Expression-Linked Patterns of Codon Usage, Amino Acid Frequency, and Protein Length in the Basally Branching Arthropod *Parasteatoda tepidariorum*

Carrie A. Whittle[1] and Cassandra G. Extavour[1,2,*]

[1]Department of Organismic and Evolutionary Biology, Harvard University,

[2]Department of Molecular and Cellular Biology, Harvard University,

*Corresponding author: E-mail: extavour@oeb.harvard.edu.

Accepted: March 22, 2016

Data deposition: The RNA-seq data used in this study are available at the SRA database under project number PRJNA253975.

Abstract

Spiders belong to the Chelicerata, the most basally branching arthropod subphylum. The common house spider, *Parasteatoda tepidariorum*, is an emerging model and provides a valuable system to address key questions in molecular evolution in an arthropod system that is distinct from traditionally studied insects. Here, we provide evidence suggesting that codon usage, amino acid frequency, and protein lengths are each influenced by expression-mediated selection in *P. tepidariorum*. First, highly expressed genes exhibited preferential usage of T3 codons in this spider, suggestive of selection. Second, genes with elevated transcription favored amino acids with low or intermediate size/complexity (S/C) scores (glycine and alanine) and disfavored those with large S/C scores (such as cysteine), consistent with the minimization of biosynthesis costs of abundant proteins. Third, we observed a negative correlation between expression level and coding sequence length. Together, we conclude that protein-coding genes exhibit signals of expression-related selection in this emerging, noninsect, arthropod model.

Key words: spider, arachnid, Chelicerata, optimal codons, amino acids.

Introduction

Arthropods are the largest animal phylum, estimated to contain at least 80% of all animal species (Akam 2000; Odegaard 2000; Regier et al. 2010). Genome-wide molecular evolutionary research in this vast taxonomic group has largely focused on holometabolous insect models of the genera Drosophila, Anopheles, Tribolium, Nasonia, and Apis, or the branchiopod crustacean Daphnia (Wiegmann and Yeates 2005; Weinstock et al. 2006; Stark et al. 2007; Richards et al. 2008; Group et al. 2010; Colbourne et al. 2011; Neafsey et al. 2015). Recent data from emerging model species, including hemimetabolous insects (the cricket *Gryllus bimaculatus* and the milkweed bug *Oncopeltus fasciatus*) and an amphipod crustacean (*Parhyale hawaiensis*), suggest that nontraditional arthropod models can provide valuable insights into the factors shaping molecular evolution of protein-coding DNA (Whittle and Extavour 2015). Expanding this research to include arthropods that belong to the most basally branching arthropod clade, the Chelicerata, is essential to furthering our understanding of genome evolution in the most speciose group of animals, and thus of animal genome evolution as a whole (Sanggaard et al. 2014; Zuk et al. 2014). An emerging model for comparative development and body plan evolution is the common house spider, *Parasteatoda tepidariorum* (previously *Achaearanea tepidariorum*) (Hilbrant et al. 2012). This taxon offers promising opportunities to address key issues in evolutionary genomics.

The spiders belong to the Chelicerata, the most basally branching subphylum of the arthropods (Regier et al. 2010; Hilbrant et al. 2012). The spider *P. tepidariorum* has historically served as a laboratory model for genetics and development, due to its rapid life cycle (<2 months), ease of culture, high fecundity, and small size (McGregor et al. 2008; Hilbrant et al. 2012). Stages of embryogenesis and postembryonic development have also been well characterized (Akiyama-Oda and Oda 2003; Mittmann and Wolff 2012). This arachnid has extensive functional genetic tools available (McGregor et al.

2008; Hilbrant et al. 2012) and has been employed to reveal properties of specific genes, including those involved in embryonic body patterning and germ line differentiation (Oda et al. 2007; Hilbrant et al. 2012; Schwager et al. 2014). It has also played a significant role in the studies of developmental evolution, as *P. tepidariorum* provides an outgroup to polarize studies of derived and ancestral body plan traits in the other branches of arthropods (Hilbrant et al. 2012). Recently, large-scale transcriptome data sets from the embryonic tissues of this spider have become available (Posnien et al. 2014), as well as a draft genome sequence (Stephen Richards and Alistair McGregor, personal communication; see Acknowledgements), expanding the utility of this system to include genome-wide molecular evolutionary research. In particular, the large-scale genomic and transcriptomic data sets now available make *P. tepidariorum* highly attractive as a model to investigate the evolution of protein-coding genes in a basally branching arthropod.

In protein-coding genes, the use of synonymous codons for amino acids appears to be nonrandom. Biases in codon usage may result from selection for biochemically efficient and accurate translation (Duret and Mouchiroud 1999; Duret 2000; Stoletzki and Eyre-Walker 2007), or from mutation (Osawa et al. 1988; Sueoka 1988; Sharp et al. 1995). The hypothesis of translational selection has been supported by findings that transfer-RNA (tRNA) abundance or tRNA gene copy number corresponds to the most common codons in the genome (Ikemura 1981, 1985; Duret 2000). Furthermore, highly transcribed genes preferentially use a subset of favored codons (optimal codons) in various eukaryotic taxa, suggesting that translational selection operates in a number of fungi, plants, and animals (Sharp et al. 1986; Stenico et al. 1994; Duret and Mouchiroud 1999; Cutter et al. 2006; Ingvarsson, 2008; Whittle et al. 2011). In the arthropods, optimal codon usage in highly expressed genes has been reported in some holometabolous insects, which are the relatively derived monophyletic group of "higher" insects that undergo true metamorphosis, such as Drosophila, Aedes, Anopheles, Nasonia, and Tribolium species (Duret and Mouchiroud 1999; Behura and Severson 2011, 2013; Williford and Demuth 2012). However, optimal codon use appears weak or absent in other Holometabola, such as Bombyx (Jia et al. 2015; Whittle et al., unpublished data). Although data from arthropods outside Holometabola are relatively uncommon, recent findings of a connection between expression levels and codon usage from two basally branching hemimetabolous insects, a cricket (*G. bimaculatus*) and a milkweed bug (*O. fasciatus*), and from the amphipod crustacean *Parh. hawaiensis* are suggestive of translational selection in those systems (Whittle and Extavour 2015). Expanding this research to include the spider *P. tepidariorum* will allow determination of whether optimal codons exist in the most basally branching arthropods, and if so, whether and how their use is connected to other genome evolution characteristics such as amino acid composition (Akashi 2003; Cutter et al. 2006; Williford and Demuth 2012).

Research on amino acid composition relative to expression level is sparse as compared with research on codon usage. Gaining insight into amino acid preferences could advance our understanding of the dynamics underlying protein evolution, and thus warrants greater attention. Although some amino acid preferences might individually have small effects on protein function, cell biology, and/or fitness, collectively such effects may govern proteome-wide patterns of amino acid composition and evolution (Akashi 2003). The few studies available to date suggest that indeed amino acid usage is under selective pressures, as it correlates with expression level in some eukaryotes (Duret 2000; Akashi 2003; Cutter et al. 2006; Williford and Demuth, 2012). Furthermore, amino acid frequency in proteins encoded by highly expressed genes appears to correspond to the most abundant tRNA in some organisms (e.g., yeast, *Caenorhabditis elegans*), consistent with selection for accurate and rapid translation (Duret 2000; Akashi 2003). In certain eukaryotes such as humans, *Sacchraomyces cerevisiae*, *Tribolium castaneum*, and *C. elegans*, amino acids encoded in highly expressed (and presumably highly translated) genes tend to be less metabolically costly (Cutter et al. 2006; Raiford et al. 2008; Williford and Demuth 2012) and/or have low size/complexity (S/C) (Dufton 1997), likely reducing the biochemical energy costs of protein synthesis and of protein stability (Dufton 1997; Cutter et al. 2006; Williford and Demuth 2012). Recently, it was found that for some arthropods, such as *G. bimaculatus* and *O. fasciatus*, elevated transcription is linked to greater usage of intermediate-sized amino acids (as compared with high cost), rather than the smallest possible amino acids. This may reflect a balance between minimizing average amino acid cost per protein while retaining those amino acids needed for stability, conformation, and/or function (Whittle and Extavour 2015). Examination of these molecular evolutionary features in *P. tepidariorum* can help reveal whether expression-mediated selection (Urrutia and Hurst 2003) on codon and amino acid usage is operative in this basally branching arthropod model, and could plausibly have been present in a last common arthropod ancestor.

In the present investigation, we study optimal codon and amino acid usage in *P. tepidariorum* using recently available large-scale genomic and RNA-seq data (supplementary table S1, Supplementary Material online). From an assessment of codon usage relative to expression levels, we provide evidence of optimization of codon usage across the majority of amino acids in this taxon. In addition, from an evaluation of amino acid frequency, we reveal that those amino acids with high S/C scores exhibit reduced usage in highly transcribed genes in favor of those with medium or low values, likely reflecting selection for reduced protein biosynthetic costs. Furthermore, we show that highly expressed genes are shorter than those expressed at lower levels, consistent with minimization of translational (and transcriptional) costs, and tend to be involved in protein synthesis and cell cycling. Collectively, these data suggest that expression-related selection acts on codon usage, amino acid frequency, and protein lengths in this spider.

