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

Running title: Gene-specific codon usage bias

Keywords: *Drosophila melanogaster*, fruit fly, codon usage bias, synonymous mutation, neutral mutation

ABSTRACT

Genomic data from diverse taxa indicate that synonymous codon usage frequencies are often not random. Although the molecular evolutionary forces that shape codon usage remain poorly understood, synonymous mutations are generally assumed to be neutral because they do not affect protein sequences. However, empirical evidence indicates 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 spatial and sex-related regulation of protein expression in *Drosophila melanogaster*. These data indicate that natural selection shapes gene-specific codon usage biases, and therefore, it represents an overlooked genomic feature that plays an important role in the spatial and temporal regulation of gene functions. Hence, the broadly accepted dogma that synonymous mutations are generally functionally neutral should be reevaluated.

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

50 and McInerney, 2013; Yang et al., 2015). In contrast, whether genetic drift and natural
51 selection play equally important roles in driving gene-specific codon usage patterns
52 remains controversial (Akashi, 1995; Rao et al., 2011; Doherty and McInerney, 2013;
53 Kessler and Dean, 2014). Specifically, since synonymous mutations do not affect protein
54 sequences, and therefore are usually assumed to have no impact on their structure or
55 biochemical function, the effect of natural selection on gene-specific codon usage bias is
56 often dismissed as weak (King and Jukes, 1969; Kimura, 1968; Kimura and Ohta, 1974;
57 Echave, Spielman, and Wilke, 2016).

58 Nonetheless, a series of biochemical and molecular studies of specific protein-coding
59 genes seem to present mechanistic challenges to the broadly accepted dogma that
60 synonymous mutations are generally functionally neutral. First, gene-specific
61 synonymous codon usage patterns have been shown to be important for the function of
62 some protein-coding genes by affecting features such as translational efficacy (Ikemura,
63 1981; Akashi, 1994; Ko et al., 2005), mRNA secondary structures (Hasegawa,
64 Yasunaga, and Miyata, 1979), mRNA splicing patterns (Pagani, Raponi, and Baralle,
65 2005; Parmley and Hurst, 2007), miRNA binding accessibility (Gu et al., 2012), co-
66 translational protein folding (Zhang, Hubalewska, and Ignatova, 2009; Zhou et al., 2013;
67 Fu et al., 2016), and post-translational modifications (Kimchi-Sarfaty et al., 2007; Zhou et
68 al., 2013; Quax et al., 2015; Fu et al., 2016). Second, the temporal regulation of protein
69 synthesis in several viruses in the *Herpesviridae* group has been shown to depend on
70 viral-specific codon usage patterns that are different from those of the host cell (Shin et
71 al., 2015). Third, high protein levels of some housekeeping genes depend on the biased
72 usage of common synonymous codons (Stenico, Lloyd, and Sharp, 1994; Ma et al.,
73 2014). Some of these anecdotal examples could be explained by the "translational

74 selection theory", which stipulates that “optimal” protein translation rate and accuracy
75 partly depend on the availability and affinity of cognate tRNAs, and therefore,
76 necessitates the evolution of codon usage patterns that match the relative abundance of
77 specific tRNAs in the relevant cellular pools (Ikemura, 1981; Shields and Sharp, 1987;
78 Slimko and Lester, 2003; dos Reis, Savva, and Wernisch, 2004). Together, these
79 studies suggest that gene-specific codon usage biases might be playing a general, yet
80 underappreciated, role in the spatial and temporal regulation of gene activity, and
81 therefore should be affected by natural selection.

82 Here we combine theoretical and experimental approaches to test the hypothesis that
83 natural selection has played a significant and important role in shaping gene-specific
84 codon usage patterns. To test this hypothesis, we developed a statistical approach for
85 identifying signatures of natural selection on gene-specific codon usage patterns. Using
86 this approach, we demonstrate that a considerable proportion of protein-coding genes
87 from diverse eukaryotic species display molecular signatures of natural selection on
88 biased gene-specific codon usage patterns. Additionally, we use the well-annotated
89 genome of *Drosophila melanogaster* to identify genome-wide signatures of selection on
90 gene-specific codon usage patterns, and demonstrate that some combinations of unique
91 rare synonymous codons are associated with specific biological functions, including
92 spatial and male-biased protein expression patterns. Together, our results indicate that
93 in contrast to the broadly accepted dogma, natural selection plays a significant role in
94 determining gene-specific codon usage biases in diverse eukaryotic species, and that
95 gene-specific codon usage biases are important for the spatial and temporal regulation
96 of some protein-coding genes. Together, these data indicate that the broadly accepted

assumption that that synonymous mutations are in general functionally and evolutionarily neutral should be reevaluated.

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 "selection-mutation-drift" model, which evaluates the relative contribution of different molecular forces to the observed codon usage bias by estimating the selection coefficients for each possible nucleotide substitution, our approach is based on statistically rejecting the null hypothesis that ‘Synonymous mutations are neutral’. While our approach does not provide specific

120 quantitative estimates for the level of selection on each individual gene, it statisticaly
121 identifies specific genes whose biased codon usage patterns have been likely impacted
122 by natural selection.

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

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

140
$$\begin{aligned}&\sum_i \frac{2s_{pqi\to pqr}}{1-e^{-2Ns_{pqi\to pqr}}} \times x_{pqi} \times \mu_{i\to r} + \sum_j \frac{2s_{jqr\to pqr}}{1-e^{-2Ns_{jqr\to pqr}}} \times x_{jqr} \times \mu_{j\to p} \\ &= \sum_i \frac{2s_{pqr\to pqi}}{1-e^{-2Ns_{pqr\to pqi}}} \times x_{pqr} \times \mu_{r\to i} + \sum_j \frac{2s_{pqr\to jqr}}{1-e^{-2Ns_{pqr\to jqr}}} \times x_{pqr} \times \mu_{p\to j}\end{aligned} \quad (1)$$

141 Variables *p*, *q*, and *r* denote the specific nucleotide identities (A, C, G, or T) of the first,
142 second, and third positions respectively in each codon. Codons synonymous to codon
143 *pqr* that vary at either the first or third positions are denoted by codons *jqr* or *pqi*
144 respectively. Therefore, $s_{pqi\to pqr}$, for example, denotes the selection coefficient of the *pqi*-
145 to-*pqr* mutation. *N* denotes effective chromosomal population size (Yang and Nielsen,
146 2008). x_{pqr} denotes the count of codon *pqr* in a focal gene across the entire population,
147 and $\mu_{i\to r}$ denotes an estimate for the *i*-to-*r* mutation rate. Consequently, the probability
148 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).
149 The amino acid serine represents a unique case because it is encoded by six
150 synonymous codons that belong to two independent codon groups, which are not
151 interchangeable via single synonymous mutations. Therefore, in our model we treat the
152 ser codon groups AGC, AGT and TCA, TCC, TCG, TCT as if they are coding two
153 independent amino acids (Gilchrist et al., 2015).

