the AGSC or HPZ between genotypes with a two-tailed unpaired Student's *t*-test.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SUPPLEMENTAL INFORMATION TITLES AND LEGENDS

Supplemental Figure Legends

Figure S1. Percentages of genes carrying signatures of natural selection on synonymous codon usage in diverse eukaryotic species, related to Figure 1. For shown species, all annotated protein-coding genes that passed data pre-processing were included. The percentages of genes carrying selection signatures were corrected by false discovery rate (FDR = 0.05). Species are stacked by phylum and kingdom. Blue, unicellular species; black, multicellular species; red, can be either unicellular or multicellular.

Figure S2. Clustered codon usage patterns across genes carrying signatures of selection on gene-specific codon usage bias in *A. thaliana*, related to Figure 2. Hierarchical clustering of all genes identified as under selection for gene-specific codon usage bias in *A. thaliana.* Analyses as in Figure 2.

Figure S3. Clustered codon usage patterns across genes carrying signatures of selection on gene-specific codon usage bias in *S. cerevisiae*, related to Figure 2. Hierarchical clustering of all genes identified as under selection for gene-specific codon usage bias in *S. cerevisiae.* Analyses as in Figure 2.

Figure S4. Clustered codon usage patterns across genes carrying signatures of selection on gene-specific codon usage bias in *C. elegans*, related to figure 2. Hierarchical clustering of all genes identified as under selection for gene-specific codon usage bias in *C. elegans.* Analyses as in Figure 2.

Figure S5. Clustered codon usage patterns across genes carrying signatures of selection on gene-specific codon usage bias in *H. sapiens*, related to figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/292938; this version posted May 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hierarchical clustering of all genes identified as under selection for gene-specific codon usage bias in *H. sapiens.* Analyses as in Figure 2.