Materials and Methods

For our sequence analyses of *P. tepidariorum*, we recently studied assembled coding sequence(s) (CDS) data derived from DNA-seq (whole genome sequencing: https://i5k.nal.usda.gov/, accessed July 2015; see Acknowledgements). In order to measure expression levels of these CDSs, we used large-scale embryo RNA-seq data sets containing 332,245,126 reads from multistage embryos representing all stages of embryogenesis (stages 1–14 as per Mittmann and Wolff 2012) (supplementary table S1, Supplementary Material online; see Posnien et al. 2014). Embryos have proven effective for studying the relationship between expression levels and molecular evolution, as they are multitissue structures that collectively express a large component of the genome (Subramanian and Kumar 2004; Whittle and Extavour 2015). For our assessment, we identified all CDSs with an ATG start and stop codon with no unknown or ambiguous nucleotides, for a total of 23,746 sequences (of a predicted 27,990 genes; https://i5k.nal.usda.gov/; accessed July 2015). We mapped reads to CDSs, and measured expression as fragments per kilobase million (FPKM) using Geneious 8.1 (http://www.geneious.com; last accessed March 17, 2016) run on a computing cluster (Harvard Odyssey).

Measuring Molecular Evolutionary Parameters

For identification of optimal codons for each of the 18 amino acids with synonymous codons, we compared codon usage for the 5% of genes with the highest and lowest expression levels. The approach of comparison of codon usage profiles among the most extremely highly and lowly expressed CDSs has proven to be an effective means to reveal preferences in codon use that are linked to transcription (Shields et al. 1988; Duret and Mouchiroud 1999; Cutter et al. 2006; Cutter 2008; Ingvarsson 2008; Wang et al. 2011; Whittle et al. 2011; Whittle and Extavour 2015). For each CDS per expression class, we determined the relative synonymous codon usage (RSCU), which represents the observed frequency of a specific codon relative to the expected frequency if all synonymous codons were used equally. RSCU values larger than 1 denote preferential usage of a codon, and greater values within a synonymous codon family indicate increased usage (Sharp et al. 1986). Optimal codons were defined as those with a statistically significant and positive $\Delta\text{RSCU} = \text{RSCU}_{\text{Mean highly expressed CDS}} - \text{RSCU}_{\text{Mean low expressed CDS}}$ (Duret and Mouchiroud 1999; Cutter et al. 2006; Ingvarsson 2008; Wang et al. 2011; Whittle et al. 2011; Whittle and Extavour 2015). RSCU was determined using CAIcal (Puigbo et al. 2008). As a supplementary assessment, correspondence analysis of codon usage was conducted using CodonW, which predicts putative optimal codons based on distribution of codons along the principle axis, and must be confirmed with expression data (Peden 1999; http://codonw.sourceforge.net/; last accessed March 17, 2016). This was conducted to assess the agreement between the ΔRSCU and correspondence approaches.

After the identification of optimal codons, the frequency of optimal codons (Ikemura 1981) was measured for all genes under study using CodonW (Peden, 1999). We also measured ENC prime (ENC′; Novembre 2002), which is a predictor of bias in codon usage that accounts for nucleotide composition; values range between 20 and 61 with lower values indicating greater bias. ENC′ was determined in the program INCA (Supek and Vlahovicek 2005).

For our assessment of amino acid preferences, we evaluated the biochemical cost of proteins as described in Dufton (1997), where each amino acid is assigned a size complexity score (S/C) based on its molecular weight and complexity. The S/C scores range in value from a low of 1 for Gly up to 73 for Trp, and represent the biochemical costs of production and stability in protein conformation (Dufton 1997; Williford and Demuth 2012). The S/C score provided by Dufton (1997) for each amino acid is provided in supplementary table S2, Supplementary Material online. We determined the difference in the frequency among the 5% highest and lowest expressed genes for each of 20 amino acids, and as a complementary assessment measured the Spearman rank correlation between FPKM and amino acid frequency across all genes (for each amino acid). With respect to CDS length, analyses were conducted as detailed in the Results and Discussion sections.

Gene Ontology Annotation

For gene ontology (GO) annotation, we used data from the taxon *Drosophila melanogaster*, which has a well-annotated genome. In particular, we compared the spider CDS list with the *D. melanogaster* protein sequence database (v. 6, flybase.org; last accessed March 17, 2016) using BLASTX (http://blast.ncbi.nlm.nih.gov/; last accessed March 17, 2016) to identify likely orthologs. A similar approach was conducted using Swissprot as the reference database (Release 2015_09; http://www.uniprot.org; last accessed March 17, 2016). The ortholog was identified as the protein with the lowest e-value, with a cutoff of 10^{-6} for *D. melanogaster* and $e < 0.01$ in Swissprot. The GO tool DAVID (Huang da et al. 2009) was used for functional classification of genes under study.

Results and Discussion

Following mapping of the complete list of RNA-seq reads to the genomic CDSs, we found that a total of 19,936 genes in *P. tepidariorum* (see Materials and Methods) were expressed in

Table 1
The List of ΔRSCU Values for Spider among the 5% Highest and Lowest Expressed CDS in *Parasteatoda tepidariorum* Embryonic Tissues

Amino Acid	Codon	ΔRSCU	*P* value
Ala	**GCT**	**+0.360**	**
Ala	GCC	−0.028	
Ala	GCA	−0.190	**
Ala	GCG	−0.190	**
Arg	**CGT**	**+0.233**	**
Arg	CGC	−0.079	*
Arg	CGA	−0.100	*
Arg	CGG	−0.074	*
Arg	AGA	−0.109	*
Arg	AGG	−0.020	
Asn	**AAT**	**+0.079**	**
Asn	AAC	−0.121	**
Asp	**GAT**	**+0.105**	**
Asp	GAC	−0.163	**
Cys	TGT	−0.078	*
Cys	TGC	−0.269	**
Gln	CAA	−0.088	**
Gln	CAG	+0.002	
Glu	GAA	+0.007	
Glu	GAG	−0.045	**
Gly	**GGT**	**+0.179**	**
Gly	GGC	−0.058	*
Gly	GGA	−0.095	*
Gly	GGG	−0.127	**
His	CAT	−0.031	
His	CAC	−0.156	**
Ile	**ATT**	**+0.164**	**
Ile	ATC	−0.028	
Ile	ATA	−0.181	**
Leu	**TTA**	**+0.086**[a]	*
Leu	TTG	+0.089	*
Leu	CTT	+0.021	
Leu	CTC	−0.090	**
Leu	CTA	−0.057	*
Leu	CTG	−0.062	*
Lys	AAA	−0.078	**
Lys	**AAG**	**+0.064**	**
Phe	**TTT**	**+0.034**	*
Phe	TTC	−0.106	**
Pro	**CCT**	**+0.356**	**
Pro	CCC	−0.057	*
Pro	CCA	−0.270	**
Pro	CCG	−0.179	**
Ser	**TCT**	**+0.343**	**
Ser	TCC	−0.106	**
Ser	TCA	−0.181	**
Ser	TCG	−0.233	**
Ser	AGT	+0.182	**
Ser	AGC	−0.065	*
Thr	**ACT**	**+0.290**	**
Thr	ACC	−0.006	
Thr	ACA	−0.147	**
Thr	ACG	−0.224	**

(continued)

Table 1 Continued

Amino Acid	Codon	ΔRSCU	*P* value
Tyr	**TAT**	**+0.053**	*
Tyr	TAC	−0.168	**
Val	**GTT**	**+0.234**	**
Val	GTC	−0.120	**
Val	GTA	−0.044	
Val	GTG	−0.106	**

Note.—Putative optimal codons are underlined and in bold.
[a]*P* value of TTA was lower than TTG.
*$P \leq 0.05$ and >0.001; **$P \leq 0.001$.

embryos (FPKM > 0), with an average expression FPKM value of 100.4 ± 4.8. This gene set was used for all our subsequent analyses.

Optimal Codon Usage

The ΔRSCU values for each of 18 amino acids with synonymous codons are shown in table 1. We found that 14 of the 18 amino acids with synonymous codons had an optimal codon with a statistically significant and positive ΔRSCU in *P. tepidariorum*. Twelve of the putative optimal codons ended in T, while two, namely Lys (AAG) and Leu (TTA; note that TTG also showed signs of being an optimal codon), ended in G and A, respectively. The strong link between high expression and the frequency of T3 codons suggests that selection contributes to codon usage in this spider species. Correspondence analysis of codons across all CDSs using CodonW predicted the 13 of 14 optimal codons in table 1, with the only exception being Lys (changed to AAA). The remaining four amino acids with no optimal codons in table 1 had putative optimal codons identified in correspondence analysis (Cys (TGT), Gln (CAA), Glu (GAA), His (CAT)), but were not included further because they were not confirmed using our expression data, as required for defining them as optimal codons (Peden 1999).