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

157
$$\sum_i \lim_{s_{pqi\to pqr}\to 0} \frac{2s_{pqi\to pqr}}{1-e^{-2Ns_{pqi\to pqr}}} \times x_{pqi} \times \mu_{i\to r} + \sum_j \lim_{s_{jqr\to pqr}\to 0} \frac{2s_{jqr\to pqr}}{1-e^{-2Ns_{jqr\to pqr}}} \times x_{jqr} \times \mu_{j\to p}$$
$$= \sum_i \lim_{s_{pqr\to pqi}\to 0} \frac{2s_{pqr\to pqi}}{1-e^{-2Ns_{pqr\to pqi}}} \times x_{pqr} \times \mu_{r\to i} + \sum_j \lim_{s_{pqr\to jqr}\to 0} \frac{2s_{pqr\to jqr}}{1-e^{-2Ns_{pqr\to jqr}}} \times x_{pqr} \times \mu_{p\to j}$$

158
$$\Rightarrow$$

159
$$\sum_i \frac{1}{N} \times x_{pqi} \times \mu_{i\to r} + \sum_j \frac{1}{N} \times x_{jqr} \times \mu_{j\to p} = \sum_i \frac{1}{N} \times x_{pqr} \times \mu_{r\to i} + \sum_j \frac{1}{N} \times x_{pqr} \times \mu_{p\to j} \quad (2)$$

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

162 $$\sum_i x_{pqi} \times \mu_{i\to r} + \sum_j x_{jqr} \times \mu_{j\to p} = \sum_i x_{pqr} \times \mu_{r\to i} + \sum_j x_{pqr} \times \mu_{p\to j} \quad (3)$$

163 Subsequently, if the null hypothesis that all synonymous mutations are neutral is true
164 then the impacts of genetic drift, represented by the effective chromosomal population
165 size *N*, and natural selection, represented by selection coefficient *s*, are canceled as
166 shown above, and therefore, the observed codon usage pattern could be explained
167 solely by mutational bias. In contrast, if the null hypothesis is rejected then we must
168 accept the alternative hypothesis, which indicates natural selection has had a significant
169 impact on the observed gene-specific codon usage bias.

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

184 Next, we used our approach to analyze the well-annotated genomes of five
185 representative eukaryotic species, *Drosophila melanogaster*, *Arabidopsis thaliana*,
186 *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Homo sapiens*. Initial analyses
187 of the overall relative synonymous codon usage (RSCU) (Sharp et al., 1988) revealed
188 that these species utilize markedly different genome-level patterns of synonymous
189 codon usage biases (Fig. 1A). Next, we searched for protein-coding genes that exhibit
190 signatures of selection on gene-specific codon usage bias by utilizing our new statistical
191 approach. We found that after correcting for false discovery rate (FDR = 0.05)
192 (Benjamini and Hochberg, 1995), the percentage of protein-coding genes that carry
193 signatures of selection on gene-specific synonymous codon usage bias varied between
194 14% and 64% across these five representative genomes (Fig. 1B). Additional analyses
195 of available eukaryotic genomes revealed that in 35 out of 40 species analyzed, at least
196 10% of the protein-coding genes carried signatures of natural selection on synonymous
197 codon usage patterns, independent of whether they are unicellular or multicellular (Fig.
198 S1). These results indicate that the overall impact of natural selection on gene-specific
199 patterns of codon usage is broader than would be expected if synonymous substitutions
200 were mostly functionally neutral. Our results also suggest that the relative impact of
201 selection on gene-specific codon usage bias is not constant across the eukaryotic
202 phylogeny, and therefore, might be the result of diverse, clade-specific selective forces.

203 As currently framed, our quantitative approach is based on a set of conservative
204 assumptions, which should make the model applicable to as many genomes as possible.
205 However, we acknowledge that some of these assumptions might not represent the
206 correct biological context for all species or ecological conditions. We also anticipate that
207 the assumption that adjacent nucleotides have no effect on the mutation rates of focal

208 nucleotides could lead to increased false-positive rates (Bérard and Guéguen, 2012).
209 However, we anticipate that this possible source for false-positives is compensated for
210 by the assumption that background nucleotide composition of nearby chromosomal
211 regions is constant (Oliver et al, 2001), which is further alleviated by the conservative
212 use of relaxed constraints for estimating μ values (Drake et al., 1998). Nevertheless,
213 false positive rates produced by our model could be further reduced by having more
214 precise estimates of the μ parameters, which could be achieved by using population-
215 level genetic variations data, and more specific information about species-specific
216 mutation rates, and how they might be impacted by the identities of adjacent nucleotides.

217

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

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

227 Due to its well-annotated genome and wealth of available genetic and phenotypic data,
228 we chose to use the genetically tractable species *D. melanogaster* for functional
229 analyses of genes that are under selection for biased codon usage patterns. Hierarchical
230 gene-clustering analysis revealed that genes with identified signatures of selection on

231 biased codon usage could be assigned to one of two major clusters (Clusters A and B;
232 Fig. 2). Further analysis of Cluster A revealed that synonymous codon usage in these
233 genes is enriched for the rare codons Lys-AAA, Glu-GAA, Gln-CAA, Phe-TTT, Tyr-TAT,
234 and His-CAT, relative to their usage in Cluster B genes. Similar clustering analyses of
235 the genomes of *A. thaliana*, *S. cerevisiae*, *C. elegans*, and *H. sapiens* revealed gene-by-
236 codon clustering in all five species, likely representing species-specific patterns of
237 selection (Figs. S2-5).

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

248 In addition, analyses of Gene Ontology terms (Gene Ontology Consortium, 2015)
249 revealed that genes that encode extracellular matrix proteins are overrepresented, and
250 genes that encode cytoplasmic proteins are underrepresented, in Cluster A (Table S1).
251 We also found that genes annotated as encoding odorant-binding proteins, a class of
252 secreted proteins, as well as extracellular space proteins, are underrepresented in
253 Cluster B. Together, these analyses suggest that in *D.* melanogaster, natural selection

254 on the increased usage of specific combinations of rare synonymous codons might be
255 important for male-specific functions of some genes.

256

257 **Biased gene-specific codon usage could modulate spatial protein expression**
258 **patterns**

259 One possible explanation for the association between certain biased gene-specific
260 codon usage patterns and tissue-specific expression is that the selected specific rare
261 synonymous codons match tissue-specific tRNA pools, and therefore, enhances the
262 translational efficiency of some protein in specific tissues or cell types (Zalucki, Gittins
263 and Jennings, 2008; Pechmann, Chartron and Frydman, 2014). Similar associations
264 between gene-specific codon usage bias and organ-enriched gene expression patterns
265 have been previously reported in humans (Plotkin, Robins, and Levine, 2004; Dittmar,
266 Goodenbour and Pan, 2006). Therefore, we next tested the hypothesis that the specific
267 pattern of codon usage bias exhibited by Cluster A genes contributes to increased
268 translation efficiency in the male reproductive system.

269 Our clustering analysis revealed that male-enriched genes in Cluster A preferentially use
270 the rare codons Lys-AAA, Gln-CAA, Glu-GAA, Phe-TTT, Tyr-TAT, and His-CAT (Fig. 2).
271 However, only the first three codons are recognized by exactly matched tRNA
272 anticodons, while the other three rare codons share the same tRNAs with their more
273 commonly used synonymous codons, with a mismatch (wobble) at the third codon
274 position (Crick, 1966; Chan and Lowe, 2009), which suggests these codons are less
275 likely to contribute to tissue-specific translation efficiency. Therefore, we next analyzed
276 the specific usage of rare codons AAA, CAA, and GAA in all genes that show enriched

277 expression in the male reproductive system, independent of whether these genes have
278 passed the initial statistical threshold for the detection of selection on overall codon
279 usage patterns. We found that these specific rare synonymous codons are indeed
280 overrepresented in male-reproductive-system-specific genes (Table 3). Thus, not only
281 are genes preferring AAA, CAA, and GAA more likely to be male-reproductive-system-
282 specific, but also genes specifically transcribed in male reproductive system use AAA,
283 CAA, and GAA more often than their expected genome average usage.