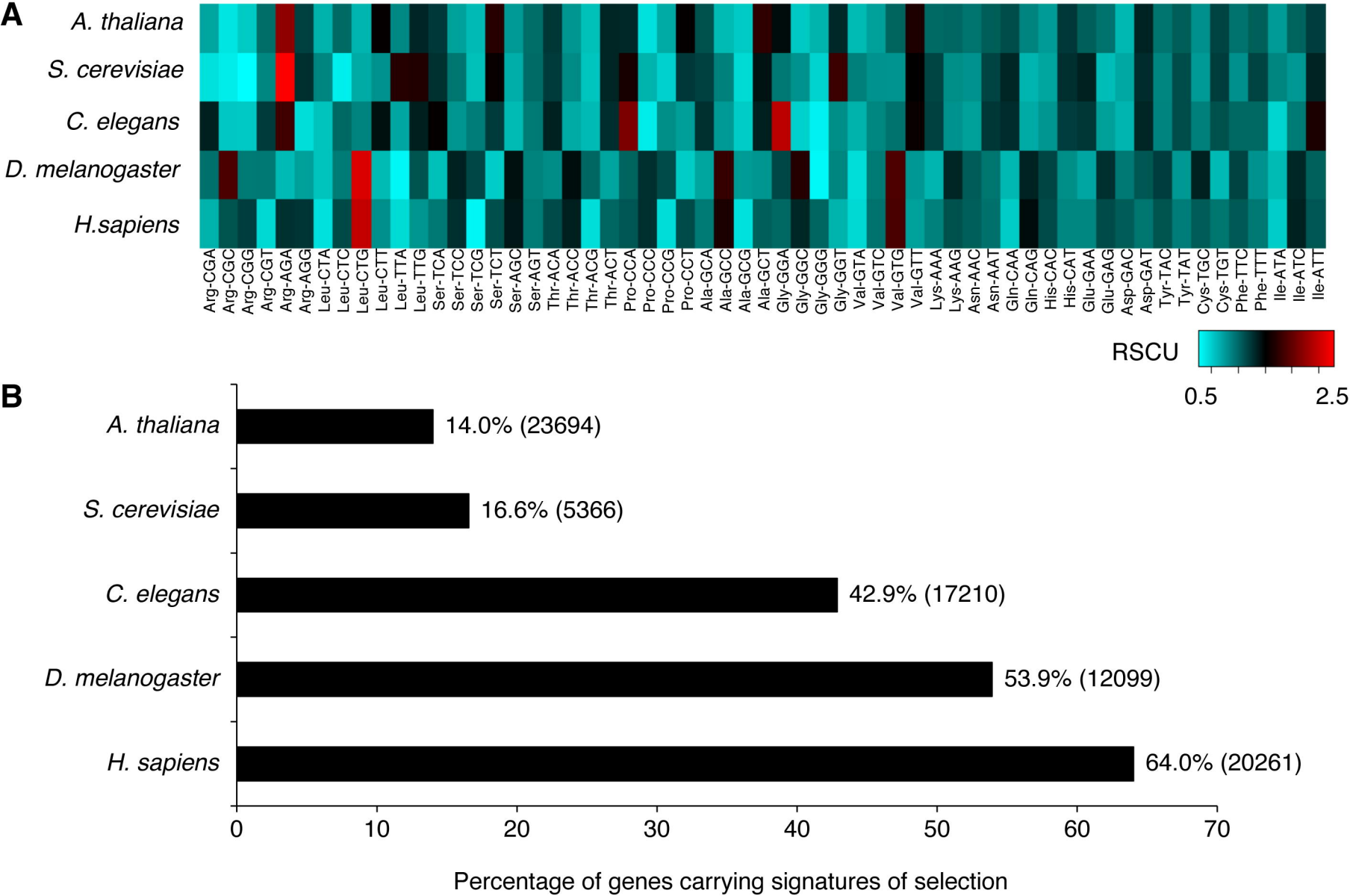
Supplemental Table Legend

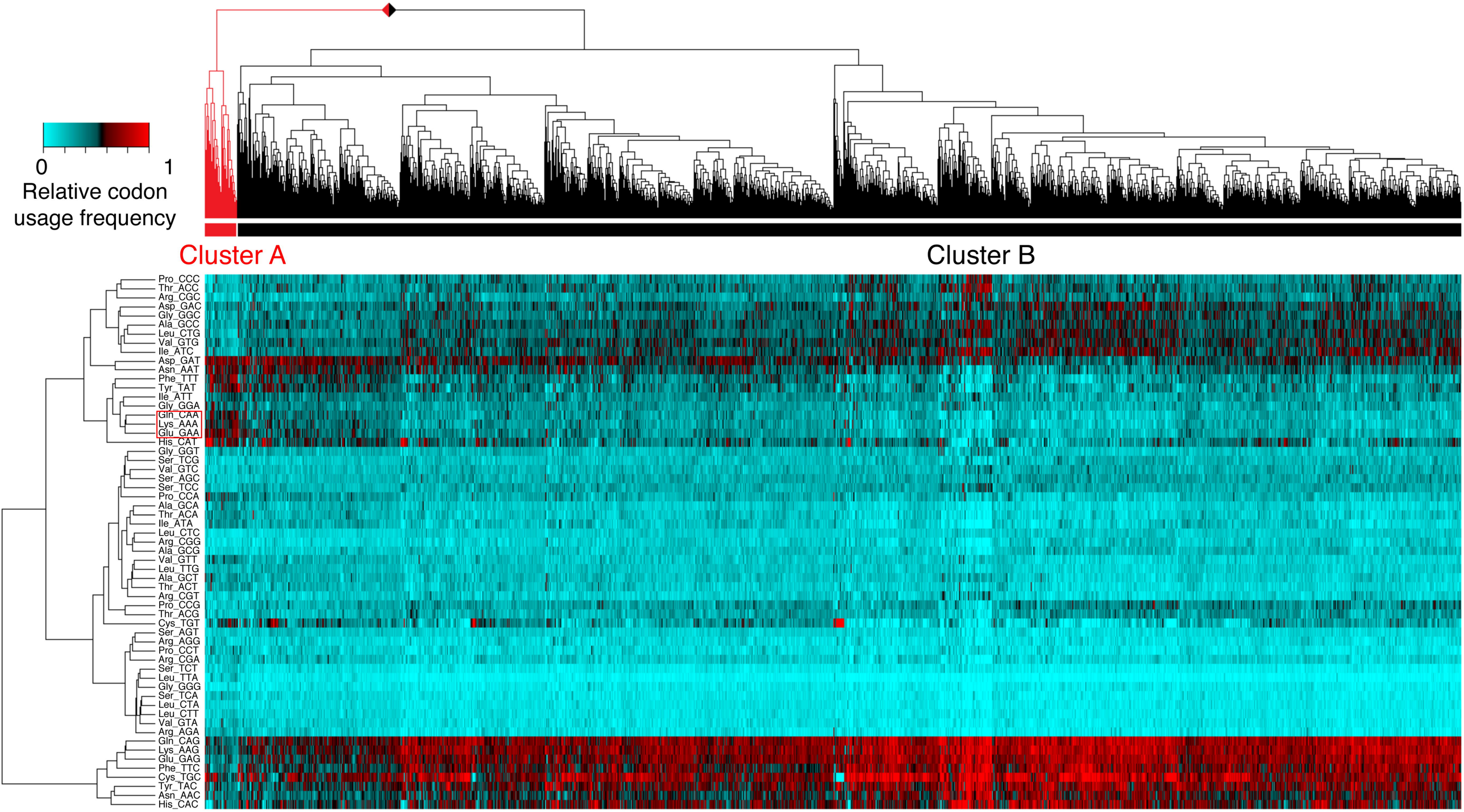
Table S1. Gene ontology analysis of gene clusters defined by codon usage patterns, related to Figure 2, Tables 1-2.