Optimal Codon Usage Varies with Degeneracy

The magnitude of bias in usage of the optimal codon varied among amino acids. Four-fold and 3-fold amino acids had very large positive ΔRSCU values. Specifically, we found values of 0.360, 0.179, 0.164, 0.356, 0.290, and 0.234 for Ala, (GCT), Gly (GGT), Ile (ATT), Pro (CCT), Thr (ACT), and Val (GTT), respectively (table 1). Variation in ΔRSCU among the 4-fold amino acids, wherein all four nucleotides can exist in the third position, is consistent with selection, rather than solely mutational pressures shaping codon usage. In addition, two of the three amino acids with six synonymous codons also had very high ΔRSCU values of +0.233 for Arg (CGT) and +0.343 for Ser (TCT); the only exception with a relatively moderate value was Leu (TTA) with +0.086 (or TTG with +0.089). In contrast, the four amino acids without putative optimal codons were those with only two alternate codons, namely Cys, Gln, Glu, and His. The other five 2-fold degenerate amino acids with optimal codons exhibited relatively low ΔRSCU, with values between 0.034 and 0.105 for Asn (AAT), Asp (GAT), Lys (AAG), Phe (TTT), and Tyr (TAT). Together, this suggests that selection pressures may be reduced for those amino acids with lower degeneracy. This may indicate that the costs of inserting a nonoptimal codon are higher when the amino acid has a greater number of synonymous codons and alternate tRNAs. These findings in the spider *P. tepidariorum* concur with recent results found in the hemimetabolous insects *G. bimaculatus* and *O. fasciatus*, and the amphipod crustacean *Parh. hawaiensis* (Whittle and Extavour 2015), which also showed variation in ΔRSCU of the optimal codon among amino acids. Similarly, in Caenorhabditis, optimal codons of all three 6-fold degenerate amino acids were found to have higher ΔRSCU including Arg, Leu, and Ser (Cutter et al. 2006) and differences in selection on optimal codons among amino acids has been suggested for Drosophila (Moriyama and Powell 1998).

Collectively, these patterns are consistent with a hierarchy of selection coefficients that has been proposed to exist among codons and amino acids (McVean and Vieira 1999; Cutter et al. 2006; Whittle and Extavour 2015). It can be speculated that the higher ΔRSCU for more degenerate amino acids might result from the need for more than one tRNA for non–2-fold amino acids (as wobble rules allow a single tRNA for 2-fold degenerate amino acids; Ikemura 1985; Percudani 2001); this could, for instance, cause greater competition among tRNAs. Combining our present findings across amino acids in *P. tepidariorum* (table 1) with those from other arthopods (Moriyama and Powell 1998; Whittle and Extavour 2015) and nematodes (Cutter et al. 2006), it appears that a propensity for greater pressure for optimal codons use in more degenerate amino acids may be common in the Arthropoda, and potentially more broadly across animals.

Selection Contributes toward Optimal Codon Usage

We subdivided the CDSs into three classes: Those with the highest 5% (high), moderate 90% (moderate), and lowest 5% (low) expression levels, and observed clear and statistically significant differences in Fop among classes (fig. 1*A*). Binning was conducted to evaluate the link between expression and Fop, as such effects might be weaker in unbinned correlative data, where the relationship would be predominantly influenced by genes with lower expression levels, as these are more abundant. Nevertheless, Fop was positively correlated with FPKM level when assessed across all expressed genes ($R = 0.29$, $P < 10^{-15}$). The clear statistical demarcations in Fop values among expression level classes (fig. 1*A*) indicate that the method of contrasting the most and least highly expressed genes, a method with extensive precedent in other organisms, was advantageous for revealing optimal codon lists in this spider (Duret and Mouchiroud 1999; Cutter et al. 2006; Ingvarsson 2008; Wang et al. 2011; Whittle et al. 2011; Whittle and Extavour 2015). We found that ENC′ decreased from the low (50.8 ± 0.14), to the moderate (49.50 ± 0.05), to the high (47.8 ± 0.21) expression classes (Mann–Whitney *U* [MWU] test, $P < 0.001$). Thus, this secondary measure, which controls for nucleotide composition, concurs with elevated biased codon usage in the high expression class in this spider.

Although a connection between Fop and expression level as reported here for *P. tepidariorum* (table 1 and fig. 1*A*) is typically explained by translational selection, an alternative possibility is that high transcription rates might be linked to a mutational bias that would result in disproportionate use of the optimal codons in highly expressed genes (Beletskii and Bhagwat 1996; Comeron 2004). For instance, it is feasible that highly transcribed genes have a mutational bias that leads to elevated C to T mutations, as has been found in *Escherichia coli*, and/or a transcription-coupled repair mutational asymmetry (Beletskii and Bhagwat 1996; Green et al. 2003). Mutational biases can be assessed by comparing T content of noncoding DNAs such as introns, which are largely selectively neutral, with T content at third sites of codons per CDS (because nearly all optimal codons end in T, and Fop and T3 were highly correlated Spearman's ranked $R = 0.85$, $P < 2.0 \times 10^{-7}$; supplementary fig. S1, Supplementary Material online) (Comeron 2004; Bachtrog 2007; Qiu et al. 2011; Williford and Demuth 2012). Elevated T in the third position of codons as compared with T of introns would be suggestive of selection.

As a further assessment of whether selection was a factor contributing to the preferential use of optimal codons in highly expressed CDSs, we identified 17,352 gene sequences expressed in embryos that had one or more introns based on the studied assembly (note: In fig. 1*B*, introns were concatenated for analysis when a gene had more than one intron; see below for unconcatenated introns). We report that the introns for embryo genes in this spider had high AT content (mean = $0.667 \pm 4.0 \times 10^{-4}$), with approximately similar levels of A ($0.336 \pm 2.0 \times 10^{-4}$) and T ($0.337 \pm 3.3 \times 10^{-4}$; fig. 1*B*), suggesting a substantial mutational AT bias. The T3 in exons increased in proportion from the low, to moderate and high expression classes, and was higher than the proportion of T in introns (T_{intron}), particularly for highest expression class (MWU tests, $P < 0.001$; fig. 1*B*). T_{intron} showed very mild variation across expression levels, and no difference between moderate and high expressed genes (fig. 1*B*). In contrast, average A3 of exons was between 0.318 and 0.320 (standard errors $<1.8 \times 19^{-3}$), which is within a similar range (or mildly lower) as introns, implying that mutational pressure may largely influence A3 in exons. Thus, these results are consistent with expression-mediated elevation of T3 codon usage in this spider. Herein, T3 is a conservative measure of optimal codon usage and selection because 12 of 18 amino acids have T-ending codons (while the proportion of T3 in CDS is measured