284 Selection on the biased usage of codons AAA, CAA, and GAA in associate with male-
285 specific gene activity should depend, at least in part, on the increased expression of their
286 cognate tRNAs in the relevant tissues. In accordance with this, northern blot analyses of
287 tissue-specific tRNA expression revealed that $tRNA^{Lys}_{TTT}$ is enriched in the male
288 reproductive organs, further supporting the hypothesis that spatial regulation of some
289 tRNA genes that correspond to rare codons could contribute to tissue-specific increase
290 in protein translation efficiency (Fig. 3A-C). The expression patterns of the tRNAs that
291 match Glu-GAA and Gln-CAA were not investigated because they are too similar to their
292 common tRNA counterparts to allow detection by hybridization probes (de Muro, 2008).
293 Nevertheless, these data indicate that some *Drosophila* genes have evolved a specific
294 codon usage bias that specifically drives their levels and/or activity in the male
295 reproductive system.

296 Next, we empirically tested whether specific combinations of rare codons are sufficient to
297 drive tissue-specific translational efficiency by generating transgenic flies that express
298 different alleles of the EGFP reporter under the control of the UAS-GAL4 binary
299 expression system (Brand and Perrimon, 1993). One allele was comprised of common
300 codons for all amino acids ($EGFP^{Common}$), and the second allele was comprised of rare

301 codons AAA, GAA, and CAA for amino acids Lys, Glu, and Gln respectively
302 (EGFP^{RareKEQ}), which represent about 18% of the total residues in EGFP. Each EGFP
303 allele was also co-expressed with an RFP reporter encoded by common codons, which
304 served as a transgene expression control. Both transgenes were driven by the
305 ubiquitous *Act5C*-GAL4 driver (Kopp, Blackman, and Duncan, 1999). To evaluate the
306 effect of codon usage bias on the spatial pattern of both EGFP alleles *in vivo*, we
307 compared the two genotypes by measuring the ratios of RFP-normalized EGFP signals
308 in the accessory gland secondary cells (AGSC) and in the hindgut proliferation zone
309 (HPZ) (Takashima et al., 2013) in the pylorus region. We found that the signal from the
310 *EGFP*^{RareKEQ} allele in the AGSCs was significantly higher than that of the *EGFP*^{Common}
311 allele. In contrast, both alleles produced similar signals in the HPZ (Fig. 3D-I). Because
312 the overall transcript levels of both alleles are similar (Fig. 3J), these results suggest that
313 some specific combinations of synonymous rare codons could impact the spatial
314 regulation of protein translation in the *Drosophila* male reproductive system.

315

316 **DISCUSSION**

317 In contrast to the broadly accepted dogma in evolutionary biology and population
318 genetics, our theoretical and empirical studies indicate that synonymous mutations
319 should not be considered neutral in general. By using multiple complementary
320 approaches, we demonstrate here that the impact of natural selection on gene-specific
321 codon usage bias is common across the eukaryotic phylogeny, and is likely to play a
322 much more prominent role in defining gene-specific codon usage bias than was
323 previously assumed. Furthermore, our data indicate that in multicellular organisms such

324 as the fruit fly, natural selection is likely to optimize codon usage patterns to match
325 specific spatial and temporal constraints on gene function, and therefore, what might
326 represent an optimal codon usage for individual genes does not necessarily mean
327 common codons should be expected at every position. Our findings also indicate that
328 previous reports about the phenotypic consequences of synonymous mutations are not
329 anecdotal rare cases. Instead, these cases likely represent a fundamental biological
330 principle associated with the spatial and temporal gene activity landscapes in eukaryotes.
331 In addition, our analyses of gene-specific codon usage biases in *D. melanogaster*
332 suggest that gene-specific codon usage bias is heterogeneous across protein coding
333 genes in a single species, and therefore, likely to be the outcome of diverse selective
334 pressures associated with different biological functions. Furthermore, our data suggest
335 that different classes of rare synonymous codons exhibit differential sensitivities to
336 natural selection.

337 Consequently, our discovery that gene-specific codon usage patterns are often shaped
338 by natural selection indicates that alternative synonymous codons serve important, yet
339 often overlooked, biological functions. Although several previous studies have used
340 statistical approaches to suggest that synonymous mutations in specific genes and
341 species might not be neutral (Berger, 1977; Eyre-Walker, 1991; Stenico, Lloyd, and
342 Sharp, 1994; Lawrie et al., 2013), our current study provides the first genome-scale
343 comprehensive view of the overall impact of natural selection on gene-specific
344 synonymous codon usage patterns across diverse eukaryotic taxa. Thus, our unbiased
345 and broadly applicable statistical approach has uncovered a fundamental genomic
346 organizational principle that is sensitive to natural selection, and likely represents an
347 important component of the genotype-phenotype axis in diverse eukaryotic species and

348 biological contexts. Therefore, it is likely that investigations, which rely on the
349 assumption that synonymous mutations are neutral, often overestimate or underestimate
350 the prevalence of selection on specific protein-coding genes, which could impact the
351 interpretations of the role of these genes in genotype-phenotype relationships at the
352 level of populations and the evolutionary timescale in particular (Chen et al., 2010), and
353 the genome evolution in general (Campos et al., 2012; Matsumoto et al., 2016).
354 Specifically, the prevailing assumption that most synonymous mutations are neutral, and
355 therefore represent a robust and quantitative measure of selection-independent
356 evolutionary rates of nucleotide substitutions, needs to be revised. This is especially
357 important in the context of the commonly used tests for identifying molecular signatures
358 of selection in specific protein-coding genes, and the classification of modes of natural
359 selection (*e.g.*, "positive selection" vs. "negative selection") (McDonald and Kreitman,
360 1991; Nielsen, 2001; Echave, Spielman, and Wilke, 2016).

361 Our data also indicate that, similarly to *Drosophila*, many human protein-coding genes
362 carry signatures of selection on their biased codon usage patterns (Fig. 1). This finding
363 could have broad implications for studies of genetic variants underlying quantitative
364 phenotypes in human populations. Consequently, assuming that synonymous mutations
365 are functionally neutral could lead to a biased view that associations between
366 nonsynonymous SNPs and specific phenotypic values are the only ones important for
367 understanding overall trait variations (Hampe et al., 2007; WTCCC and TASC, 2007; Xu
368 and Taylor, 2009; Fong et al., 2010; Huang et al., 2012). This is supported by a recent
369 data survey of human GWAS data from over 2000 independent studies, which
370 suggested that synonymous and nonsynonymous SNPs have similar likelihoods and
371 effect sizes in terms of association with disease phenotypes (Chen et al., 2010).

372 Although our analysis of human gene-specific codon usage bias was based on data from
373 a single reference genome, the simplest interpretation of our results is that both
374 synonymous and nonsynonymous mutations are likely influenced by natural selection in
375 human populations, and therefore, many synonymous SNPs could still have a significant
376 quantitative and/or qualitative impacts on gene functions and associated phenotypes in
377 health and disease.