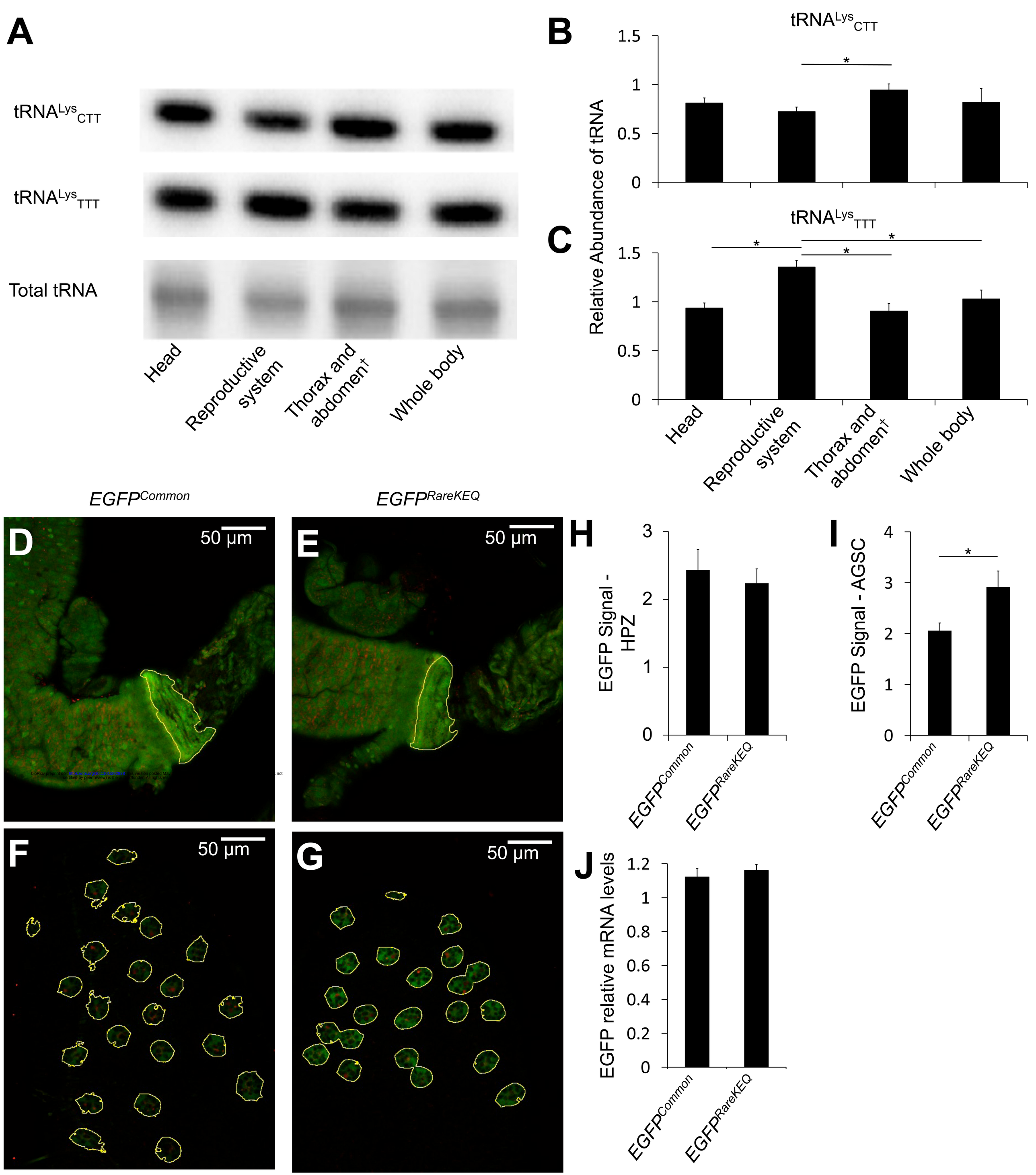
Clusters A and B are defined by their codon usage patterns, and only genes carrying signatures of natural selection on synonymous codon usage are included, as shown in Fig. 2. Molecular function, cellular component, and biological process GO terms are analyzed. Only significant over- or under-representations are shown. Fold enrichment is relative to the entire genome. We did not obtain informative results when analyzing biological process GO terms of Cluster A genes.