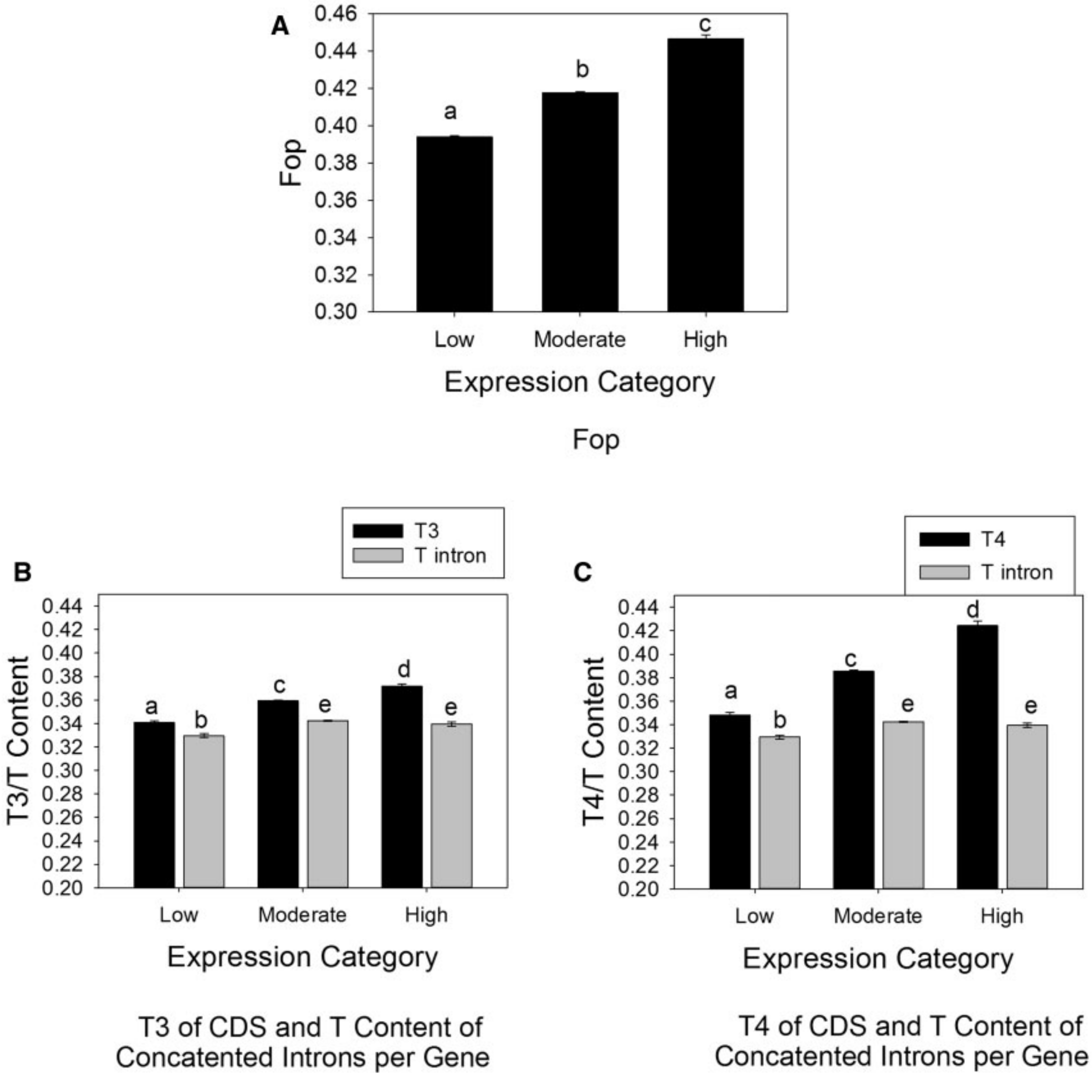

Fig. 1.—(*A*) The mean Fop value for CDS with low (below the 5th percentile), moderate (between 5th and 95th percentile), and high (>95th percentile) FPKM values for genes expressed in embryos. Different letters indicate statistically significant differences using ranked ANOVA ($P < 0.001$) and Dunn's paired contrasts ($P < 0.05$). (*B*) The mean proportion of *T* at third codon positions (T3) per CDS and *T* from associated introns for each expression category. (*C*) The mean proportion of *T* at third codon positions for 4-fold amino acids (T4) per CDS and *T* from associated introns for each expression category. For genes with more than one intron, the introns were concatenated. Different letters in B and C indicate a difference using MWU tests ($P < 0.001$ for all pairs).

across all amino acids, rather than only those utilizing optimal codons), and because third codon sites are not entirely synonymous, but rather include some nonsynonymous sites (potentially affecting the proportion of T3). We thus assessed the frequency of T in the third nucleotide site across the five 4-fold amino acids (Ala, Gly, Pro, Thr, and Val), described hereafter as T4. Each 4-fold amino acid has a T3 optimal codon defined herein (table 1), and the third nucleotide site is always synonymous, thus providing a robust measure of selection for T-ending codons. The results showed that T4 was markedly higher (>20%) than T_{intron} for the highest expression class, and to a lesser extent for the moderate and low classes (MWU tests, $P < 0.001$; fig. 1C), thus providing a stronger signal (than T3) of expression-mediated selection. Notably, analysis of solely those genes with introns (instead of all 19,936 genes under study) yielded nearly identical results for Fop, T3, and T4 (data not shown). In sum, it is evident that selection is a contributing factor involved in shaping codon usage under high expression in this taxon.

As it has been reported in some eukaryotes that short introns are nearly entirely selectively neutral (Farlow et al. 2012), while long introns can be subject to selection, as a complementary assessment we classified all introns of embryo expressed genes (unconcatenated, $N = 90{,}851$) into two classes, short and long, divided by the median across the expressed gene set (Haddrill et al. 2005). For the long introns, we also removed the 50 bp at each end of the intron, which are most prone to be involved in regulatory processes (Williford and Demuth 2012). We report that T_{intron} for short and long introns was lower than T4 in all expression classes, with the greatest differences observed in the highly expressed genes (supplementary fig. S2*A* and *B*, Supplementary Material online; MWU tests, $P < 0.001$ for all paired contrasts), further concurring with selection on codon usage in this spider.

Optimal Codon Usage in P. tepidariorum as Compared with Other Organisms

Our results showing that most putative optimal codons end in T in this spider differ from those based on studies of some other animal models studied to date, including Drosophila spp., *T. castaneum*, and *C. elegans*, and from the well-studied eukaryotic model yeast (*Schizosaccharomyces pombe*) (Duret and Mouchiroud 1999; Akashi 2001; Cutter et al. 2006; Williford and Demuth 2012), wherein optimal codons typically end in G or C. However, our findings are similar to results in chicken (*Gallus gallus*), where it was reported that each of the 11 optimal codons identified in that taxon ended in T (Rao et al. 2011). Furthermore, recent data from the hemimetabolous insects cricket (*G. bimaculatus*) and milkweed bug (*O. fasciatus*) suggest that putative optimal codons nearly always end in AT in those organisms (Whittle and Extavour 2015). In plants, species of Picea have been shown to have AT3 optimal codons (De La Torre et al. 2015), while species of Populus have GC3 optimal codons (Ingvarsson 2007, 2008), consistent with divergence in the types of optimal codons in that plant taxonomic group. Together, the growing data from nontraditional animal models, including spider, show that animals can exhibit marked differences in optimal codon usage, and that A- or T-ending codons may be favored across the majority of amino acids in some animal systems.

It has been suggested that six codons (UUC, UAC, AUC, AAC, GAC, and GGT) are universally optimal across the tree of life, based on analysis of bacteria, yeast, and *D. melanogaster* (Sharp and Devine, 1989). We found herein, however, that only one of these codons was optimal in the common house spider (GGT for glycine; table 1). This agrees with results from chicken, cricket, and milkweed bug, wherein GGT was the only one of six "universally optimal" codons found to be optimal in those taxa (Rao et al. 2011; Whittle and Extavour 2015). In this regard, it is evident that the universal codon concept does not appear to hold in all animals, including the basally branching arthropod *P. tepidariorum*.

Table 2

The Mean Frequency (percentage and standard error, SE) of Amino Acids in the 5% of Most Highly and Lowly Expressed Genes in *Parasteatoda tepidariorum* Embryos

Amino Acid	Mean Frequency (SE)		Percent Difference	*P* Value	S/C Score	Types of Codons
	High Expression	*Low Expression*				
Met	3.087 (0.056)	2.324 (0.037)	24.71	**	64.68	ATG
Gly	6.002 (0.123)	5.038 (0.102)	16.06	**	1	GGN
Lys	8.514 (0.130)	7.298 (0.081)	14.29	**	30.14	AAA/AAG
Ala	6.035 (0.092)	5.228 (0.068)	13.37	**	4.76	GCN
Glu	6.941 (0.116)	6.226 (0.070)	10.30	**	36.48	GAA/GAG
Val	6.371 (0.088)	5.857 (0.060)	8.057	**	12.28	GTN
Gln	3.752 (0.073)	3.540 (0.052)	5.648		37.48	CAA/CAG
Arg	5.336 (0.098)	5.107 (0.063)	4.275		56.34	CGN/AGA/AGG
Asp	5.278 (0.086)	5.120 (0.058)	2.983		32.72	GAT/GAC
Pro	4.557 (0.093)	4.443 (0.077)	2.496		31.8	CCN
Leu	8.840 (0.108)	9.440 (0.081)	−6.79	**	16.04	CTN/TTA/TTG
Tyr	3.293 (0.065)	3.575 (0.054)	−8.56	**	57	TAT/TAC
Phe	4.330 (0.082)	4.770 (0.058)	−10.1	**	44	TTT/TTC
Thr	5.077 (0.076)	5.666 (0.062)	−11.5	**	21.62	ACN
His	2.160 (0.056)	2.427 (0.052)	−12.3	**	58.7	CAT/CAC
Ser	7.044 (0.102)	8.089 (0.075)	−14.8	**	17.86	TCN/AGT/AGC
Ile	5.810 (0.081)	6.712 (0.065)	−15.5	**	16.04	ATT/ATC/ATA
Asn	4.622 (0.073)	5.474 (0.065)	−18.4	**	33.72	AAT/AAC
Trp	1.019 (0.044)	1.212 (0.027)	−18.9	**	73	TGG
Cys	1.915 (0.071)	2.443 (0.059)	−27.5	**	57.16	TGT/TGC

Note.—Amino acids are listed from the largest positive percent difference to largest negative percent difference among high and low expressed genes.

**$P < 0.001$ using MWU-tests. The types of codons used for each amino acid are shown.

Amino Acid Usage

We evaluated the biochemical cost of proteins as described by Dufton (1997), wherein amino acids are assigned a size complexity score (S/C) based on their molecular weight and complexity. Using an approach parallel to that used to identify optimal codons, we identified amino acids preferred under high expression by comparing their frequency in the 5% highest and lowest expressed CDSs in embryos. The results are shown in table 2. In addition, we further assessed the relationship between amino acid frequency and FPKM using Spearman rank correlation coefficients (*R*) across all genes (supplementary table S3, Supplementary Material online), which is a correlative method comparable with that used in prior studies (Urrutia and Hurst 2003; Williford and Demuth 2012). The results from the two approaches strongly agree. Specifically, nine of the ten amino acids (Met, Gly, Lys, Ala, Glu, Val, Gln, Asp, Pro; exception Arg) with increased use frequency under high (top 5%) versus low (lowest 5%) expression (table 2; note: Not all statistically significant) had a positive *R* value, and all ten amino acids with decreased use frequency under high expression (Cys, Trp, Asn, Ile, Ser, His, Thr, Phe, Tyr, Leu) had a negative and statistically significant *R* value (supplementary table S3, Supplementary Material online). The *R* values were all of mild or moderate magnitude (absolute values all <0.12), likely due to the inclusion of all genes in correlations, and to inherent variation in the relationship between expression and amino acid frequency at intermediate use values, where selection pressures are apt to be weaker than in the highest expression class. The magnitude of *R* observed here is in line with the *R* values reported in other organisms such as beetles (*T. castaneum*) and humans (Urrutia and Hurst 2003; Williford and Demuth 2012).