378 We also found that gene-specific codon usage bias is not uniform across the genome.
379 Instead, our data suggest that different groups of genes in different species often exhibit
380 enrichment for specific combinations of rare or common codons, which suggests that
381 different combinations of synonymous codons may have been selected for independent
382 biological reasons. In this regard, we show that at least one group of genes with
383 enriched expression in the reproductive system of male *Drosophila* preferentially use the
384 rare codons Lys-AAA, Gln-CAA, and Glu-GAA. By using allelic variants of a reporter
385 gene, we show that the selective usage of these specific three rare codons is sufficient
386 for generating spatially biased protein expression patterns. These findings are consistent
387 with some of the assumptions about the possible roles of codon usage bias, first put
388 forward by the "translational selection theory" (Ikemura, 1981; dos Reis, Savva, and
389 Wernisch, 2004; Plotkin, Robins, and Levine, 2004; Dittmar, Goodenbour, and Pan,
390 2006). Consequently, our study provides a broader context to this fundamental
391 evolutionary theory by emphasizing the possible role of gene-specific codon usage bias
392 in the spatial regulation of proteins in multicellular eukaryotes. However, regulation of
393 spatial protein expression cannot explain all cases of selection on gene-specific codon
394 usage bias because genomes of some unicellular eukaryotes such as the
395 choanoflagellate *Monosiga brevicollis* and the green algae *Chlamydomonas reinhardtii*

396 also include a significant proportion of protein-coding genes exhibit signatures of
397 selection on their codon usage patterns (Fig. S1).

398 Our studies also have important practical implications for experimental systems that
399 depend on the use of heterologous transgenic gene expression. We found that for some
400 genes, the selective usage of synonymous rare codons could increase their translational
401 efficacy in specific tissues and cell types, and therefore, designing experiments with the
402 assumption that transgene codon optimization necessitates the usage of only common
403 synonymous codons could lead to unexpected phenotypic outcomes, and could affect
404 data interpretation, especially for genes with restricted spatial and/ or temporal
405 expression patterns.

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

418 Although our empirical work here has mostly focused on a specific set of rare codons
419 that bind unique tRNA anti-codons, it should be noted that genes in Cluster A (Fig. 2)
420 also preferentially use the rare codons Phe-TTT, Tyr-TAT, and His-CAT. These rare
421 synonymous codons share their tRNA anticodons with their more common synonymous
422 codons via a wobble mechanism (Chan and Lowe, 2009). Theoretical models and
423 empirical data suggest that this specific class of rare codons is likely to slowdown
424 translation due to the lower affinity of common tRNAs to the rare synonymous codons
425 (Ikemura, 1981; dos Reis, Savva, and Wernisch, 2004; Savir and Tlusty, 2013; Brule
426 and Grayhack, 2017). Therefore, we speculate that in contrast to rare codons that bind
427 unique perfectly-matching tRNAs, rare codons that share tRNAs with their more
428 common synonymous codons may have evolved to broadly restrict translational
429 efficiency of some low-abundance proteins, or proteins that require slow folding kinetics
430 associated with specific post-translational modifications, or the binding of other protein
431 and small-molecule partners, chaperones, or co-factors (Komar, Lesnik and Reiss, 1999;
432 Weygand-Durasevic and Ibba, 2010; Angov, 2011; Zhou et al., 2013; Fluman et al.,
433 2014; Pechmann, Chartron and Frydman, 2014; Chaney and Clark, 2015; Quax et al.,
434 2015; Yu et al., 2015; Fu et al., 2016; Zhao, Yu, and Liu, 2017).

435 In summary, the results presented here suggest that in contrast to a widely accepted
436 dogma, synonymous mutations should not be considered functionally and evolutionarily
437 neutral by default. Instead, as was previously shown for a few specific protein-coding
438 genes (Hasegawa, Yasunaga, and Miyata, 1979; Ikemura, 1981; Akashi, 1994; Stenico,
439 Lloyd, and Sharp, 1994; Ko et al., 2005; Pagani, Raponi, and Baralle, 2005; Kimchi-
440 Sarfaty et al., 2007; Parmley and Hurst, 2007; Zhang, Hubalewska, and Ignatova, 2009;
441 Gu et al., 2012; Zhou et al., 2013; Ma et al., 2014; Shin et al., 2015; Fu et al., 2016), our

442 data indicate that selection on synonymous codon usage is likely to serve a more
443 general function in diverse biological processes across Eukaryotes. Nevertheless,
444 although our data indicate that 'synonymous' and 'neutral' mutations are not necessarily
445 the same, we do not suggest that the overall concept of evolutionary neutrality, first
446 described by Darwin in the *On the Origin of Species* as "*Variations neither useful nor*
447 *injurious would not be affected by natural selection*", should be rejected. Instead, we
448 argue that the commonly accepted specific assumption that all synonymous mutations
449 are generally neutral should be reevaluated.

450

451 **ACKNOWLEDGEMENTS**

452 Z.P. dedicates this manuscript to Dr. Yang Zhong of Fudan University, a prominent
453 botanist and evolutionary biologist who has had a great influence on his choice to pursue
454 a research career in mathematical and evolutionary biology. The authors also wish to
455 thank Alan Templeton, Allan Larson, and members of the Ben-Shahar and Zaher
456 laboratories for helpful discussions. They also thank Paula Kiefel, Dianne Duncan, and
457 Ian Duncan for technical assistance with generating transgenic animals and microscopic
458 imaging. Z.P. is supported by the training grant 2T32HG000045 from the NIH. This
459 research has been supported in part by NIH grant NS089834, and NSF grants 1545778
460 and 1707221 to YB, and NIH grant 5R01GM112641 to HZ.

461 **AUTHOR CONTRIBUTIONS**

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

465

466 **CONFLICT OF INTEREST STATEMENT**

467 The authors declare that there are no conflicts of interest with respect to the authorship
468 and/or publication of this article.

469

470 **REFERENCES**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

509 Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L., & McAdams, H. H. (2004). Codon usage
510 between genomes is constrained by genome-wide mutational processes. *Proceedings of*
511 *the National Academy of Sciences of the United States of America, 101*(10), 3480-3485.

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

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

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

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

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

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

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

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

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

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

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

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

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

540 Fu, J., Murphy, K. A., Zhou, M., Li, Y. H., Lam, V. H., Tabuloc, C. A., Chiu, J. C., & Liu, Y.
541 (2016). Codon usage affects the structure and function of the Drosophila circadian clock
542 protein PERIOD. *Genes & Development*, *30*(15), 1761-1775.

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

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

547 Gilchrist, M. A., Chen, W. C., Shah, P., Landerer, C. L., &Zaretzki, R. (2015). Estimating
548 Gene Expression and Codon-Specific Translational Efficiencies, Mutation Biases, and
549 Selection Coefficients from Genomic Data Alone. *Genome biology and evolution*, *7*(6),
550 1559-1579.

551 Gingold, H., Tehler, D., Christoffersen, N. R., Nielsen, M. M., Asmar, F., Kooistra, S. M.,
552 Christophersen, N. S., Christensen, L. L., Borre, M., Sørensen, K. D., Andersen, L. D.,
553 Andersen, C. L., Hulleman, E., Wurdinger, T., Ralfkiær, E., Helin, K., Grønbæk, K.,
554 Ørntoft, T., Waszak, S. M., Dahan, O., Pedersen, J. S., Lund, A. H.,&Pilpel, Y. (2014). A
555 dual program for translation regulation in cellular proliferation and differentiation.*Cell*,
556 158(6), 1281-1292.