Analytical Solutions of Equation Systems Used to Estimate *μ* Values, Related to Figure 1.

Sequences of Fluorescent Reporter cDNAs, Related to Figure 3.


A
A. thaliana
S. cerevisiae
C. elegans
D. melanogaster
H.sapiens
Arg-CGA
Arg-CGC
Arg-CGG
Arg-CGT
Arg-AGA
Arg-AGG
Leu-CTA
Leu-CTC
Leu-CTG
Leu-CTT
Leu-TTA
Leu-TTG
Ser-TCA
Ser-TCC
Ser-TCG
Ser-TCT
Ser-AGC
Ser-AGT
Thr-ACA
Thr-ACC
Thr-ACG
Thr-ACT
Pro-CCA
Pro-CCC
Pro-CCG
Pro-CCT
Ala-GCA
Ala-GCC
Ala-GCG
Ala-GCT
Gly-GGA
Gly-GGC
Gly-GGG
Gly-GGT
Val-GTA
Val-GTC
Val-GTG
Val-GTT
Lys-AAA
Lys-AAG
Asn-AAC
Asn-AAT
Gln-CAA
Gln-CAG
His-CAC
His-CAT
Glu-GAA
Glu-GAG
Asp-GAC
Asp-GAT
Tyr-TAC
Tyr-TAT
Cys-TGC
Cys-TGT
Phe-TTC
Phe-TTT
Ile-ATA
Ile-ATC
Ile-ATT
RSCU
0.5
2.5
B
A. thaliana
14.0% (23694)
S. cerevisiae
16.6% (5366)
C. elegans
42.9% (17210)
D. melanogaster
53.9% (12099)
H. sapiens
64.0% (20261)
0
10
20
30
40
50
60
70
Percentage of genes carrying signatures of selection


0
1
Relative codon usage frequency
Cluster A
Cluster B
Pro_CCC
Thr_ACC
Arg_CGC
Asp_GAC
Gly_GGC
Ala_GCC
Leu_CTG
Val_GTG
Ile_ATC
Asp_GAT
Asn_AAT
Phe_TTT
Tyr_TAT
Ile_ATT
Gly_GGA
Gln_CAA
Lys_AAA
Glu_GAA
His_CAT
Gly_GGT
Ser_TCG
Val_GTC
Ser_AGC
Ser_TCC
Pro_CCA
Ala_GCA
Thr_ACA
Ile_ATA
Leu_CTC
Arg_CGG
Ala_GCG
Val_GTT
Leu_TTG
Ala_GCT
Thr_ACT
Arg_CGT
Pro_CCG
Thr_ACG
Cys_TGT
Ser_AGT
Arg_AGG
Pro_CCT
Arg_CGA
Ser_TCT
Leu_TTA
Gly_GGG
Ser_TCA
Leu_CTA
Leu_CTT
Val_GTA
Arg_AGA
Gln_CAG
Lys_AAG
Glu_GAG
Phe_TTC
Cys_TGC
Tyr_TAC
Asn_AAC
His_CAC


A
tRNA$^{Lys}_{CTT}$
tRNA$^{Lys}_{TTT}$
Total tRNA
Head
Reproductive system
Thorax and abdomen†
Whole body
B
tRNA$^{Lys}_{CTT}$
C
tRNA$^{Lys}_{TTT}$
Relative Abundance of tRNA
1.5
1
0.5
0
*
EGFP^{Common}
EGFP^{RareKEQ}
D
E
F
G
50 µm
H
EGFP Signal - HPZ
3
2
1
0
I
EGFP Signal - AGSC
4
J
EGFP relative mRNA levels
1.2
0.8
0.6
0.4
0.2