Importantly, five of the six amino acids with mid-range S/C scores (between 30.14 and 37.48; Lys, Glu, Gln, Asp, and Pro) exhibited a greater usage in high than low expressed CDS (table 2), and had a statistically significant and positive correlation with expression level (all *R* values statistically significant; supplementary table S3, Supplementary Material online), consistent with their preferential usage under high transcription. In terms of amino acids exhibiting decreased frequency under high expression, we found a statistically significant decline in frequency (between 8.6% and 27.5%) in at least five of the seven amino acids with a large S/C score (>40), namely Cys (S/C score = 57.16), Trp (73.00), His (58.70), Phe (44.00), and Tyr (57.00) (and possibly Arg, S/C score = 56.34; table 2 and supplementary table S3, Supplementary Material online). An exception appears to be Met, which exhibits both a high S/C score and greater usage under elevated expression (table 2 and supplementary table S3, Supplementary Material online). Collectively, the comparison of high and low expressed genes, and the correlation methods used (table 2 and supplementary table S3, Supplementary Material online), indicate that elevated transcription is linked to higher frequency of amino acids with very low (Gly and Ala) and mid-range metabolic costs (Lys, Glu, Gln, Asp, Pro), and decreased frequency of high-cost amino acids (Tyr, Phe, Trp, His, Cys, and possibly Arg; table 2 and supplementary table S3, Supplementary Material online). We note that while we do not exclude the possibility that protein function contributes to reduced or greater usage of specific amino acids, it appears unlikely that the expression levels of the most highly expressed genes is unrelated to their level of usage of five (and possibly six) of the amino acids with the highest S/C scores (>40).

Remarkably, the correlations between amino acid usage and expression levels reported here in spider (table 1) are similar to those reported in humans (Urrutia and Hurst 2003). For instance, Met, Gly, Lys, Ala, Glu, and Val all show a statistically significant positive correlation with expression level in spiders and in humans, while the two amino acids with the greatest negative correlation with expression level in spiders (Cys, Trp) also have among the largest negative correlations in humans. In the red flour beetle *T. castaneum*, it was reported that high expression level correlates with increased usage of the moderate and the lowest cost amino acids, believed to minimize metabolic costs of protein biosynthesis (Williford and Demuth 2012). These amino acids included Gly (S/C score = 1), which was also found to exhibit elevated usage under high expression herein. For the hemimetabolous insects *G. bimaculatus* and *O. fasciatus*, moderate-sized amino acids were preferred under high expression (Whittle and Extavour 2015), while in the Nematoda amino acid usage in the genus Caenorhabditis is also correlated with expression level (Cutter et al. 2006). A common feature of these divergent invertebrate and vertebrate studies is that amino acid usage appears to be subject to selective pressures in highly transcribed genes. As such data remain sparse for metazoans, further studies in other animals, including other nontraditional arthropod models, will be essential to furthering our understanding of how amino acid preferences are involved in shaping the evolution of animal coding DNA (Akashi 2003).

It is noteworthy that certain amino acids showing the greatest increase in frequency with expression level have codons with no T nucleotides in the first or second position, for instance Gly (GGN), Lys (AAA, AAG), and Ala (GCN), while some amino acids with markedly decreased usage under high expression, such as Cys (TGT, TGC), Trp (TGG), and Ile (ATT, ATC, ATA), have a T nucleotide in the nonsynonymous first (a position where 96% of substitutions are nonsynonymous) or second positions (tables 1 and 2). This suggests that while high transcription promotes the usage of T3 optimal codons, it does not promote greater usage of amino acids with T in first or second position under high expression in this spider. As shown in table 2, the amino acid's greatest shifts in frequency (between high vs. low expression) exhibit no consistent preferences for GC-rich or AT-rich codons. Together, this suggests that elevated expression does not favor increased usage of amino acids with specific nucleotides, but does favor/disfavor specific amino acids based on S/C score.

Spider Protein Length and Molecular Evolution

Highly Transcribed Genes Encode Short Proteins

Protein length has been negatively correlated with expression level in various eukaryotes including multiple pancrustaceans (insects and crustaceans), humans, yeast, and some plants (Akashi 2003; Urrutia and Hurst 2003; Comeron 2004; Lemos et al. 2005; Ingvarsson 2007; Whittle et al. 2007; Williford and Demuth, 2012). However, no such effect was observed in one study of *Arabidopsis thaliana*, *C. elegans*, and *D. melanogaster* (Duret and Mouchiroud 1999). We asked whether such a correlation was detectable for the spider by using our prior classifications of expression levels, namely high (above the 95th percentile), moderate (between the 5th and 95th percentile), and low (below the 5th percentile). We report that CDS length decreased markedly with increasing transcript expression level in *P. tepidariorum*. As shown in supplementary figure S3, Supplementary Material online, CDS length decreased progressively from the low expression class (mean CDS length [number of codons] = 457.0 ± 14.8), to the moderate (322.5 ± 2.4), and to the highest expression class (160.9 ± 3.7) (P-value ranked ANOVA $< .001$, all Dunn's paired contrasts $< .05$). Furthermore, within each of the three expression classes, we found a negative correlation between CDS length and FPKM (high expression Spearman's ranked $R = -0.109$, $P = 5.3 \times 10^{-4}$; moderate expression $R = -0.098$, $P = 2.0 \times 10^{-7}$; low expression $R = -0.193$, $P = 4.6 \times 10^{-12}$). In this regard, highly expressed genes generally encode short proteins in this basally branching arthropod, similar to numerous other multicellular eukaryotes (Akashi 2003; Urrutia and Hurst 2003; Comeron 2004; Lemos et al. 2005; Ingvarsson 2007; Whittle et al. 2007; Williford and Demuth 2012). The shorter length of highly transcribed CDS is thought to result from selection to reduce the costs of transcription, translation, and amino acid biosynthesis and transport for abundant proteins (Moriyama and Powell 1998; Akashi 2001, 2003). Accordingly, our collective data herein suggest that CDS length, Fop, and amino acid usage (table 2 and fig. 1) have evolved in concert to promote economical protein synthesis of highly expressed genes in *P. tepidariorum*.

Given that moderately and lowly expressed CDSs are longer (supplementary fig. S3, Supplementary Material online), their reduced Fop (as compared with the highest expression class; fig. 1) might be partly explained by genetic linkage effects including selective sweeps, background selection, and interference among synonymous codon mutations, which may have greater effects in longer coding regions (Comeron et al. 1999; Akashi 2001; Betancourt and Presgraves 2002; Loewe and Charlesworth 2007; Whittle and Johannesson 2013). As an example, under selective sweeps, positive selection on amino acid mutations may drag linked nonoptimal mutations to fixation (Betancourt and Presgraves 2002; Kim 2004; Andolfatto 2007; Whittle and Johannesson 2013). Genes with lower expression levels have also been shown to be subject to weaker purifying selection (Subramanian and Kumar 2004). This may allow an excess of slightly deleterious nonoptimal codon mutations to linger (at polymorphic sites) in populations and ultimately be fixed by selective sweeps, potentially contributing to the low Fop in the moderate and low expressed CDS (which are longer) observed herein (fig. 1). In addition to genetic linkage effects, another potential explanation for reduced Fop among moderately and lowly expressed CDS (as compared with highly expressed CDS) may be that they are nonessential or contribute little or not at all to an organism's fitness, which can lead to relaxed selection (Mank and Ellegren 2009) and the fixation of nonoptimal codon mutations via genetic drift. As selection on synonymous codons is believed to be weak, with NeS ~1 (Akashi 1995; Akashi and Schaeffer 1997), even a modest reduction in selection pressure due to lower expression levels may lead to fixation of nonoptimal codons via neutral drift, and thereby reduce Fop. Similar to optimal codon usage, genetic linkage and/or genetic drift may contribute toward the reduced usage of favored amino acids observed under lower expression (table 2). Further genomic data for a wider range of Parasteatoda species, and population level data, will allow additional tests of the role of genetic linkage, including selective sweeps (Betancourt and Presgraves 2002; Andolfatto 2007; Whittle et al. 2007; Whittle and Johannesson, 2013), and the influence of relaxed selection (Akashi and Schaeffer 1997; Whittle et al. 2012) in shaping codon and amino acid usage in moderately/lowly expressed genes in this spider genus. Nevertheless, our finding that highly expressed CDSs encode short proteins in *P. tepidariorum* (supplementary fig. S3, Supplementary Material online) is consistent with selection to reduce the biochemical costs of transcription and/or translation of major proteins; this highly transcribed subset of genes is apt to be least affected by genetic linkage or relaxed selection due to the shortness of CDS and their high expression, respectively.

Optimal Codon Usage: Interaction between CDS Length and Expression Level Effects

A number of studies in multicellular eukaryotes including animals, plants, and fungi have reported that within expression classes, shorter CDSs have higher codon bias (e.g., *D. melanogaster*, *C. elegans*, and other species of Caenorhabditis, *A. thaliana*, *Silene latifolia*, *Populus tremula*, *Neurospora tetasperma*, and *Neurospora discreta*; Duret and Mouchiroud 1999; Ingvarsson 2007; Cutter 2008; Zeng and Charlesworth 2009; Qiu et al. 2011; Whittle et al. 2011).