557 Graveley, B.R., Brooks, A.N., Carlson, J.W., Cherbas, L., Choi, J., Davis, C.A., Dobin, A.,
558 Duff, M., Eads, B., Hansen, K.D., Landolin, J., Langton, L., Malone, J., Miller, D.,
559 Roberts, J., Sandler, J., Sturgill, D., Tang, H., van Baren, M.J., Wan, K.H., Xiao, S.,
560 Yang, L., Zhang, D., Zhang, Y., Zou, Y., Andrews, J., Brenner, S.E., Brent, M., Cherbas,
561 P., Dudoit, S., Gingeras, T.R., Hoskins, R., Kaufman, T., Oliver, B, &Celniker, S.E.
562 (2010.03.15). The *D. melanogaster*transcriptome: modENCODE RNA-Seq data

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

566 Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr,
567 G., De La Vega, F. M., Briggs, J., Günther, S., Prescott, N. J., Onnie, C. M., Häsler, R.,
568 Sipos, B., Fölsch, U. R., Lengauer, T., Platzer, M., Mathew, C. G., Krawczak, M., &
569 Schreiber, S. (2007). A genome-wide association scan of nonsynonymous SNPs
570 identifies a susceptibility variant for Crohn disease in ATG16L1. *Nature genetics, 39*(2),
571 207.

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

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

577 Huang, X., Kurata, N., Wang, Z. X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H., Li, W.,
578 Guo, Y., Lu, Y., Zhou, C., Fan, D., Weng, Q., Zhu, C., Huang, T., Zhang, L., Wang, Y.,
579 Feng, L., Furuumi, H., Kubo, T., Miyabayashi, T., Yuan, X., Xu, Q., Dong, G., Zhan, Q.,
580 Li, C., Fujiyama, A., Toyoda, A., Lu, T., Feng, Q., Qian, Q., Li, J., & Han, B. (2012). A
581 map of rice genome variation reveals the origin of cultivated rice. *Nature, 490*(7421), 497.

582 Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer
583 RNAs and the occurrence of the respective codons in its protein genes: a proposal for a
584 synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of*
585 *molecular biology, 151*(3), 389-409.

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

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

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

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

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

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

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

602 Komar, A. A., Lesnik, T., & Reiss, C. (1999). Synonymous codon substitutions affect
603 ribosome traffic and protein folding during in vitro translation. *FEBS letters*, *462*(3), 387-
604 391.

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

608 Kraft, D. (1988).A software package for sequential quadratic programming. Technical
609 report DFVLR-FB. InstitutfürDynamik der Flugsysteme; Oberpfaffenhofen, pp. 88–28.

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

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

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

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

Matsumoto, T., John, A., Baeza-Centurion, P., Li, B., & 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., &Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila.*Nature, 351*(6328), 652-654.

Meurer A., Smith C. P., Paprocki M., Čertík O., Kirpichev S. B., Rocklin M., Kumar A., Ivanov S., Moore J. K., Singh S., Rathnayake T., Vig S., Granger B. E., Muller R. P., Bonazzi F., Gupta H., Vats S., Johansson F., Pedregosa F., Curry M. J., Terrel A. R., Roučka Š., Saboo A., Fernando I., Kulal S., Cimrman R., Scopatz A. (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., &Leder, P. (1964). RNA codewords and protein synthesis.*Science, 145*(3639), 1399-1407.

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

Pagani, F., Raponi, M., &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.

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

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

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

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

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

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

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

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

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

Savir, Y., &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., & Wright, F. (1988). Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomycespombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic acids research*, *16*(17), 8207-8211.

Sharp, P. M., Emery, L. R., &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., & Sharp, P. M. (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Research, 15*(19), 8023-8040.

Shin, Y. C., Bischof, G. F., Lauer, W. A., &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., Tinevez, J.-Y., White, D.J., Hartenstein, V., Eliceiri, K., &Tomancak, P. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods, 9*(7), 676.

682 Slimko, E. M., & Lester, H. A. (2003). Codon optimization of
683 CaenorhabditiselegansGluCl ion channel genes for mammalian cells dramatically
684 improves expression levels. *Journal of neuroscience methods*, *124*(1), 75-81.

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

689 Stenico, M., Lloyd, A. T., & Sharp, P. M. (1994). Codon usage in *Caenorhabditiselegans*:
690 delineation of translational selection and mutational biases. *Nucleic Acids*
691 *Research*, *22*(13), 2437-2446.

692 Takashima, S., Paul, M., Aghajanian, P., Younossi-Hartenstein, A., & Hartenstein, V.
693 (2013). Migration of Drosophila intestinal stem cells across organ boundaries.
694 *Development, 140*(9), 1903-1911.

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

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

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

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., & 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) & 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., &Ibba, M. (2010).New roles for codon usage. *Science*, *329*(5998), 1473-1474.

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

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

Xu, Z., & 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., & 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.

723 Yang, X., Ma, X., Luo, X., Ling, H., Zhang, X., &Cai, X. (2015). Codon usage bias and
724 determining forces in Taeniasolium genome. *The Korean journal of parasitology, 53*(6),
725 689.

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

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

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

734 Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K.,
735 Cummins, C., Gall, A., Girón, C.G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier,
736 T., Izuogu, O.G., Janacek, S.H., Juettemann, T., To, J.K., Laird, M.R., Lavidas, I., Liu, Z.,
737 Loveland, J.E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D.N., Newman,
738 V., Nuhn, M., Ogeh, D., Ong, C.K., Parker, A., Patricio, M., S., H., Schuilenburg, H.,
739 Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A.,
740 Frankish, A., Hunt, S.E., Kostadima, M., Langridge, N., Martin, F.J., Muffato, M., Perry,
741 E., Ruffier, M., Staines, D.M., Trevanion, S.J., Aken, B.L., Cunningham, F., Yates, A.,
742 &Flicek, P. (2017). Ensembl 2018. *Nucleic acids research, 46*(D1), D754-D761.

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

Zhao, F., Yu, C. H., & 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., & Liu, Y. (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature*, *495*(7439), 111-115.

752 **FIGURE LEGENDS**

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

761

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

771

772 **Figure 3. Gene-specific codon usage bias affects spatial protein expression in *D.***
773 ***melanogaster*.** (A) Representative northern blots and (B-C) summary data of the
774 relative abundance of $tRNA^{Lys}_{CTT}$ (common) and $tRNA^{Lys}_{TTT}$ (rare) codons across male

775 tissues. *, $p<0.05$ (n = 4, ANOVA followed by SNK *post hoc* tests). Error bars denote
776 standard deviation. †: Reproductive system is excluded. (D-E) Representative images
777 showing EGFP and RFP expression in HPZ (yellow lines marks tissue boundaries). (F-G)
778 Representative images showing EGFP and RFP expression in the AGSC (yellow lines
779 marks secondary cells). (H-I) Summary data of normalized EGFP signals in HPZ and
780 AGSC. *, $p<0.01$ (n=5, unpaired Student's *t*-test). Error bars denote standard deviation.
781 (J) Real-time qRT-PCR mRNA expression data of $EGFP^{Common}$ and $EGFP^{RareKEQ}$
782 (unpaired Student's *t*-test, n = 4, NS). Error bars denote SEM.

783 **TABLES**

	Cluster A		Cluster B	
Tissue	Fold enrichment	p-value	Fold enrichment	p-value
Adult Central Nervous System			1.13	0.040
Midgut			1.19	0.003
Testis			0.56	<0.001
Male Accessory Gland	4.45	<0.001	0.36	<0.001
Carcass			0.34	<0.001
Salivary Gland			0.49	0.024
Eye			0.80	0.050

784

785 **Table 1. Tissue-specific expression patterns of genes in Clusters A and B.** Genes
786 included in Clusters A and B are as in Fig. 2. Only tissues with a significant over- or
787 underrepresentation in at least one of the two clusters are shown. Fold enrichment is
788 relative to the entire genome.

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

789

790 **Table 2. Sex-biased expression patterns of genes in Clusters A and B.** Genes

791 included are as in Table 1. Fold enrichment is relative to the entire genome.

792

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

804 **METHODS**

805 **Genomic and transcriptomic data**. Protein-coding DNA sequences were from
806 Ensembl (Zerbino et al., 2017) (https://www.ensembl.org/). Reference coding sequences
807 included in the analysis of gene-specific codon usage bias were chosen according to the
808 following criteria: 1) The sequence length is a multiple of three. 2) The sequence uses
809 standard genetic code. 3) For each gene, only the longest mRNA isoform was used for
810 analysis. If there were multiple isoforms of the same length, then the first record shown
811 in the fasta file was used. 4) The encoded protein includes all 19 amino acids that have
812 degenerate codons; for the amino acid serine, the two-fold and four-fold degenerate
813 codon groups were treated as if they encoded two different amino acids. Transcriptomic
814 data were from the FlyAtlas microarray database (Chintapalli, Wang, and Dow, 2007)
815 and the modENCODE RNA-seq database (Graveley et al., 2010; Gelbart and Emmert,
816 2013).

817

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

826 Category one – Codons for Arginine:

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

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:

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

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

843
$$\begin{cases} Ex_{AAA} = y_{Lys} \times \frac{x_{AAA}}{x_{AAA}+x_{AAG}} = y_{Lys} \times \frac{x_{AAA}}{x_{AAA}+x_{AAA}\times\frac{\mu_{A\to G}}{\mu_{G\to A}}} = y_{Lys} \times \frac{1}{1+\frac{\mu_{A\to G}}{\mu_{G\to A}}} \\ Ex_{AAG} = y_{Lys} \times \frac{x_{AAG}}{x_{AAA}+x_{AAG}} = y_{Lys} \times \frac{x_{AAA}\times\frac{\mu_{A\to G}}{\mu_{G\to A}}}{x_{AAA}+x_{AAA}\times\frac{\mu_{A\to G}}{\mu_{G\to A}}} = y_{Lys} \times \frac{\frac{\mu_{A\to G}}{\mu_{G\to A}}}{1+\frac{\mu_{A\to G}}{\mu_{G\to A}}} \end{cases}, \quad (10)$$

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

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

849
$$\Theta = (\mu_{A\to C}, \mu_{C\to A}, \mu_{A\to G}, \mu_{G\to A}, \mu_{A\to T}, \mu_{T\to A}, \mu_{C\to G}, \mu_{G\to C}, \mu_{C\to T}, \mu_{T\to C}, \mu_{G\to T}, \mu_{T\to G}). \quad (11)$$

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

853 Since equations generated from Equation (3) form a system of homogeneous linear
854 equations, it is meaningless to infer exact μ values from the minimization of $\chi^2(\Theta)$; rather,
855 only the relative magnitudes of different μ values are important for calculating the
856 expected codon counts later. As we used the assumption that the lowest μ value is at

857 least 1/100 of the highest, we set the range of μ values between 0.001 and 0.1 during
858 the minimization of $\chi^2(\Theta)$. For each gene, the minimum $\chi^2(\Theta)$ is used to calculate p-value.

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

861

862 **Codon usage heatmaps**

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

864 $$f_{gd} = \frac{n_{gd}}{y_{ga}}, \quad (12)$$

865 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
866 synonymous codons in *g*. It should be noted that codons for Ser are not treated as two
867 codon groups in codon usage heatmaps. As the mechanism of recognizing stop codons
868 is fairly different from recognizing other codons (Brown et al., 2015), and methionine and
869 tryptophan respectively have only one codon, the analysis is restricted to the other 59
870 codons. Therefore, a 59-dimension vector B_g is used to describe the codon usage
871 pattern of *g*,

872 $$B_g = \left(f_{g1}, f_{g2}, f_{g3}, \cdots, f_{gd}, \cdots, f_{g59}\right)^T. \quad (13)$$

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

875 $$H_d = \left(f_{1d}, f_{2d}, f_{3d}, \cdots, f_{gd}, \cdots, f_{Md}\right). \quad (14)$$

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

878 The "gplots" (Warnes et al., 2015) package in R was used to generate heatmaps. The
879 method of hierarchical clustering was complete linkage with Euclidean distance
880 measuring the dissimilarities between codons and Spearman's correlation coefficient
881 measuring the similarities between genes.

882

883 **Identifying genes with tissue-specific expression patterns**

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

894

895 **Identifying genes with sex-biased expression patterns**

896 Adult fly expression data from the modENCODE RNA-seq database were used to
897 identify genes with either male- or female-biased expression patterns. Only genes that

898 showed at least 10-fold higher expression in one sex relative to the other were defined
899 as sex-biased.

900

901 **Hypergeometric tests for gene set enrichment**

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

904

905 **Gene ontology analysis**

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

910

911 **Analyses of tRNA gene expression analyses**

912 Northern blots were used to measure relative tRNA abundance in different body parts.
913 Total RNA was extracted from four pools of 10 dissected male tissues (head,
914 reproductive system, and remaining thorax and abdominal parts) and 10 whole male
915 flies with the TRIzol reagent (Invitrogen Catalog # 15596-026). Probe sequences were:
916 $tRNA^{Lys}_{CTT}$, 5'-AACGTGGGGCTCGAACCCACGACCCTGA-3'; $tRNA^{Lys}_{TTT}$, 5'-
917 GAACAGGGACTTGAACCCTGGACCCTTG-3'. Probes were labeled by ^{32}P using T4
918 PNK (NEB Catalog # M0201S). Signals were measured and normalized to total tRNA

919 signals using the BIO-RAD Quantity One 1-D analysis software. ANOVA followed by
920 SNK *post hoc* test was used to compare tRNA levels between samples.

921

922 **Animals**

923 Fruit flies (*D. melanogaster*) were raised on corn syrup-soy food (Archon Scientific) at
924 25°C and 60% relative humidity with a 12-hour light/dark cycle. Synthetic cDNAs
925 encoding EGFP and mCherry fluorescent proteins were from IDT Inc. Full sequences of
926 *mCherry*^{Common}, *EGFP*^{Common}, and *EGFP*^{RareKEQ} alleles are in the Supplemental
927 Information (Sequences of Fluorescent Reporter Genes). To generate animals that
928 express each transgenic allele under UAS control, synthetic cDNAs were cloned into the
929 EcoRI/NotI cloning sites of the pUASTattB plasmid (Bischof et al., 2007). Since the
930 *EGFP*^{RareKEQ} allele contains one EcoRI site in the middle, digestion time was shortened
931 to less than 20 minutes to allow incomplete digestion. The UAS-*RFP*^{Common} transgene
932 was inserted into a chromosome II landing site (*y*^{1} *M{vas-int.Dm}ZH-2A w*; M{3xP3-*
933 *RFP.attP}ZH-51D*; Bloomington #24483), and the UAS-*EGFP*^{Common} and UAS-
934 *EGFP*^{RareKEQ} transgenes were inserted into the same chromosome III landing site
935 (*y*^{1}*M{vas-int.Dm}ZH-2A w*; M{3xP3-RFP.attP}ZH-86Fb*; Bloomington #24749). Double
936 homozygotic lines UAS-*RFP*^{Common}; UAS-*EGFP*^{Common} and UAS-*RFP*^{Common}; UAS-
937 *EGFP*^{RareKEQ} were generated and then crossed to *Act*5*C*-GAL4/*SM*6 (Duncan lab,
938 Washington University).