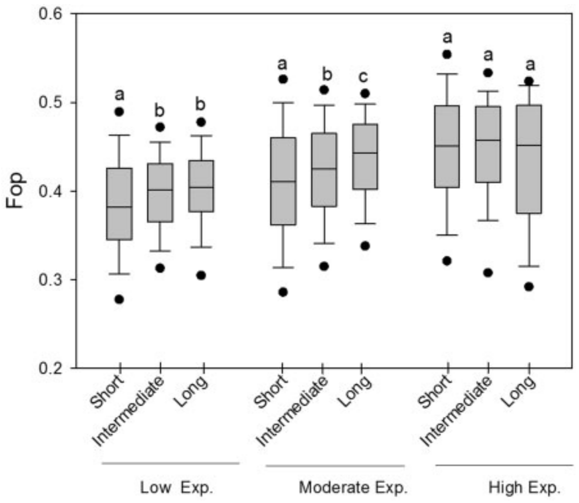

Fig. 2.—Box and whisker plots of Fop relative to expression (Exp.) and CDS length (short, intermediate, long). Different letters within each expression class indicate a statistically significant difference between lengths using ranked ANOVA ($P < 0.001$) and Dunn's paired contrasts ($P < 0.05$).

However, no such gene-length effects were observed in other organisms (e.g., some species of Caenorhabditis and Populus; Cutter 2008; Ingvarsson 2008). Thus, we next considered whether CDS length influences codon usage within expression classes in this spider. To determine this, we conducted a two-way ANOVA with Fop as the dependent variable and gene length (divided into three discrete classes: Short [<250 codons], intermediate [between 250 and 500 codons], and long [>500 codons]) and expression level (high, moderate, low) as categorical parameters. The results revealed that Fop is indeed depended on the interaction between expression level and CDS length ($P = 0.007$; supplementary table S4, Supplementary Material online). Accordingly, we examined Fop for each combination of expression level and CDS length. We report that for the highest expression class, CDS length had no observable effect (fig. 2; ranked ANOVA $P > 0.05$); thus, expression level, rather than the short average length of highly transcribed genes (supplementary fig. S3, Supplementary Material online), shapes Fop in this class (fig. 1). However, gene length did have an effect within the lower expression classes. Specifically, for moderately expressed CDS, Fop increased from the short, to intermediate, to long CDS. Similarly, for lowly expressed CDS, Fop was lower for the short than intermediate and long length classes (ranked ANOVA and Dunn's $P < 0.05$ for all significant contrasts; fig. 2). This implies that longer CDSs have elevated Fop, at least for genes with moderate or low expression levels, and are opposite to results reported in most multicellular eukaryotes studied to date (Duret and Mouchiroud 1999; Ingvarsson 2007; Cutter 2008; Zeng and Charlesworth 2009; Qiu et al. 2011; Whittle et al. 2011). Thus, this spider may be an anomalous multicellular eukaryote in this respect, wherein longer CDSs (among genes with similar expression) exhibit high Fop. The findings suggest that the high costs of longer CDSs in the moderate and low expression classes may be mitigated by enhanced optimal codon usage, and is similar to results reported for some prokaryotes such as *E. coli* (Eyre-Walker 1996; Moriyama and Powell 1998; Akashi 2001). In this respect, these results concur with Eyre-Walker (1996) and Moriyama and Powell (1998) who proposed that for proteins translated at the same level, selection favoring optimal codons should be stronger for long proteins.

It is important to note that, as shown in figure 2, Fop increased from the low, to moderate, to the highest expression class for the short, intermediate, and for the long CDS classes (ranked ANOVA and Dunn's $P < 0.05$; only one paired contrast was nonsignificant for the moderate and high expressed long CDS). These trends further confirm the relationship found between expression level and Fop regardless of CDS length (fig. 1). In this regard, it is evident that expression plays the more pervasive role in shaping codon usage in spider, and that CDS length has a minor, though not irrelevant, effect that is restricted to the moderately and lowly expressed CDS classes. Taken together, we argue that the greatest effect of CDS length on spider fitness might not be its role in codon usage, but rather in the fact that highly transcribed genes are shorter than their moderate and low expressed counterparts (supplementary fig. S3, Supplementary Material online) reducing the biosynthetic costs of abundant transcripts and proteins.

Orthology and Function of Spider Genes under Study

We next wished to identify *D. melanogaster* orthologs for our genes under study. The objective in identification of these orthologs was to allow us to (1) examine the likely functions of the spider genes studied herein, and (2) determine whether the orthologous genes in a divergent arthropod displayed similar relationships between codon use and expression levels as those found in the spider. To this end, we identified the likely *D. melanogaster* orthologs of *P. tepidariorum* genes, using the top hit resulting from BLASTX to the *D. melanogaster* protein list with a cutoff value of 10^{-6} (see Methods). We note that using a reciprocal best BLAST hit criterion to identify putative orthologs yielded the same results (data not shown). Out of the 19,936 spider CDS under study herein, we identified putative *D. melanogaster* orthologs for 10,830 of these (54.3%), reflecting the substantial divergence among these arthropods. In addition to using *D. melanogaster* protein sequences, we attempted to identify additional putative orthologs for the spider CDS set using the protein database Swissprot (http://www.uniprot.org; last accessed March 17, 2016), but found little improvement ($N = 11,141$, compared with the 10,830 obtained by

Table 3

Functional Clustering of Highly Expressed CDS (all CDS above the 95th percentile of FPKM) in *Parasteatoda tepidariorum* Embryos Using Their Orthologs in the Model *Drosophila melanogaster* and the Gene Ontology System DAVID (Huang da et al. 2009)

	***P* value**
Cluster 1: Enrichment score 35.68	
Cytosolic part	4.00×10^{-48}
Cytosolic ribosome	1.80×10^{-44}
Ribosomal protein	6.60×10^{-42}
Ribosome	1.10×10^{-39}
Structural constituent of ribosome	4.60×10^{-38}
Ribosomal subunit	6.00×10^{-31}
Ribosome	8.40×10^{-29}
Structural molecule activity	3.40×10^{-19}
Cluster 2: Enrichment score 24.41	
Mitotic spindle organization	6.40×10^{-36}
Spindle organization	3.50×10^{-32}
Microtubule cytoskeleton organization	4.00×10^{-30}
Microtubule-based process	3.70×10^{-28}
Mitotic cell cycle	2.60×10^{-27}
Cytoskeleton organization	2.90×10^{-26}
Cell cycle	2.30×10^{-18}
Cell cycle phase	1.10×10^{-17}
Cell cycle process	1.10×10^{-17}
M phase	1.30×10^{-17}
Cluster 3: Enrichment score 9.26	
Mitochondrial ATP synthesis coupled electron transport	3.10×10^{-12}
ATP synthesis coupled electron transport	1.20×10^{-11}
Respiratory electron transport chain	3.20×10^{-11}
Electron transport chain	4.20×10^{-10}
Cellular respiration	2.40×10^{-9}
Energy derivation by oxidation of organic compounds	1.50×10^{-8}
Respiratory chain	1.50×10^{-8}
Mitochondrial respiratory chain	3.30×10^{-8}
Cluster 4: Enrichment score 7.32	
Hydrogen ion transmembrane transporter activity	6.00×10^{-9}
Monovalent inorganic cation transmembrane transporter activity	9.20×10^{-9}
Inorganic cation transmembrane transporter activity	1.90×10^{-6}

Note.—*P* values are from a modified Fisher's test, wherein lower values indicate greater enrichment. The four GO clusters with the highest enrichment scores are shown.

comparing with *D. melanogaster* alone). This may be because even more divergent reference species are included in that database (the top five species represented are human [*Homo sapiens*], mouse [*Mus musculus*], the plant thale cress [*A. thaliana*], rat [*Rattus norvegicus*], and yeast [*S. cerevisiae*]; http://web.expasy.org/docs; last accessed March 17, 2016). The proportion of likely orthologs found here for spider genes using fly for comparison is in line with results observed for other animal taxa sharing a most recent last common ancestor in a similar time frame; for instance, about 70% of genes in humans are suggested to have likely orthologs in zebrafish (Howe et al. 2013). We speculate that the subset of spider genes without clearly defined orthologs in *D. melanogaster* or Swissprot has evolved rapidly, precluding likely orthology assignment via sequence similarity alone, and/or that this subset contains some genes specific to the spider lineages.

Spider Gene Orthologs and Functional Annotation

Having identified likely *D. melanogaster* orthologs of a majority of the spider genes under study herein, we next studied GO for the highly expressed CDS set. We used *D. melanogaster* for this GO annotation as it is well annotated (www.flybase.org; last accessed March 17, 2016). Using the GO tool DAVID (Huang da et al. 2009), we found that the highly expressed CDSs from *P. tepidariorum* were enriched in genes involved in protein synthesis (see cluster 1; table 3), including ribosomal proteins, cytosolic ribosomes, and ribosomal subunits, consistent with high transcription rates reported for these essential genes in other organisms (Wang et al. 2011; Whittle and Extavour 2015). The second group (cluster 2; table 3) included genes involved in the cell cycle, including functions classified as mitotic cell cycle, spindle organization, and cell cycle processes. To a lesser extent, the high expression class also contained various housekeeping genes such as those involved in mitochondrial activity and ATP synthesis (clusters 3 and 4; table 3). Thus, we conclude that the highly expressed genes in *P. tepidariorum*, which are most prone to translational selection on codon usage (table 1 and fig. 1), shifts in amino acid composition (table 2), and protein lengths, are largely those involved in protein synthesis and cell cycling. This concurs with recent findings from other emerging arthropod models (Whittle and Extavour 2015).