939

940 **Real-time qRT-PCR**

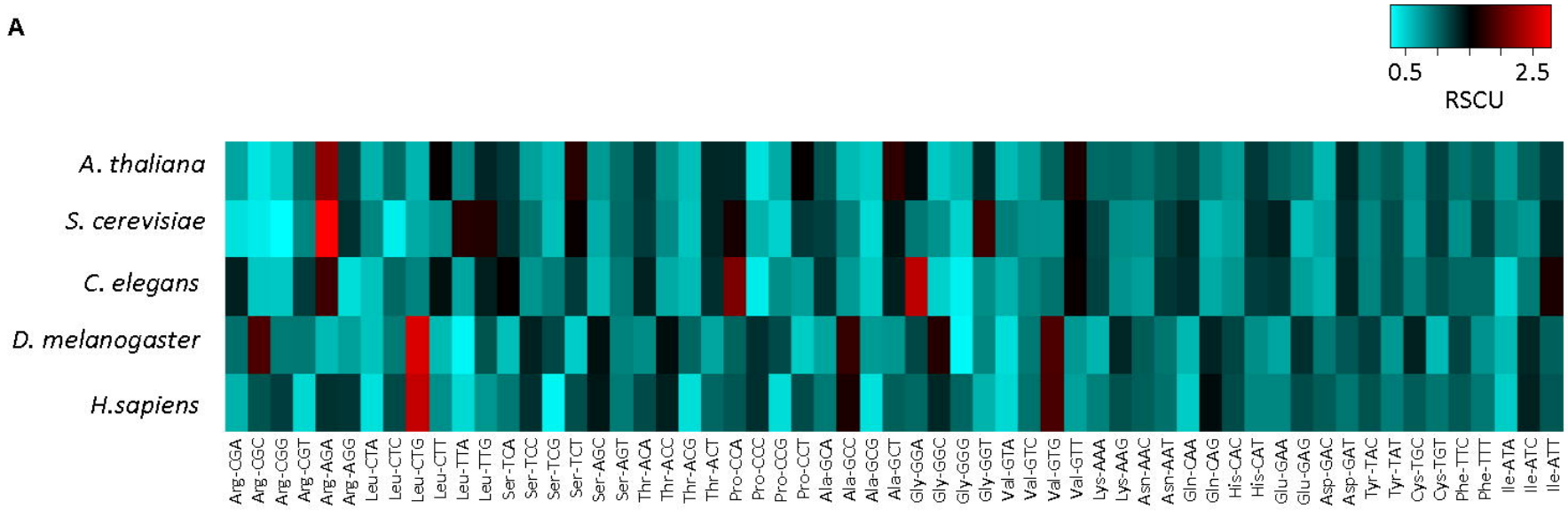
941 The mRNA expression levels of reporter genes in whole four-day-old male flies were
942 quantified by using real-time qRT-PCR, following previously published methods (Lu et al.,
943 2012; Hill et al., 2017). For *RpL*32, the forward primer was 5'-
944 CACCAAGCACTTCATCCG-3', and the reverse primer was 5'-
945 TCGATCCGTAACCGATGT-3'. For $EGFP^{Common}$ and $EGFP^{RareKEQ}$, the forward primers
946 were respectively 5'-AACTTCAAGATCCGCCACAAC-3' and 5'-
947 AACTTCAAAATCCGCCACAAC-3', while these alleles shared the same reverse primer
948 5'-GTGCTCAGGTAGTGGTTATCG-3'.

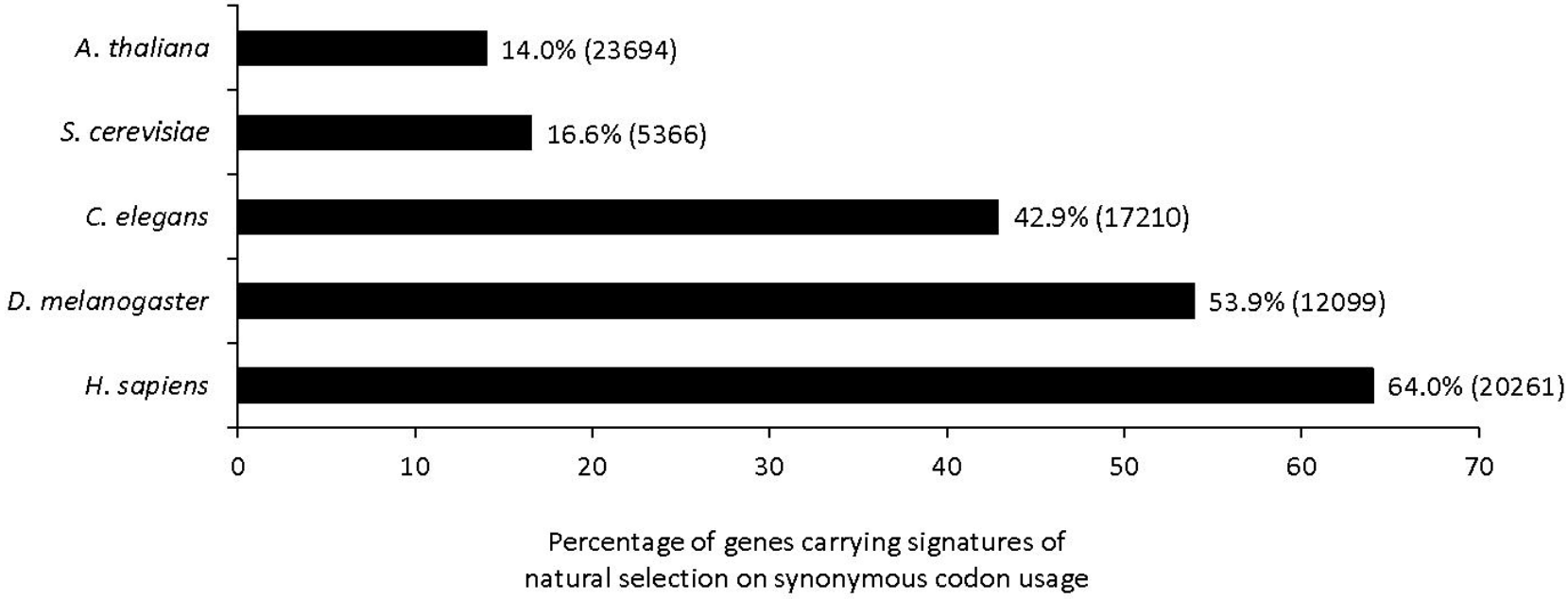
949

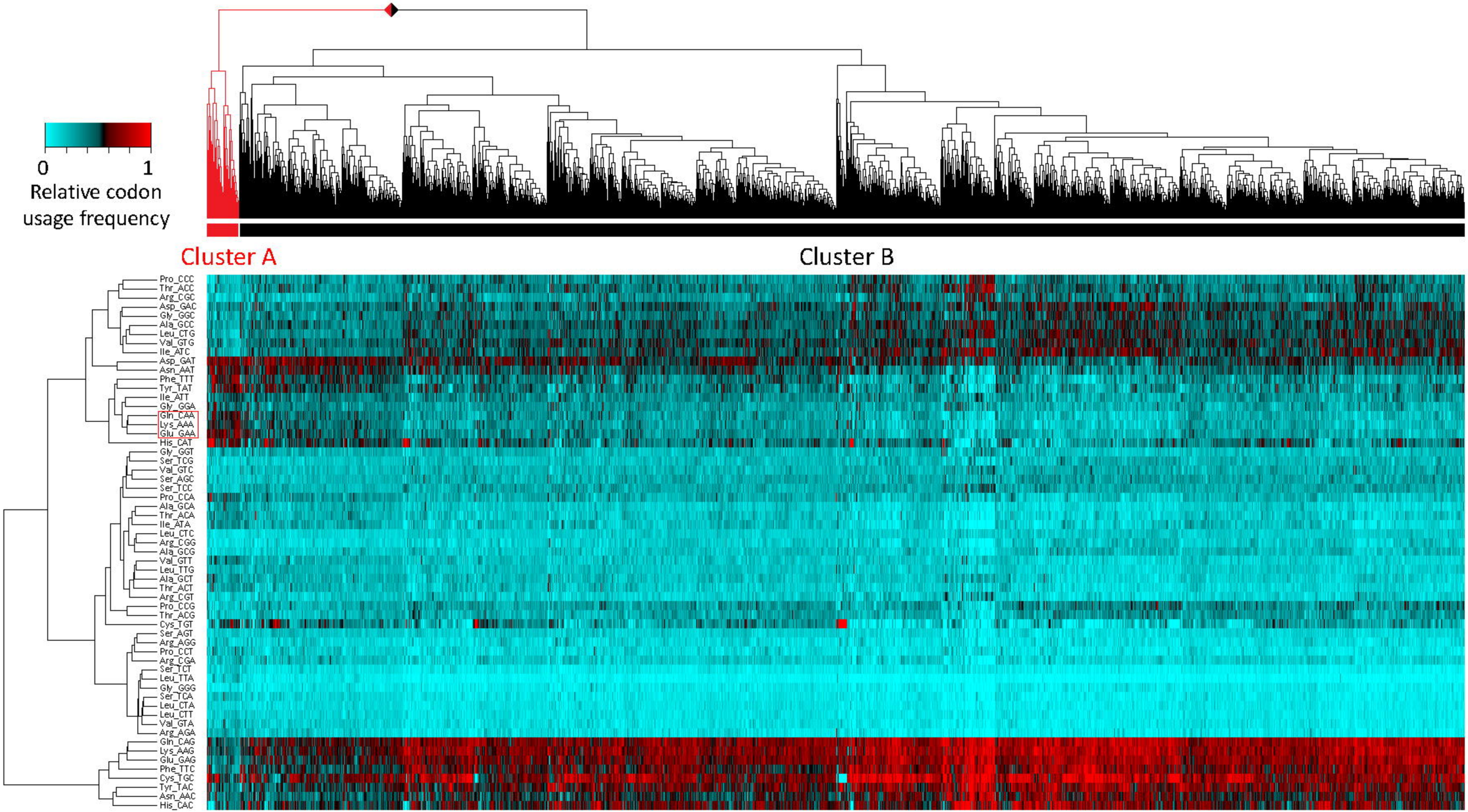
950 **Quantitative reporter gene imaging**

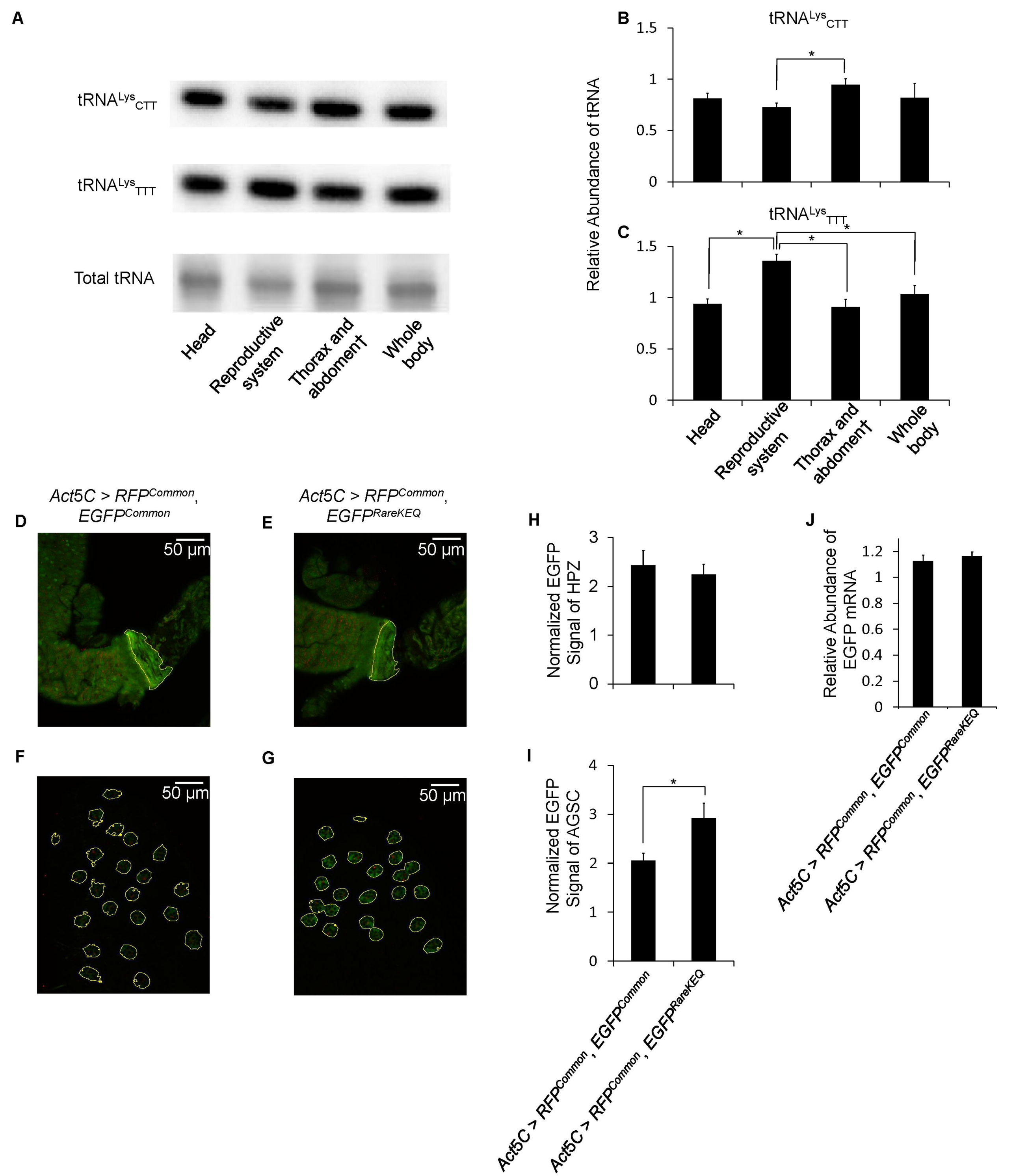
951 Male reproductive and gut tissues from four-day-old adult male flies that express either
952 the $EGFP^{Common}$ or $EGFP^{RareKEQ}$ were dissected in chilled PBS and mounted for imaging
953 on a Nikon A1Si laser scanning confocal microscope with a 20X oil objective (n=5
954 samples per genotype).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

962


A
0.5
2.5
RSCU
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


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
natural selection on synonymous codon usage


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
*
D
Act5C > RFP^{Common}, EGFP^{Common}
E
Act5C > RFP^{Common}, EGFP^{RareKEQ}
50 μm
F
G
H
Normalized EGFP Signal of HPZ
3
2
1
0
I
Normalized EGFP Signal of AGSC
4
J
Relative Abundance of EGFP mRNA
1.2
0.8
0.6
0.4
0.2
Act5C > RFP^{Common}, EGFP^{Common}
Act5C > RFP^{Common}, EGFP^{RareKEQ}