Arthropod embryos are highly complex structures, spanning an array of developmental stages and tissue/cell types (Diez-Roux et al. 2011; Mittmann and Wolff 2012; Combs and Eisen 2013; Schwager et al. 2014; Donoughe and Extavour 2016). Although many of the moderately and lowly expressed genes studied herein may play specific roles in embryonic development, they appear to be less prone to selective pressures, perhaps because they are innately less costly due to lower transcription and translation levels. Furthermore, some of these genes might be more dispensable to embryogenesis, or have minimal fitness effects (Hirsh and Fraser 2001; Mank and Ellegren 2009), particularly those from the lowest expression class, as compared with those from the highest expression class (table 3).

Codon Usage in Drosophila Orthologs

Finally, we assessed whether spider genes and their orthologs in the fly *D. melanogaster* shared similar optimal codon usage patterns. Although this spider has T3 optimal codons (table 1),

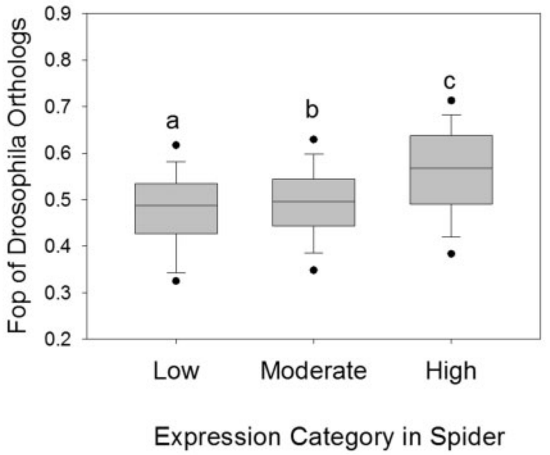

Fig. 3.—The Fop of *Drosophila melanogaster* orthologs of *Parasteatoda tepidariorum* genes with low, moderate, and high expression levels. The *D. melanogaster* optimal codons, which preferentially end in G or C, were taken from Duret and Mouchiroud (1999). Different letters indicate a statistically significant difference using ranked ANOVA ($P < 0.001$) and Dunn's paired contrast ($P < 0.05$).

the fly taxon is known to have GC3 optimal codons (Duret and Mouchiroud 1999; Maside et al. 2004). We used the optimal codon lists from Duret and Mouchiroud (1999) and calculated Fop for the fly orthologs to spider CDS from the low, moderate, and high expression classes. As shown in figure 3, we found that indeed *D. melanogaster* orthologs of genes that were highly expressed in the spider had statistically significant higher Fop than those orthologs from the moderate or low expression classes (ranked ANOVA $P < 0.001$, Dunn's paired contrast $P < 0.05$), indicating that the optimization of codon usage is conserved across these orthologous gene sets, despite the fact that these two organisms have markedly different types of optimal codons (T3 versus GC3). Although the greater usage of optimal codons, which mostly end in GC3 in the fly, automatically implies higher expression, we verified that these orthologs had qualitatively similar expression profiles to those found in *P. tepidariorum*. For this, we used the RNA database from modEncode (Graveley et al. 2011) and studied expression (given in modEncode as RPKM) for fly embryos, using data spanning an array of developmental stages (every 2 h between 0 and 24 h). Indeed, we found that the expression level of fly orthologs to genes in the high, moderate, and low expression classes in *P tepidariorum* decreased in an anticipated manner: The mean RPKM values were 417.2 ± 12.4, 120.9 ± 2.0, and 90.9 ± 7.6, respectively (*P* value ranked ANOVA $< .001$, each Dunn's contrast $P < 0.05$). Thus our results provide an important example showing that optimized codon usage is conserved across orthologs from two highly divergent systems, even when the types of optimal codons are markedly different.

The distinct optimal codons (T3) in *P. tepidariorum* as compared with *D. melanogaster* suggest that a major shift in codon usage has occurred since the divergence of these arthropods. There are several feasible mechanisms that could explain this shift. It is possible that this shift was mediated by reductions in population size, which weaken selection on codon usage, allowing new optimal codons and tRNA populations (often measured as number of tRNA genes; Duret 2000) to emerge upon population expansion. For example, in *Drosophila willistoni*, it has been found that the most preferred codons are T3, unlike other Drosophila species, which have G3 or C3 preferred codons. It has been proposed that this might arise from population size changes, or from shifts in tRNA pools, over the evolutionary history within this single genus (Vicario et al. 2007). Accordingly, it is reasonable to infer that comparable shifts might have occurred in population sizes and tRNA pools since the divergence of spiders and flies, which are from different arthropod subgroups (Chelicerata and Pancrustacea, respectively). An alternate explanation for the differences in optimal codons between spiders and flies is that the expression of tRNA genes differs markedly across taxa (as their expression can even vary among tissues; Moriyama and Powell 1998). Moreover, it is worth noting that for 2-fold degenerate sites and many 3-, 4-, and 6-fold sites, a shift from C3 optimal codons (in Drosophila) to T3 optimal codons (in Parasteatoda) could be achieved without any changes in tRNA pools due to wobble positions (e.g., for Asp, GAT and GAC may be matched to the same tRNA anticodon, GTC). Future analyses of tRNA populations in spider will help ascertain the causes of divergence in the types of optimal codons among *D. melanogaster* and *P. tepidariorum* and the putative role of translational selection.

Conclusions

We have shown that the emerging spider model *P. tepidariorum* provides a valuable system to study factors underlying the molecular evolution of protein-coding genes in a noninsect arthropod. Our assessment revealed not only that codon usage and protein lengths vary with expression in this basally branching arthropod, but also that amino acid usage, which remains poorly studied in the literature, is connected to transcription level. A propensity to utilize specific amino acids in highly transcribed genes, if ultimately reported in a broad range of organisms, would suggest that expression-mediated selection on amino acid frequency is a significant factor in genome evolution that can contribute to minimizing costs of protein biosynthesis and stability, and shaping the relative abundance of amino acids in the genome. Findings of expression-related preferences for specific codons, inexpensive amino acids, and short CDS lengths, as observed herein, are consistent with translational selection to promote cost-efficient protein synthesis (Duret and Mouchiroud 1999; Duret 2000; Akashi 2003; Urrutia and Hurst 2003; Cutter et al. 2006; Ingvarsson 2008; Williford and Demuth 2012) but might also arise from selection for cost-efficient transcription

(Urrutia and Hurst 2003; Trotta 2011). Our data suggest that protein synthesis and cell cycling genes are most prone to selection on their codons, amino acids, and CDS lengths, likely enhancing organism-level fitness by reducing the costs of synthesizing these abundant proteins, which are inherently essential to all tissue types and stages of development. Expanding research on codon usage, amino acid frequency, and protein lengths to a wider range of nontraditional model organisms will help further elucidate the dynamics underling the evolution of protein coding DNA, and the importance of expression-mediated selection in shaping the evolution of eukaryotic genomes.

Supplementary Material

Supplementary tables S1–S4 and figures S1–S3 are available at *Genome Biology and Evolution online* (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank the Harvard Faculty of Arts and Sciences Research Computing and Informatics groups for their Odyssey support services, particularly Michele Clamp. In addition, the authors are grateful to Stephen Richards and the BCM-HGSC i5K pilot project, and Alistair McGregor and the *P. tepidariorum* Genome Consortium for the use of prepublication genome and annotated gene set data (manuscript in preparation). The data can be found at the National Agricultural library (https://i5k.nal.usda.gov/Parasteatoda_tepidariorum; last accessed March 17, 2016) as well as at the NCBI. The comments by the anonymous reviewers are greatly appreciated. This work was partially supported by NIH grant number 5R01-HD073499.

Literature Cited

Akam M. 2000. Arthropods: development diversity within a (super) phylum. Proc Natl Acad Sci U S A. 9:4438–4441.

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics 139:1067–1076.

Akashi H. 2001. Gene expression and molecular evolution. Curr Opin Genet Dev. 11:660–666.

Akashi H. 2003. Translational selection and yeast proteome evolution. Genetics 164:1291–1303.

Akashi H, Schaeffer SW. 1997. Natural selection and the frequency distributions of "silent" DNA polymorphism in *Drosophila*. Genetics 146:295–307.

Akiyama-Oda Y, Oda H. 2003. Early patterning of the spider embryo: a cluster of mesenchymal cells at the cumulus produces *Dpp* signals received by germ disc epithelial cells. Development 130:1735–1747.

Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. Genome Res. 17:1755–1762.

Bachtrog D. 2007. Reduced selection for codon usage bias in *Drosophila miranda.* J Mol Evol. 64:586–590.

Behura SK, Severson DW. 2011. Coadaptation of isoacceptor tRNA genes and codon usage bias for translation efficiency in *Aedes aegypti* and *Anopheles gambiae.* Insect Mol Biol. 20:177–187.

Behura SK, Severson DW. 2013. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. Biol Rev Camb Philos Soc. 88:49–61.

Beletskii A, Bhagwat AS. 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. Proc Natl Acad Sci U S A. 93:13919–13924.

Betancourt AJ, Presgraves DC. 2002. Linkage limits the power of natural selection in *Drosophila*. Proc Natl Acad Sci U S A. 99:13616–13620.

Colbourne JK, et al. 2011. The ecoresponsive genome of *Daphnia pulex.* Science 331:555–561.

Combs PA, Eisen MB. 2013. Sequencing mRNA from cryo-sliced *Drosophila* embryos to determine genome-wide spatial patterns of gene expression. PLoS One 8:e71820.

Comeron JM. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics 167:1293–1304.

Comeron JM, Kreitman M, Aguade M. 1999. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics 151:239–249.

Cutter AD. 2008. Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. Mol Biol Evol. 25:778–786.

Cutter AD, Wasmuth JD, Blaxter ML. 2006. The evolution of biased codon and amino acid usage in nematode genomes. Mol Biol Evol. 23:2303–2315.

De La Torre AR, Lin YC, Van de Peer Y, Ingvarsson PK. 2015. Genome-wide analysis reveals diverged patterns of codon bias, gene expression, and rates of sequence evolution in Picea gene families. Genome Biol Evol. 7:1002–1015.

Diez-Roux G, et al. 2011. A high-resolution anatomical atlas of the transcriptome in the mouse embryo. PLoS Biol. 9:e1000582.

Donoughe S, Extavour CG. 2016. Embryonic development of the cricket *Gryllus bimaculatus.* Dev Biol. 411:140–156.

Dufton MJ. 1997. Genetic code synonym quotas and amino acid complexity: cutting the cost of proteins? J Theor Biol. 187:165–173.

Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. 16:287–289.

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

Eyre-Walker A. 1996. Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol Biol Evol. 13:864–872.

Farlow A, Dolezal M, Hua L, Schlotterer C. 2012. The genomic signature of splicing-coupled selection differs between long and short introns. Mol Biol Evol. 29:21–24.

Graveley BR, et al. 2011. The developmental transcriptome of *Drosophila melanogaster.* Nature 471:473–479.

Green P, Ewing B, Miller W, Thomas PJ, Green ED. 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet. 33:514–517.

Group NGW, et al. 2010. Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science 327:343–348.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. Genome Biol. 6:R67.

Hilbrant M, Damen WGM, McGregor A. 2012. Evolutionary crossroads in developmental biology: the spider *Parasteatoda tepidariorum.* Development 139:2655–2663.

Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. Nature 411:1046–1049.

Howe K, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4:44–57.

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 synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol. 151:389–409.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 2:13–34.

Ingvarsson PK. 2007. Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula.* Mol Biol Evol. 24:836–844.

Ingvarsson PK. 2008. Molecular evolution of synonymous codon usage in Populus. BMC Evol Biol. 8:307.

Jia X, et al. 2015. Non-uniqueness of factors constraint on the codon usage in *Bombyx mori.* BMC Genomics 16:356.

Kim Y. 2004. Effect of strong directional selection on weakly selected mutations at linked sites: implication for synonymous codon usage. Mol Biol Evol. 21:286–294.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. 22:1345–1354.

Loewe L, Charlesworth B. 2007. Background selection in single genes may explain patterns of codon bias. Genetics 175:1381–1393.

Mank JE, Ellegren H. 2009. Are sex-biased genes more dispensable? Biol Lett. 5:409–412.

Maside X, Lee AW, Charlesworth B. 2004. Selection on codon usage in Drosophila americana. Curr Biol. 14:150–154.

McGregor A, et al. 2008. *Cupiennius salei* and *Achaearanea tepidariorum*: spider models for investigating evolution and development. BioEssays 30:487–498.

McVean GA, Vieira J. 1999. The evolution of codon preferences in *Drosophila*: a maximum-likelihood approach to parameter estimation and hypothesis testing. J Mol Evol. 49:63–75.

Mittmann B, Wolff C. 2012. Embryonic development and staging of the cobweb spider *Parasteatoda tepidariorum* C. L. Koch, 1841 (syn.: *Achaearanea tepidariorum*; Araneomorphae; Theridiidae). Dev Genes Evol. 222:189–216.

Moriyama EN, Powell JR. 1998. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli.* Nucleic Acids Res. 26:3188–3193.

Neafsey DE, et al. 2015. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 *Anopheles mosquitoes.* Science 347:1258522.

Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol. 19:1390–1394.

Oda H, et al. 2007. Progressive activation of Delta-Notch signaling from around the blastopore is required to set up a functional caudal lobe in the spider *Achaearanea tepidariorum.* Development 134:2195–2205.

Odegaard F. 2000. How many species of arthropods? Erwin's estimate revised. Biol J Linn Soc. 71:583–597.

Osawa S, et al. 1988. Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. Proc Natl Acad Sci U S A. 85:1124–1128.

Peden JF. 1999. Analysis of codon usage [ph.d. thesis]. [Nottingham (UK)]: University of Nottingham.

Percudani R. 2001. Restricted wobble rules for eukaryotic genomes. Trends Genet. 17:133–135.

Posnien N, et al. 2014. A comprehensive reference transcriptome resource for the common house spider *Parasteatoda tepidariorum.* PLoS One 9:e104885.

Puigbo P, Bravo IG, Garcia-Vallve S. 2008. CAIcal: a combined set of tools to assess codon usage adaptation. Biol Direct. 3:38.

Qiu S, Bergero R, Zeng K, Charlesworth D. 2011. Patterns of codon usage bias in *Silene latifolia.* Mol Biol Evol. 28:771–780.

Raiford DW, et al. 2008. Do amino acid biosynthetic costs constrain protein evolution in *Saccharomyces cerevisiae*? J Mol Evol. 67:621–630.

Rao Y, et al. 2011. Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. DNA Res. 18:499–512.

Regier JC, et al. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463:1079–1083.

Richards S, et al. 2008. The genome of the model beetle and pest *Tribolium castaneum.* Nature 452:949–955.

Sanggaard KW, et al. 2014. Spider genomes provide insight into composition and evolution of venom and silk. Nat Commun. 5:3765.

Schwager EE, Meng Y, Extavour CG. 2014. *vasa* and *piwi* are required for mitotic integrity in early embryogenesis in the spider *Parasteatoda tepidariorum.* Dev Biol. 402:276–290.

Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF. 1995. DNA sequence evolution: the sounds of silence. Philos Trans R Soc Lond B Biol Sci. 349:241–247.

Sharp PM, Devine KM. 1989. Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do 'prefer' optimal codons. Nucleic Acids Res. 17:5029–5039.

Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14:5125–5143.

Shields DC, Sharp PM, Higgins DG, Wright F. 1988. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol. 5:704–716.

Stark A, et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. Nature 450:219–232.

Stenico M, Lloyd AT, Sharp PM. 1994. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res. 22:2437–2446.

Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol Biol Evol. 24:374–381.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168:373–381.

Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci U S A. 85:2653–2657.

Supek F, Vlahovicek K. 2005. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics 6:182.

Trotta E. 2011. The 3-base periodicity and codon usage of coding sequences are correlated with gene expression at the level of transcription elongation. PLoS One 6:e21590.

Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. Genome Res. 13:2260–2264.

Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of *Drosophila*. BMC Evol Biol. 7:226.

Wang B, et al. 2011. Optimal codon identities in bacteria: implications from the conflicting results of two different methods. PLoS One 6:e22714.

Weinstock GM, et al. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443:931–949.

Whittle CA, Extavour CG. 2015. Codon and amino acid usage are shaped by selection across divergent model organisms of the Pancrustacea. G3 (Bethesda) 5:2307–2321.

Whittle CA, Johannesson H. 2013. Evolutionary dynamics of sex-biased genes in a hermaphrodite fungus. Mol Biol Evol. 30:2435–2446.

Whittle CA, Malik MR, Krochko JE. 2007. Gender-specific selection on codon usage in plant genomes. BMC Genomics 8:169–179.

Whittle CA, Sun Y, Johannesson H. 2011. Evolution of synonymous codon usage in *Neurospora tetrasperma* and *Neurospora discreta*. Genome Biol Evol. 3:332–343.

Whittle CA, Sun Y, Johannesson H. 2012. Genome-wide selection on codon usage at the population level in the fungal model organism *Neurospora crassa*. Mol Biol Evol. 29:1975–1986.

Wiegmann B, Yeates DK. 2005. The evolutionary biology of flies. New York: Columbia University Press.

Williford A, Demuth JP. 2012. Gene expression levels are correlated with synonymous codon usage, amino acid composition, and gene architecture in the red flour beetle, *Tribolium castaneum*. Mol Biol Evol. 29:3755–3766.

Zeng K, Charlesworth B. 2009. Estimating selection intensity on synonymous codon usage in a nonequilibrium population. Genetics 183:651–662. 651SI-623SI.

Zuk M, Garcia-Gonzalez F, Herberstein ME, Simmons LW. 2014. Model systems, taxonomic bias, and sexual selection: beyond *Drosophila*. Annu Rev Entomol. 59:321–338.

Associate editor: Laurence Hurst